AI in Drug Discovery and Development

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

Week-04
Lecture-18

Welcome to the course on "AI in drug discovery and development." In today's session, we will talk about omics data integration for target discovery. So, by the end of this lecture, you will be able to learn the basics of omics and understand the role of omics in drug discovery. Explore AI-driven big data handling in omics, apply network pharmacology for target discovery, and also discover emerging trends in omics-based drug discovery. So, if we talk about omics, the term omics is derived from the Greek word 'ome,' meaning a complete set or whole body of knowledge. So, it was first used in a genome, which refers to the complete set of genes in an organism.

So, the suffix "omics" signifies a large-scale holistic approach to studying biological molecules. The omics discipline uses high-throughput technologies and computational tools to analyze biological systems comprehensively. So, here you can see the old types of omics data integration. So, here you can see that the genome, which has around 20,000 to 25,000 genes in our cells that comprise our genome, and then you have the genes; after transcription, you get the mRNA level.

So, it consists of mRNA, microRNA, and non-coding RNA. So, this comprises the transcriptome, which consists of more than a million of them in our cells. And then this transcriptomics after translation it leads to the formation of the protein's synthesis of the proteins and then that forms the proteomics. In our cell, we can have more than 1 million different types of proteins, which comprise the proteome. And then those protein molecules, due to the enzymatic reactions, are metabolized into different substances.

This comprises the metabolomics, where around 5000 metabolites have been identified in our cells, and those comprise the metabolome. So, the changes in the genes at the epigenome level, like DNA methylation or histone modification, occur. So, it comprises the epigenomics. So, let's talk about genomics. So, it is a study of an organism's entire genome, including its structure, function, evolution, and mapping, that helps in a thorough understanding of genetic variation, gene regulation, and their role in disease and evolution.

So, there are several techniques; for example, whole genome sequencing, which determines the complete DNA sequence of an organism. and whole exome sequencing, which focuses on the protein-coding regions of the genome, is called the exome. And then

there are genome-wide association studies called GWAS, which identify genetic variants linked to the diseases. And then there is a term called comparative genomics, which studies genomic differences between species for evolutionary insights. And if we talk about transcriptomics, it is the study of the entire set of RNA transcripts, or transcriptome, produced by the genome under specific conditions.

and provides insights into gene expression, patterns, regulation, and functional roles in health and disease. So, some of the techniques that are being used, for example, RNA sequencing, short-named RNA-seq, are high-throughput sequencing methods to quantify and analyze RNA molecules. And then microarrays, these are, you know, like the latest technologies where you can use hybridization-based techniques for measuring gene expression levels. And then you have the qRT-PCR, quantitative real-time polymerase chain reaction, which is a targeted approach to validate gene expression levels in a sample. So, if you are interested in evaluating the expression of a single or multiple genes.

So, you can do that with the help of qRT-PCR. And then, if we talk about epigenomics, it is the study of heritable changes in gene expression without altering the DNA sequences. So, key epigenetic modifications are DNA methylation, where a methyl group is added to the DNA, which leads to the silencing of gene expression. And then there are histone modifications like acetylation, methylation, and phosphorylation, which alter chromatin structures. And then there are non-coding RNAs that regulate gene expression post-transcriptionally.

So, some of the techniques that are being used to obtain this data are bisulfite sequencing, which detects the DNA methylation patterns. We have ChIP-seq, which is chromatin immunoprecipitation sequencing that studies histone modifications. And then we have the ATAC-seq, which is the assay for transposase-accessible chromatin and which can map the open chromatin regions. And then we have proteomics, which is a large-scale study of the entire set of proteins, also known as the proteome, in a biological system that focuses on protein expression, structure, function, interaction, and post-translational modification. So, some of the technologies that are being used for obtaining protein data are mass spectrometry, two-dimensional gel electrophoresis, protein microarrays, western blotting, and ELISA.

