AI in Drug Discovery and Development

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

Week-10
Lecture-50

Welcome to the course "AI in Drug Discovery and Development." In this session, we will talk about how to do network pharmacology with the help of AI. So, by the end of this lecture, you will be able to understand the system-based shift in interactive discovery and the fundamentals of network pharmacology. Explain the workflow and construction methods used in network pharmacology, interpret key network matrices, and apply enrichment tools to identify relevant pathways. Discuss current challenges in network pharmacology and the role of AI in addressing them. So, we have seen a paradigm shift in drug discovery where we are moving from the traditional, you know, drug discovery approach to the systems biology and network-based approach.

So, because the traditional drug discovery approach follows a linear and reductionist model. So, because it believes that the philosophy of one target, one drug, and one disease is the foundation, we are only one drug binding to a single target, and that is curing the single disease. And it prioritizes single molecular targets identified through hypothesis-driven research; it is effective for monogenic or infectious diseases with well-defined pathology. However, it ignores the, you know, complex interplay of genes, proteins, and pathways, and therefore, it has a quite high attrition rate in clinical phases due to off-target effects or limited efficacy.

On the other hand, the systems biology and network-based approaches embrace a holistic view of the biological system. So, they believe that a drug is, you know, interacting with multiple targets, and then those targets are interlaced into a kind of network where all of them are affected by all of the others, actually. and ultimately the net you know contribution from all of the targets that leads to the disease symptoms. So, it recognizes that most diseases, especially the chronic and complex ones like cancer, Alzheimer's, or diabetes, are multigenic and multifactorial, and it focuses on the interconnectivity and redundancy within the cellular networks. So, it aims to understand how multiple targets, pathways, and cellular modules interact with each other and leverages the network model to identify potential drug intervention points beyond a single molecule.

So if we talk about network pharmacology, it's a systems-level approach that integrates pharmacology, bioinformatics, and network science to understand drug action across biological networks rather than isolated targets. So it explores multitarget drug actions and

how they affect complex disease mechanisms involving interconnected molecular pathways. So, it has emerged from the convergence of systems biology, which studies the complex biological interactions, classical pharmacology, which focuses on the molecular targets of drugs, and network science, which is the analysis of complex networks in biology and medicine. So, it was first formalized in 2007 by Hopkins, who introduced the term network pharmacology as an evolution from the reductionist paradigm. So, if you look up, look at the development.

So, in 2007, when it was, you know, developed as network-based drug combination discovery, network pharmacology, and drug target networks were identified, or, you know, this hypothesis was developed. And then, in 2050, it mapped the unexplored target space and therapeutic potential of natural products. In 2019, the transition from a single drug target to synergistic network pharmacology was considered. And then in 2022, network pharmacology is curing causal mechanisms instead of treating the symptoms. So, we are talking about a drug that acts holistically on the whole system and cures the disease instead of just, you know, relieving the symptoms by targeting a single target.

So, what kind of role is network pharmacology playing in modern drug discovery? So, you know one of the important aspects is that it reveals polypharmacology. So, it identifies how a single drug interacts with multiple targets, which uncovers hidden therapeutic or adverse effects. It also enables the rational design of multi-target drugs to enhance efficacy and reduce resistance. It also helps in identifying synergistic drug combinations by predicting synergistic mechanisms through network proximity and co-targeting. It also guides the design of combination therapies for complex diseases like cancer, TB, and COVID-19.

It maps the disease modules as well as clusters disease-associated genes or proteins through clustering disease-associated genes or proteins into functional modules for mechanistic insight. By identifying hub nodes and bottlenecks as potential therapeutics or biomarker targets. So it also helps us understand the drug side effects by analyzing the network propagation of drug action leading to off-target effects and by mapping adverse outcome pathways to predict toxicity and adverse drug reactions in silico. So, if we look at the applications of network pharmacology, the application of network pharmacology is huge; as I said, we can predict the side effects. We can do the, you know, drug repurposing, and through that, we can do a faster drug discovery.

We can find new uses for the old drugs, and by predicting the side effects, we can also address the off-target effects. So, if we identify the off-target side effects, that can also be a starting point for the development or repurposing of the drugs. And then it can help us, you know, personalize the medicine, where it can tailor the treatment plan, and it can also

help us develop patient-specific therapies. And then it can also help us develop synergistic drug combinations where drug synergy discovery can be done. And then, you know, with a low dose, high efficacy can be obtained if you know that there is a synergistic drug combination, meaning the drugs are working together to achieve the same effect at a lower dose, actually.

