AI in Drug Discovery and Development

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

Week-02
Lecture-07

Welcome to the course "AI in Drug Discovery and Development." In today's session we will have a look at the overview of AI technologies. So, by the end of this lecture you will be able to understand the basic concept and components of artificial intelligence. Explore key AI technologies such as machine learning, deep learning, and natural language processing. Learn about the types of AI systems and their core functionalities and also build a foundation of understanding how AI technologies work. So, as we have seen earlier as well that AI is a field of science concerned with building computers and machines that can reason, learn and act in such a way that would normally require human intelligence or that involve data whose scale exceeds what humans can analyze.

So, it has always been a dream for humans to develop such a machine because humans are considered as the most intelligent animal in this universe actually, at least on the earth. Because we don't know about the you know, there might be some aliens on some some planet which is like very very far from us. But at least we are the most intelligent animals on on earth and then what if we could develop a machine or an algorithm which can Think and act like us. Which can you know analyse the data and then take a decision on the basis of the analysis of data.

So that was always the dream of humans actually. And then it has been, you know, it has led to the development of AI, I would say. So, it is a broad field that encompasses many different disciplines, including computer science. data analytics and statistics, hardware and software engineering, linguistics, neuroscience, and even philosophy and psychology. So, the key features of AI are the ability to process vast amount of data and autonomous decision making and continuous learning and adaptation.

And some of the examples we are using AI in our daily life is like we are using the virtual assistants like Siri or Alexa. We are getting the recommendations on Netflix or YouTube. Or if we are using the Google Maps, we are getting the shortest route to destination A to destination B. Or we are getting the weather forecasting. So there are hundreds of examples where we are using AI every day.

So, if you look at the types of AI, so we can differentiate AI on the basis of their capability into three types. One is called as the narrow AI, which is also known as the weak AI. And

these weak AIs are, they are specifically designed for a single task actually or for a specific task. So, if we talk about like the drug discovery applications where these narrow AI are being used. for example, we can, when we are performing the virtual screening, so that AI model can predict which molecules are likely to bind to a target.

And that we can do with, you know, for example, either, you know, using the structure-based design like molecular docking, using auto-dock tool with the enhancements with the AI. Like you have the G-NINA or you have S-MINA or you can use the Schrodinger tools for that. Or you can use the QSAR, which is a ligand-based screening method where you build models to predict activity from chemical structure. So, those models will be specifically suitable for this specific purpose. So, if we are developing it for molecular docking that cannot be used for the other purpose.

And then we have, for example, the AlphaFold, which is another example of narrow AI, which is used for predicting the protein structure. So, that can predict the protein three dimensional structure with very high accuracy by just using the sequence of the protein. So, then we have the general AI or that is also known as the strong AI. So, this AI is considered as having human like intelligence. So, it can understand, it can learn and it can apply knowledge across a wide range of tasks like a human would do.

So, this is to be honest this is not a reality yet, but maybe very soon we will see the development of general AI agents. And then this can be in, if you look at the drug discovery application or tasks which can be done by using general AI. So this could be end-to-end discovery from predicting novel targets to designing molecules, optimizing synthetic routes and planning clinical trials, all with minimal human intervention. And there are companies, for example, InSilico medicine, which have been doing excellent in this area. So they have been developing all those end to end kind of platforms for discovery of new drugs.

And then adaptive drug design can also be one example where general AI can be used. So, that can be an AI that can autonomously shift between different types of data like genomics, structure, patient data to guide the personalized drug design. And then we have this super intelligent AI. So that is absolutely not feasible yet. So if we can develop this super intelligent AI, so that is a kind of a futuristic thing.

So, which surpasses human intelligence in all the fields. So, if we talk about the application in drug discovery and development, so it can revolutionize biomedical research, it could hypothesize new biological mechanisms. Predict side effects before clinical trials and even invent new classes of drugs beyond human imaginations. So, just imagine if we can develop a super intelligent AI, so how easy the drug discovery and development could

become. Okay, so when we talk about the core components of AI, so it is basically these are three pillars of AI.