So, some of the tools that are being used are MaxQuant and Proteome Discoverer for protein identification and quantification, and Cytoscape and String for protein-protein interaction analysis. Some of the databases that contain this information are Uniprot, Pride, and Peptide Atlas, which include information about the protein sequences and the expression data. So, if we talk about the proteomics data collection pipeline, So, the first step is to isolate the protein, meaning extracting the protein and then digestion,

trypsinization, which leads to converting the long polypeptide chain into smaller fragments by breaking the peptide bonds in them. and then the separation of those peptides using LC-MS, 2D gel electrophoresis, followed by identification and quantification, and then bioinformatics analysis. So, this is how we can determine what kind of proteins are actually in a sample, and then we can compare them with the healthy control group.

So, to determine what kind of changes have occurred in a diseased person at the proteomics level. Then, we come to metabolomics. So, it involves the comprehensive study of metabolites, the small molecules less than 1500 Daltons involved in cellular processes, and provides insights into biochemical pathways, disease mechanisms, and drug metabolism. So, it helps in biomarker discovery, precision medicine, and systems biology. So, some of the techniques that are being used highly for metabolomics analysis are nuclear magnetic resonance spectroscopy, mass spectrometry with chromatography like GC-MS or LC-MS, and capillary electrophoresis mass spectrometry.

And then some of the tools that are being used are XC-MS metabo-analyst. And some of the databases, HMDB, which contains the Human Metabolome Database, KEGG, which contains the metabolic pathways, and METLIN, which contains the metabolic identification. So, what is multi-omics, actually? So, it is an integrative approach that combines data from multiple omics disciplines such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics to gain a comprehensive understanding of biological systems, disease mechanisms, and drug responses. So, it provides a holistic and systems-level view of cellular processes, enabling more precise drug target discovery, biomolecular identification, and personalized medicine. So, let us see why we need the multi-omics approach for drug target discovery.

So, you can see here all these, you know, omics technologies. So, what we can find is similar to metabolomics, which highlights the metabolic pathways altered in the disease state, revealing the enzyme-based drug targets. And then the proteomics identifies dysregulated proteins and post-translational modifications affecting the cellular pathways. Transcriptomics reveal the differential gene expression pattern across healthy and diseased tissues. And then epigenomics investigates DNA methylation, histone modification, and chromatin accessibility to find regulatory targets.

And genomics identifies genetic mutations, SNPs, and copy number variations linked to disease risk. So, altogether they can help us identify a suitable drug target for drug discovery and development. We will see an example of that as well as how multiomics data have been used to identify global drug targets. Okay, so if we compare single-omic versus multi-omic approaches for target discovery, the challenge is related to limited biological context. So, the single-omic challenge, which can analyze only one layer, genomics or

transcriptomics, provides an incomplete view of the disease mechanism.

While the multiomics approaches integrate multiple layers like genes, proteins, and metabolites for a holistic understanding, that is how our human body actually works. So, neither the organ system nor the cell is working in silos. So, they are actually working in harmony with each other, and the body acts as, you know, a kind of system. So, that is how these multiomics approaches can give us the systems level view. And that is how they can help us deal with drug discovery and development.

So, the single omics approach also provides incomplete mechanistic insight because, for example, genomics is only able to reveal mutations but does not indicate whether those mutations are leading to any functional impact or not. Likewise, the transcriptomics show gene expression but do not confirm whether the protein level changes are occurring in the cell or not. And then proteomics identifies proteins but lacks information about their regulatory mechanisms. So, whether it is like a temporary increase in protein concentration or a kind of permanent change that is leading to the disease pathophysiology. So, by using the multi-omic approach, we can overcome these limitations by connecting the genetic variation to functional changes at the transcript, protein, and metabolite levels, and then relate them to the difficulty in prioritizing drug targets.

So, the identified targets may lack validation across different biological layers, leading to a high failure rate in drug development if we use a single omic approach. While the multi-omics approach enables cross-validation by integrating genomic, proteomic, and metabolic evidence to prioritize clinically relevant targets. So, as well as there is another, you know, difference, which is the lack of personalized insights. The disease heterogeneity makes it difficult to identify patient-specific targets using only one data type. While the multi-omics approach can stratify patients based on molecular profiles, it enables a precision medicine approach as well.