And then it can help us understand the mechanism of action of the drugs, as well as where the multi-target insights can be obtained through the pathway interaction. And then again, polypharmacology, where we can use a kind of, you know, develop them as multi-target drug designs; we can develop drugs as, you know, targeting multiple targets. As well as we can use them for treating complex diseases because, you know, polypharmacology is usually successful in complex diseases where not only a single pathology is at work. There are multiple pathologies working simultaneously, and that is how we can use, you know, the network pharmacology in this aspect, actually. So if we talk about the general workflow of network pharmacology, we start by defining the biological question.

So whether our purpose is drug repurposing or we want to, you know, do disease module discovery or target prioritization. Once we have defined the question, the next step is data collection. Where we can do that from the different databases, or we can use multi-omics data sets, or we can do it from the literature. And then, after the data collection, the next step is the network construction, where we define the nodes as network types. And then we do the network analysis, such as topological functional analysis and scoring, followed by AI and ML learning integration.

AI ML integration allows us to predict associations, rank targets, and generate a hypothesis where we prioritize candidates for testing. And then perform the experimental validation where we use the in vitro and in vivo methods to experimentally validate our hypothesis, which we have generated through analyzing the networks. and then finally, we you know iteratively read refine the you know the whole process actually. So, we update the model with new data and further go on until we get, you know, satisfying wet lab results, and this is a kind of iterative process where it keeps on going until we get optimal results, actually. So let us see what those steps are and how we do that in detail.

So the first step is defining the research objective; this is a foundational step in network pharmacology that begins by identifying a focused biological or therapeutic question. So this step guides the entire workflow by setting the scope and relevance of the study. So some of those common objectives include understanding the mechanism of action of a known drug or a natural product. Exploring potential repositioning opportunities for the approved drugs or illustrating how complex herbal formulas influence multiple molecular targets. And then some of the tools that we use for defining the research objective are

scientific literature databases like PubMed and Google Scholar.

Clinical data sets, curated case studies, disease ontologies, and systems biology reviews. So, after we have defined the research question, the second step is data collection, which means building the knowledge base. So, this step focuses on gathering high-quality multidimensional data to form the foundation of the network. So, we collect relevant data for all network components, such as the drug compounds, the name structure, known targets, and all the information about, you know, the drug molecules. And then gene and protein-related data, where we try to get information on how those genes and proteins are involved in disease or drug response.

And disease and pathways data where we try to find how those pathways are related to the compound or the formula means the formulation that we are using for identifying its target. And the goal is to integrate data across different biological levels to ensure that a comprehensive network can be constructed. So, these are, you know, some of the resources where you can get the data from. If you are interested in getting the drug target data, you can obtain it from DrugBank, ChEMBL, or BindingDB. For disease gene, DisGeNET, OMIM, or CTD.

So, it contains the disease gene dataset, and then protein-protein interaction data can be obtained from STRING, BioGRID, or IntAct. And then your pathway dataset can be obtained from KEGG or Reactome. And then herbal medicine data can be obtained from TCMSP, TCMID, and ETCM. Omics data can be obtained from GEO, which is part of NCBI, TCGA, or LINCS. And then literature mining can be done using, you know, Pubtator, BioBert, or LitVar.

So these are some NLP-based tools that can be used to obtain data from the literature and from literature mining. So these are not only resources—there are plenty of them—but I've shown just some of them here to give you an idea of what those resources are and where we can get the data from. So once we get this data, we actually build the knowledge base. The next step is network construction. So, network construction involves organizing collected data into a structured graph-based model where networks are built by defining the nodes and the edges.

And where nodes, you know, those are the biological entities like genes, proteins, drugs, or diseases. and the edges represent the relationship or interactions between these entities. So there could be several types of networks, such as homogeneous networks, where we have a single type of nodes, like protein-protein interactions. Where we can have a heterogeneous kind of network with different types of nodes, such as a drug-gene-disease interaction network. And then we have a bipartite or multipartite network where we have

a specialized heterogeneous network with strict pairing rules for both within and across the types.