The first pillar is the data, which is the foundation for training AI models. And then this data could be in the structured format, like the spreadsheets, where you have the, everything is arranged in rows and columns. And it could be unstructured as well, like it could be the images from, you know, histopathology or the X-ray images or, you know, MRI images or the videos actually. And then the second pillar of AI is the algorithms. So, these are the logics or rules that guide an AI system to learn and make decisions.

And the third and very important pillar is the computing power. So, the high performance hardwares like GPU and TPUs for processing complex algorithms. So, we will have a look at them one by one. So, if we talk about the data which is absolutely the foundation of AI in Drug Discovery and if we talk about the data being used in drug discovery by all those AI models. So, it could be the chemical structure where the chemical structure can be represented as molecular structure or it can be represented as the SMILES string.

So, SMILES is a 1D representation of chemical structure and then it can also be the compound libraries which can be in different forms actually. So, you have like different formats where you can represent all those chemical structures in the form of libraries. And then another kind of data is the biological data. So, that could be the protein sequences, the binding site information or the genomics data or even the multiomics data where we have the data from proteomics, epigenomics, metabolomics, lipidomics, etc. So, all that data is coming from the biological source actually.

And then we have the pharmacological data. So, that could be the ADMET properties. For example, when a drug is going to go into the human body or animal body. So, that is interacting with the body. So, how the drug is interacting with the body and how the body is interacting with the drug.

So, all those kinds of data can be also be used for training all those models like the data could be related to the absorption of drug, distribution of the drug, metabolism of the drug, excretion of the drug. or related to the toxicity of the drug, or it could be the biological activity data or bioactivity data, which is when that drug molecule binds to its desired drug target. So, for example, it can be an enzyme, it can be a receptor, it can be an ion channel. So, when it binds to the drug target, so what kind of activity it leads to, so that is called as bioactivity data. So, then it can, the data could also be like clinical trial outcomes as well.

And then there can be data from the literature and patents as well because every research data is usually published either in the form of the research article or the patent actually. So,

that data is also can be used for, you know, training all those models. So, some of the challenges in using the data is the data sparsity, of course, these challenges we will talk about later in detail actually. So, the data is coming from the you know, for example, for some rare diseases, we have very limited data. And then noise and error, the data is inconsistent coming from different assays across different labs across different you know, continents.

If we combine all that data, we do not know what kind of errors it is bringing into the data. And then the multimodal data integration, that is again a challenge. Can we combine chemical, biological and clinical data sets? So, all the challenges as I said like we will talk them about later. And then the second pillar is the algorithms. So, the algorithms are the step-by-step instructions that AI uses to learn patterns, make prediction and improve over time.

So, in drug discovery, usually we use these machine learning techniques like supervised learning, which learns from labelled data, where we can use, for example, predicting IC_{50} values from chemical structure. Or we can use unsupervised learning which finds the hidden pattern in unlabelled data like clustering similar compounds to identify potential lead compounds. Or reinforcement learning where we can learn through the trial and error maximizing rewards like we use it for the you know like REINVENT the AstraZeneca molecular AI tool for the de novo molecular generation. And of course the deep learning which uses neural nets to handle complex unstructured data like it predicts protein, it can predict protein ligand binding, generate new molecular structures or analyze the biomedical images as well. And then the third pillar of AI is the computing power.

So, for training and running those AI models, especially in drug discovery, it requires enormous computational resources. And the key components, they include CPUs, the central processing units, which handles general tasks efficiently, and also being used for basic data processing, lightweight model training, etcetera. And then we have GPUs, which are nowadays becoming very important for training all those models, especially those deep learning models. So, they can handle parallel computation, and they can speed up the deep learning tasks, and thus suitable for training large neural nets for tasks like molecular docking prediction or image analysis.

And then we have the TPU. These are specifically designed by Google for deep learning tasks. So, they can do faster matrix multiplication. And then we have the cloud computing like the Amazon Web Services or Google Cloud or Azure. So, the idea is that you don't need to set up your own hardware stuff at your own place. So, instead you can use it from the cloud or the web.