So, if we look at the pipeline of multi-omics data integration analysis. So, the first step is data collection, where we collect different kinds of data; for example, genomics, proteomics, transcriptomics, epigenomics, and metabolomics data. And then the second step is preprocessing and dimensionality reduction. So, where we deal with the missing values, also called non-values, and then the next step is feature extraction. And then the feature selection using, for example, the principal component analysis or t-SNE technique.

So these are techniques to reduce dimensionality, meaning you have thousands of features, but which of those features are relevant to your analysis? So, you just extract those features and select those features by using these dimensionality reduction approaches. And then, when you integrate this multi-omics data, you can have multiple types of integration, such

as early integration, joint integration, late integration, and mixed integration. And then, using this multi-omics integration data, you can use it for classification, regression, or clustering. You can use it for making a model prediction, or you can use it for method evaluation, leading to interpretability and explainability. You can identify, for example, which of the drug targets is suitable for a disease or not.

So, some of the types of omics data integration can be of two kinds: one is horizontal integration, and the other is vertical integration. So, the horizontal integration combines data sets with the same type of variable. For example, gene expression or protein abundance across different study conditions or subjects enhances statistical power, improves reproducibility, and enables meta-analysis. For example, merging multiple transcriptomics datasets from different cohorts to identify a robust gene expression signature. While in the case of vertical integration, it integrates various types of omics data for the same biological sample or system.

So, this provides a multilayered understanding of biological processes by linking different molecular levels. For example, studying how genetic variation influences gene expression, protein levels, and metabolite changes in a disease model. So, some of the multiomics databases are like the Cancer Genome Atlas, known as TCGA, as well. So, it is a multiomics database with genomics, transcriptomics, epigenomics, and clinical data from over 11,000 patient samples across 33 cancer types. Managed by the National Cancer Institute and the National Human Genome Research Institute.

And then you have the Genotype-Tissue Expression (GTEx) program, which studies the genetic variants and their impact on gene expression across human tissues, including sex-based differences. And then you have the ProteomicsDB, which is a large-scale proteomics database with over 19,000 LC-MS/MS experiments for quantitative protein analysis. And then you have the MetaboLights, which is a global metabolomics repository containing raw data, metabolite structures, reference spectra, and biological roles across various species and techniques. So, how can we use AI in big data handling in omics? So, when we talk about big data. So, there are three attributes that stand out in defining the big data characteristics.

For example, the volume of data is significant. So, it is called 3 V, where the first V is the volume of data, the second V is the variety of data, and the third V is the velocity of data. AI is revolutionizing omics by efficiently handling high-dimensional biological data, enabling precision medicine and accelerating drug discovery. So, scalability is one of the biggest advantages of, you know, utilizing AI for it. So, it processes vast multi omics data set from next generation sequencing, mass spectrometry and single cell omics.

And then you can use it for data integration as well. So, AI-driven multiomics fusion integrates genomics, proteomics, and metabolomics for a holistic understanding of disease. And then you can use it for automated feature selection as well. So, the ML models extract meaningful patterns from thousands of gene proteins and metabolites. This is a summary of how you can use it for the integration of AI in multiomics data analysis.

So, for example, here you have cancer patients. So, you take all the data, multiomics data for those cancer patients, which can comprise genomics, transcriptomics, and proteomics. And then you use similar data from the healthy control group: the genomics, transcriptomics, and proteomics data. By using different AI algorithms, such as supervised machine learning methods, deep learning methods, or graph neural networks, for determining the biological networks. So, by using all these algorithms, what you can do is identify the subtypes; for example, they have some kind of similar characteristics.

So, you can segregate them, or you can cluster them into subtype 1, and then other patients have characteristics of cancer that are different from those in subtype 1. So, you can classify them into subtype 2 based on all this multi-omics data analysis. And then you can identify the therapeutic targets as well. So, you can identify targets that can be used to treat this disease. The third thing is that you can predict the disease prognosis and survival probability.