And then some of the tools that can be used for network construction are Cytoscape, where we can have interactive visualization as well. Then we have NetworkX, which is a Python-based, you know, network construction tool. And then we have Gephi or NDX, which is a kind of tool for large-scale exploration. So there are, you know, multiple network construction methods that exist and can be used. So the first one is, you know, a knowledge-based strategy that builds biological networks using curated data from established sources like OMIM, DrugBank, KEGG, STRING, HIT, and Medline.

And then it employs text mining techniques like co-occurrence frequency, bubble bootstrapping, and NLP to extract the relationships. So it manually integrates protein-protein interactions, drug targets, and disease gene associations. And the strength of this knowledge base strategy is that it has high reliability and interpretability due to expert-verified data. Because the data we are using is coming from those databases, those databases are curated carefully to obtain only reliable data. And then you can use the experiment-based strategy in which we construct the networks using high-throughput omics data like microarray, proteomics, or metabolomics.

So, here we focus on identifying differentially expressed genes or proteins and linking them through known PPIs or co-expression patterns. An example is a network built from microarray data for the Qishen-Yuki formula in myocardial infarction. So the strength of the experimental-based strategy is that it is useful for uncovering context-specific biological responses. And then we have a computation-based strategy where we leverage these computational algorithms to predict novel interactions and construct networks in silico. So it uses methods like molecular docking to predict compound target binding, ML methods like SVM and random forest for DTI prediction, and similarity-based modeling based on chemical target similarity.

An example is predictive modeling of a compound target network for Qi-enriching versus a blood-tonifying herbal compound. And then the strength of this computationally based strategy is that it is highly scalable and cost-efficient for exploring large chemical spaces. Okay, so after the network building, the next step is network analysis. So, analyzing the structure of the network reveals biologically meaningful insights such as identifying essential genes or proteins and diseases, discovering potential multi-target drugs, or locating functional disease modules. So, some of the topological matrices are used to identify key players.

So these are degree centrality, which are the nodes with the most interactions considered

as the hubs, and then betweenness centrality, where we are using the nodes connecting different regions of the network. And then closeness centrality, where we identify the nodes with the shortest path to the others, and modularity, where we cluster modules representing the functional units. So some of the tools we use for network analysis again are Cytoscape plugins like Network Analyzer, MCODE, which is a module detection tool, and ClueGO, which is used for pathway enrichment. And then there are different packages, such as R packages like igraph and bionet, and Python packages like NetworkX and PyVis. Okay, after doing that, the next step is functional and enrichment analysis.

The functional enrichment analysis involves statistically assessing whether specific biological annotations, like a pathway or process, are overrepresented among a selected list of genes compared to a reference background. For example, the whole genome or transcriptome, so the main objective of, you know, functional enrichment analysis is to translate abstract network data into biological meaning. To identify biological processes, signaling pathways, and disease associations enriched among the genes and proteins in the network, and to prioritize biologically significant modules or targets for downstream validation or therapeutic intervention. So there are multiple types of enrichment analyses. For example, one is the gene ontology enrichment analysis, where we can do the analysis for biological processes like apoptosis and inflammation.

So which of those biological processes are actually involved in that network? And then you have molecular functions based on, for example, kinase activity, receptor binding, and cellular components, like where those pathways or genes are present, such as the nucleus or plasma membrane. So, it helps to contextualize what the genes or proteins do and where they act. So, these basically tell us the identified genes and proteins, where they are involved, where they are present, and what kind of functions they are performing or for which they are responsible. And then we can do the pathway enrichment analysis, where we can map the genes or proteins to known signaling and metabolic pathways. And some of the common databases for pathways are KEGG, Reactome, and WikiPathways.

These are useful for identifying key cascades modulated by a drug or formula. And then we can perform the disease association enrichment, where we can identify the diseases or syndromes associated with the genes or proteins, and it is useful for understanding polypharmacological effects or off-target implications. And then we can do a custom module enrichment when networks are clustered into modules; each module can be enriched separately to find its distinct biological role. So these are some of the tools, you know, which are used for this enrichment analysis, like you have DAVID, where you can perform gene ontology and pathway enrichment analysis, and it is very simple and widely used. You have the G profiler, where you can do gene enrichment and KEGG and Reactome analysis, and it also offers background customization.