So, it's ideal for start-ups and labs without on-site hardwares and it can be used for running large-scale virtual screening with millions of compounds. So, now there are services which you can use to screen millions of compounds, even billions of compounds in very short time without using your own hardware just in the cloud actually. So coming to the question, why do we need AI in drug discovery or drug discovery and development? So if you look at this data, so we can see that at every step, we are adding 2 into the previous number. So the next number, after 8, everyone can guess that it will be 10, actually. Because that was very easy, it was very easy to identify the pattern that every time we are, you know, adding 2 into the previous number and we are getting 10 in the end.

But can we guess which of those compounds can cross the blood brain barrier? Even for, you know, a chemist, so it can be a difficult problem until unless that chemist have been working for maybe 30 years in the CNS drug discovery. And have seen these structures before so he or she can just tell but maybe I don't know with 100% accuracy or not but is it so that we are not able to tell which of those molecules can cross blood brain barrier. And the reason is that we are not able to see the hidden pattern in this data and that is where we can use the AI or machine learning because they are faster, they are quicker and then they can process a large amount of data. So, if we use, if we visualize this data and then try to develop a model so that model can predict molecules which can tell us which molecule can have a better chance of crossing the blood brain barrier. And this is why we usually require all those fancy stuff like AI and machine learning in drug discovery and development.

Okay, coming to the machine learning. So, what we are doing here is we are like using the known data to develop a model to predict the unknown data. And then here the known data is it can be the big enough archive or the previous observations or the past data. And the model is you know the known data plus the algorithm the machine learning algorithms. And then unknown data here is the missing data, unseen data or not existing or the future data. For example, we wanted to predict the solubility of chemicals.

So, we need to have the experimental solubility data for a large number of chemicals so that we can use that known data and then we make a model out of that known data. And then with the help of that model, we can predict solubility of a new compound. So, that is we can say the unseen or missing data or not existing data so that we can predict with the help of that. So coming to the core concept of AI. So, AI is an umbrella term that encompasses all techniques that allow machine to mimic human intelligence.

So, it include both rule-based system, those traditional programming expert system and learning-based models like ML/DL. So, it using known data to develop a model to predict the unknown data and this we have already seen in the last slide as well. So, you can see

here, for example, we have this large circle and everything here is the artificial intelligence, okay. And the original idea is just to mimic the human intelligence. Can we mimic the human intelligence? So, okay, so then a machine learning, which is a specific approach within AI that enables machine to learn from data without explicitly programmed.

So, these are not the you know, the not the rule-based system. So, they can they can function a little bit independently. And then it is a subset of artificial intelligence. So, it focuses on creating algorithms that improve with experience rather than relying on predefined rules. And then you can see from here that all the ML techniques are AI, but not all the AI is ML, right? Because ML is a subset of AI.

And then if we go further deeper, so we see the deep learning, which is a specialized field of ML that uses artificial neural networks to process large amount of data and recognize complex patterns. So, the idea of development of artificial neural network was to mimic the activity of human brain. So, like how the human brains is working? Can we develop an algorithm which is mimicking the human brain activity? Like in our brain, we have neurons. So, and that one neuron is connected to several other neurons. So, likewise, we develop those, you know, perceptrons and then make this artificial neural net.

So, it is particularly useful for handling images, speech, and unstructured data. And then we can see here all the DL is ML and AI, but not all the ML is DL. So again, DL is a subset of ML and AI. So, you can see now the difference between all these three actually. Okay coming to machine learning so as I said like it is subset of AI where machine learn patterns from the data to make predictions or decision without explicit programming.

So, here you can see a kid actually that is playing with you know cars and trucks. So, what the kid is doing is it is trying to learn the features it is trying to learn by seeing by touching and by exploring okay. So, it is learning the features and the shapes and color of the cars and trucks. We can see here there are two toys.

One is basically a car; another one is truck. So, once the kid is learning, so then he will be able to differentiate between them. Okay, so let us let us come to these types and then we will discuss it as well. So, the type is the types of ML is like supervised learning where the machine is learning from the label data. If this kid is not knowing what is what like if it is not knowing that this is a car or this is a truck and if we as a parent are telling it that okay this is a truck and then this is a car. So, then we have labeled the data means we have already told the kid that this is a car and this is a truck so that he can learn the features of car and features of the truck and then he can remember it right.