Okay, so now here I am giving you a wonderful example of what is known as Panda Omics. So, it is a tool developed by Insilico Medicine. So, it is a cloud-based software platform that applies AI in bioinformatics techniques to multimodal omics and biomedical text data for therapeutic target and biomarker discovery. So, it's developed by Insilico Medicine, a leading AI-driven biotechnology company specializing in drug discovery and aging research. So, it generates novel and repurposed therapeutic targets and biomarker hypotheses with the desired properties and is available through licensing or collaboration.

So, let us see how this Panda-omics works. So, it takes on the omics data set, searching for different kinds of multi-omics data, and then it processes and analyzes the data, followed by sample clustering and comparison creation. So, next is the gene and pathway level analysis, which further aggregates multiple comparisons into a meta-analysis. And then it also takes advantage of the prior knowledge and trends extracted from the publications, grants, and clinical trials. So, all the data is also fed into this model, using this information as well. So, it leads to the identification of the therapeutic target and biomarker discovery.

Once we have identified a target, the next step is to identify the disease-relevant active compound, which is also known as hit identification. And then indicate prioritization and

expansion, determining the biological knowledge graph. And then it uses GPT to interpret the results, and then it has an automated lab called a robotic lab. So, which can synthesize molecules, and those molecules have been, you know, designed by using Chemistry 42, which is a generative AI model. So, it can it can generate it is a de novo drug design model based on generative AI.

So, this Chemistry 42 can design molecules. So, it can design small molecules with chemistry 42, and then those molecules can be used for engaging this target that has been identified. So, InSilico medicine has identified several drug targets in its pipeline by using this multi-omics approach. So, some of the other related AI tools in omics are like SOPHIE, which separates common and specific transcriptional responses using a generative neural network; you can go through this link for more details about that. And then you have the scEMAIL, which is a universal and source-free generator for single-cell RNA-seq data with novel cell type perception. And then you have the TIST, which can analyze transcriptome data and histopathological images integratively for spatial transcriptomics.

And then you have TripletGO, which predicts gene function by integrating transcript expression profiles with protein homology inferences. And then you have the DrSim, which enables transcriptional phenotypic drug discovery by similarity learning. And the link for all these tools has been given here. So, you can go through them in detail if you are interested. Okay, so then there is another term called network pharmacology.

So, which is a computational and systems biology approach that studies the complex interactions between drugs, targets, diseases, and biological networks rather than just relying on the single target or single drug paradigm. So, some applications of network pharmacology include drug repurposing, polypharmacology, precision medicine, and toxicity prediction. So, if you look at the flow of data in network pharmacology. So, we start with the omics data acquisition and pre-processing, and identify the key genes from the omics data. And then, by using those genes, you can identify the functional pathways in which those genes are involved, and then you can see the enrichment analysis as well.

And then after that, you can make a drug target mapping and prioritization, followed by experimental validation and systems biology integration. So, then we come to the omics data acquisition pre-processing. So, we can identify disease-associated genes, proteins, and metabolites from high-throughput omics experiments. So, the data sources, as I said earlier, include genomics, such as WGS (whole genome sequencing) and genome-wide association studies. An example of using GWAS in Alzheimer's disease is that APOE-E4 was identified as a genetic risk factor in Alzheimer's disease using WGS studies.

And then you have transcriptomics like RNA-seq or microarray, which contain the

differentially expressed genes. For example, in AD, the upregulation of GSK3 beta and CDK5 involved in tau hyperphosphorylation was observed using transcriptomic analysis. And then comes the proteomics, where we use mass spectrometry or protein expression profiling. So, in Alzheimer's disease, the increased levels of beta amyloid and tau proteins can be identified using proteomics analysis. And then the metabolomics, like LC-MS and GC-MS, can detect the altered glucose metabolism, which is a hallmark of AD, through the metabolomics.

So, if we talk about the AI's role in multiomics data integration and analysis, it can help in feature selection and dimensionality reduction to extract the significant markers. So, how does one construct biological networks? So, we can establish the relationship between identified genes, proteins, and disease-relevant molecular networks. So, those network types can be protein-protein interaction networks, which identify the direct and indirect interactions between proteins using STRING or BioGRID datasets. And then you have the gene regulatory network that identifies the transcription factors regulating disease genes. And then you have the metabolic networks that identify biochemical pathways involved in disease progression using KEGG or Reactome.