You have the enricher, which has extensive libraries and is quite fast, and it includes transcription factor and drug enrichment as well. You have the Metascape. This is a kind of multi-tool integration you can use, and it usually produces beautiful visual outputs. You have the clue to go where it is a kind of site for Cytoscape plug-ins for gene enrichment or pathway analysis. And then it can show the network visualization of the enriched terms as well.

And you have the DisGeNET, which is used for disease gene mapping and clinical translation. After that, we have, you know, the next step, which is where we predict the new drug targets or disease drug interactions using AI or ML. To enhance traditional network pharmacology by applying ML or AI for predicting unknown, novel, and hidden relationships between drug targets and diseases, especially in large, complex data sets where manual analysis is not feasible. But why do we need AI in network pharmacology? Because biological networks are high-dimensional and non-linear, making them ideal for AI-based pattern recognition. And then the traditional methods, like working or co-expressions, often miss the non-obvious associations because those AI methods can, you know, learn very small features as well.

So, that is how they can identify the relationship between disease and drug and identify a new target as well. And those AI models can learn from existing biological and pharmacological data to predict the new interactions and it enables drug repurposing, polypharmacology discovery, side effect prediction and synergistic combinations. So the type of predictions we can make using AIR for drug target interaction is something we have studied in detail in earlier sessions, where we can predict which proteins a compound is likely to bind to. And then we can do the drug target association studies where we can identify, you know, potential therapeutic uses for the existing drugs. It was formerly known as repurposing, and then we can do the target pathway gene disease links as well, where we can infer new targets involved in disease progression or therapy.

And we can, you know, predict the drug combinations as well, where we can predict synergistic or antagonistic interactions between the drug pairs. The techniques and algorithms we use are the classical machine learning methods, such as supervised models like random forest, SVM, and XGBoost, to classify drug-target pairs based on known interactions using molecular descriptors. Or we can use deep learning, where we can learn hierarchical features from raw data like SMILES or gene expression, and use CNNs for molecular graphs, autoencoders, and recommenders for the latent representation. And then we can use graph-based deep learning methods that can work on molecular or biological interaction graphs to predict novel links in drug-disease target networks. And then we can use representation learning and embedding techniques like node2vec and DeepWalk to

convert nodes into low-dimensional vectors, preserving the relationships for link prediction.

And these are some of the popular Python tools, libraries, and platforms for predictive modeling, like Deep Purpose. Although we have studied them in detail in earlier sessions, I just want to highlight some of them here. So we have the deep purpose we can use to do the end-to-end DTI prediction, and then we have DeepCAM where we can develop the ML model for drug design and property prediction as well. PyTorch, Geometric, or DGL, which is a graph neural network implementation. Scikit-learn, where we can use classical ML algorithms, BioBERT, which is an NLP-based relationship extraction tool, and ChemProp, which is a property prediction tool from SMILES, can all be used.

Okay, once we have done that, the next step is the hypothesis generation. So a hypothesis is a scientifically grounded assumption or prediction that explains how a drug compound or molecular interaction might influence a biological system or disease process. So it is typically based on network features, topological relevance, or AI-predicted associations. So the sources of hypothesis generation are the hub nodes identified via network centrality matrices, modular clusters and reached in disease pathways. predicted drug target or drug disease links overlap between drug interaction networks and disease modules and perturbation signatures from the omics data.

So, we can have you know multiple kind of hypothesis is for example, we can have target based hypothesis where we can consider that an exam and a protein X is a potential of target for drug Y possibly contributing to the anti inflammatory effect. We can have you know drug repurposing hypothesis as well where for example an antifungal drug shows high proximity to the neuroinflammatory genes and may be repurposed for the Alzheimer disease. We can have polypharmacology hypothesis for example a natural compound Z interacts with both kinase A and receptor B indicating potential for synergistic cancer inhibition. and we can have combination therapy hypothesis where for example combining drug A which is a PI3K inhibitor with drug B which is an mTOR modulator may effectively block parallel survival pathways in resistant tumors and then we can have you know pathway level hypotheses where for example activation of nrf2 pathway mediates the protective effect of a flavonoid rich extract against oxidative stress in the liver cells so ultimately even we have made a hypothesis and we have got like exciting results so there is no alternate to the experimental validation so once predictions have been made through computational network based approaches it is essential to validate these hypotheses through experimental or clinical methods so this step ensures that predicted drug target interactions disease associations or synergistic effects are biologically and therapeutically relevant So why validation is essential because it confirms the biological activity and specificity of predicted targets or pathways, eliminates false positives generated from in

silico modeling, provides credibility and transfer translatability of computational findings into real world therapeutics and supports preclinical development of new drugs or combinations. So, there are multiple approaches to experimental validation.