So, the role of AI in constructing the biological network is that you can use a graph neural net, which can predict unknown interactions in the PPI networks. And the AI clustering algorithms, like k-means or spectral clustering, can group functionally related proteins. So, once the network is constructed, we have to identify the key pathways that are involved in the disease and analyze them. So, for that we use enrichment analysis and pathway mapping. So, the enrichment analysis is done by identifying the biological processes and molecular functions that are enriched in the diseased genes.

For example, in AD, this pathway includes oxidative stress response and neuroinflammation. So, some of the tools that can be used are DAVID, Metascape, and Gene Ontology Analysis. And then for pathway mapping. So, what we can do is map the disease-associated genes to known biological pathways, like the PI3K-AKT pathway, which is crucial for neuronal survival and is dysregulated in AD. So, you can use tools such as the Kyoto Encyclopedia of Genes and Genomes, KEGG, or Reactome for that.

So, we can also identify the drug targets via network pharmacology. So, it can identify druggable targets and potential therapeutics using network-based approaches. So, the network-based approaches contain drug target interaction prediction where we can use AI models like Deep Purpose or Deep Affinity, which identify interactions between other drugs and the targets. And then you have network-based drug repurposing, which identifies existing drugs that modulate key targets. For example, the repositioning of metformin or pioglitazone for Alzheimer's disease. And then you can do multi-target drug discovery,

which identifies drugs that act on multiple network nodes for better efficiency as well.

So, once we have identified those targets using that pathway analysis. So, the next step is to validate those AI-generated targets using in vitro and in vivo experiments. So, some of the techniques that we can use, which we have seen earlier as well in drug discovery and development, like gene knockdown and CRISPR studies, RNAi, siRNA, or shRNA, and CRISPR-Cas9, confirm gene function in disease models. And then we can use the in-silico studies, such as molecular docking, molecular dynamic simulations, MMPBSA, and MMGBSA calculations, which ensure binding stability and affinity. And then we can use in vitro studies like cell viability studies (MTT, XTT, or Alamar Blue assay), western blot, TURT, PCR, immunofluorescence, confocal microscopy, flow cytometry, ELISA, etc.

And then we can do the in vivo validation as well, where we need to produce a disease-specific animal model and integrate multi-omics, including proteomics or metabolomics, to correlate with the AI-predicted drug target. I want to summarize, and we want to just look at the emerging trends in omics-based target discovery. So, there are techniques called spatial omics that are being developed nowadays. So, which can analyze the biological molecules in their native tissue locations, providing specially resolved molecular insights. You also have the single-cell multi-omics data, which integrates DNA, RNA, and protein data at the single-cell level, enabling the identification of rare cell types.

So, because now you are identifying and getting all those multi-omics data from single cells. So, we can identify the rare cell type, do the cellular diversity mapping, trace the lineage, and understand the cross-omics regulatory mechanisms as well. So, these advancements they enhance precision medicine by uncovering disease mechanisms at unprecedented resolution. So, summarizing this session: So, the omics technologies play a crucial role in drug target discovery by integrating genomics, transcriptomics, proteomics, and metabolomics. So, the traditional single omic approach it has limitations making multi omics integration essential for improving target selection accuracy.

And then various computational tools, network pharmacology approaches, and AI-driven platforms help analyze omics data for drug target discovery. However, there are some challenges, like data standardization and cost, which can be addressed by cloud computing and using AI. And then there are some emerging trends, which include spatial omics and single-cell multi-omics. These are shaping the future of precision medicine and therapeutic developments.

Then I have a question for you to think about. If a novel disease emerged with no known genetic, proteomic, or metabolomic markers, how would you design an omic-driven strategy to predict and validate potential drug targets? So, what challenges might arise, and

how could AI and machine learning assist in overcoming them? And you can go through all these references to get more details about this topic. So, you have, for example, the first paper, which is about Pandaomics, an AI-driven platform for therapeutic target and biomolecular discovery. And with that, thank you very much.