So, we can use the in vitro validation which can be where you can test the drug target interaction in controlled cell or biochemical system. So, some of the common assays like binding assays like SPR or ITC or fluorescence polarization can use the enzyme inhibition assays, cell viability and proliferation assays, reporter gene assays for pathway activation inhibition. And they can use the tools like ELISA, Western blot, QPCR, luciferase reporters, etc. And then you can perform in vivo validation as well where you can confirm efficacy, safety and pharmacological effect in the living organisms. And you can use the models like disease specific animal models like genograph model in cancer, knock-in knock-out mice model for gene validation.

Behavioral or histological analysis can be done, for example, related to neurological pathways or inflammatory pathways and then you can measure the outcomes like therapeutic efficacy target modulation toxicity or admet profile so once we validate our compounds by using in vitro and in vivo validation then only we can say that okay we have got the you know, we have successfully utilized the network pharmacology to either identify a new target for an existing drug that is repurposing or to identify, you know, like polypharmacology. For example, a drug is, single drug is acting on multiple targets. The next step is iterative refinement. So once we have got, so as I said, like this, this workflow is, you know, iterative.

So it's cyclical, not linear. So we get the validated data and new discoveries are then fed back into the pipeline to update the interaction database and retrain AI models with better ground truth through feedback mechanism. So and it improves the network construction and enrichment accuracy. So this ensures that the network pharmacology models evolve with emerging knowledge and technological advances because during you know the for example during validation we got some exciting data so that data can be then fed into the model in the beginning and then further it can be you know, you know upgraded and then we can get the outcome from that model. Okay, so then let us talk about some of the bottlenecks in network pharmacology and how AI can help in that. So some of the bottlenecks are like incomplete and sparse biological data.

So many interactions like drug target, gene disease are missing due to experimental limitations. And by using AI what we can do is we can use for example graph based deep learning, which can infer missing links in biological networks. And we can use like matrix factorization and embedding methods like node2vec and deepwalk which can predict unknown interactions from the known patterns as well. So, if there is like noisy or

conflicting interaction data set. So, where the data set contains false positive or negative or different sources often contradict each other.

So, in that case data deoising algorithms and semi supervised learning can filter unreliable edges and AI driven confidence scoring for integrating and weighing data from multiple sources can be used. And then we have like manual or outdated curation of pathways where the pathway databases rely on expert annotation which can be slow and inconsistent. So in that case we can use the NLP tools like BioBERT or GPT based models that can mine recent literature to update the pathways in near real time and AI can auto curate biological processes from millions of papers reducing reliance on manual annotations. and integration of heterogeneous data so combining diverse data types like omics phenotypic and chemical is complex due to different formats and semantics so in that case multimodal learning can jointly analyze the heterogeneous data sources and ontology aware ai frameworks can align and map concepts across biology chemistry and clinical domains and AI can transform raw or mixed data into network ready formats using feature extractions and dimensionality reduction as well. So these are you know the solutions which are provided by AI based on the bottlenecks which exist in the network pharmacology.

okay coming to the summary so the network pharmacology transforms drug discovery by targeting biological systems as interconnected network rather than the isolated molecule because as we say like from one drug one drug one disease one target hypothesis to we are moving to the you know a systems level approach so it it helps us to you know do the understand the systems level effect of the drugs on the body and it suppose multi target strategies suited for complex diseases especially like cancer and neurodegeneration and by integrating pharmacology systems biology and computational modeling it maps drug gene pathway interactions and AI enhances this approach through predictive modeling pattern recognition and integration of high dimensional data However, despite challenges in data quality and interpretation, AI-powered network pharmacology offers holistic, efficient, and personalized route to therapeutic innovation. So in the end, I have an open question for you in a world where AI can uncover hidden multi-target drug actions across complex disease networks. How might this reshape our definition of a successful drug? Should it treat symptoms, modulate network or rewire the disease biology itself? So just think over it. And I have some you know suggestions for further reading where you can go through these articles if you wanted to learn more about you know this topic And with that, thank you.