So, likewise in drug discovery you can have like you can predict the solubility of compound

based on the experimental data. So, we have labeled data means the data with the experimental values. And then we can use the data to train a model and that will be called a supervised learning. But if we talk about unsupervised learning, so here the model is learning without any intervention actually. So, for example, again, now this kid is learning by seeing, touching and exploring his toys, right? So, now if we do not tell him this is a toy, this is a car and this is a truck.

So, he automatically learns based on their shape actually. For example, the truck is like a bit larger than a car, truck has this open, you know, kind of a carrier in the back while car is like a closed, okay. So, now the kid is learning without your intervention from the parents. So, that is called as unsupervised learning. So, in this case, what the kid can be doing is he or she can cluster or segregate cars on one side and trucks on one side based on the features which he or she has learned.

So likewise, the customer segmentation marketing is done by using unsupervised learning. Or we can say like clustering of molecules based on their structural features can be done by unsupervised learning in drug discovery and development. And then we have reinforcement learning, where it is learning by interacting with an environment to maximize rewards. For example, now we say that, okay, now the kid is learning, but then we ask, then we give it a new toy and ask whether is it a car or a truck. And if the kid is telling that it is a truck while it is a car, so then he or she is making wrong prediction.

So, then we will give a punishment to the model. Like we will say that, okay, we will not go for playing in the park today so, that will be you know a kind of punishment. And then on the other hand if he or she is predicting it correctly then we will give him a him or her a reward actually like a chocolate or may be outing to a Disneyland or something like that. So, that will be that will be you know based on the reward and punishment. So, that is called as reinforcement learning. So, here we can take an example of training robots to navigate obstacles so that they are you know the reinforcement learning can come in come handy.

So, coming to deep learning, so it is a subset of ML that uses artificial neural network modeled after the human brain to analyze complex data. So, as I said like it the deal is the idea is that you mimic how the brain works. So, you have the multiple layer of neural network which extract increasingly complex features from data. And then the application in the drug discovery and development we can see is like image recognition like detecting cancer using mammographic images or speech recognition converting spoken words into the text. Then these neural networks which are being used in deep learning, so they can be of different types actually.

They can be Multi-layer Perceptron which is a fully connected network used for general tasks such as classification regression. They can be convolutional neural network which are specialized for image and video analysis. Then we can have the recurrent neural nets which are designed for sequential data like from the time series or natural language processing. So, coming to the key AI technologies, so right now we have been seeing you know a lot of developments in AI. So, the most of the development is happening in natural language processing and computer vision and generative AI of course.

So, we will have a look at all these key technologies like NLP, computer vision, reinforcement learning, generative AI, expert system, knowledge representation reasoning and autonomous system. Of course, we will have a look at them from the perspective of drug discovery and development. So, the NLP it enables machine to understand, interpret and generate human like human language converting unstructured text into the meaningful data. We have seen that the example of using NLP in for example, a translator or speech to text recognition software or you know, how to say signage to text as well. So, but how can we use them in drug discovery and development is like they can play a key role in literature mining which can extract insights from scientific papers, patents and clinical trial reports.

They can be used in drug target discovery as well, where they can identify relationships between disease proteins and compounds. They can be used in the clinical data analysis as well, processing electronic health records to uncover patient insights. And they can be used in the case of adverse event detection as well, where they can identify side effects from medical reports and trial data. And then for knowledge graph building as well, where they can map the connections between biological entities for hypothesis generation. And then we have the computer vision, which enables machines to interpret and analyze visual data, which can be in the form of images, video, etc. to extract meaningful information.

So, the key functions of computer vision in the drug discovery and development can be protein structure analysis, where they can identify binding sites and molecular conformations from cryo-electron microscopy images. Or the histopathological image analysis where they can detect cancerous cells and tissue abnormalities. Drug manufacturing monitoring where they can ensure the quality control through real-time image inspection. And the phenotypic screening where they can analyze cellular images to assess drug effects.

And then we have the reinforcement learning. So, that learning paradigm where agent learns optimal actions through trial and error maximizing cumulative rewards. So, the key functions where it can play important role in you know AI drug discovery and development are the molecular generation which where it can help in designing novel compounds by rewarding desirable chemical properties. Protein folding prediction optimizes folding

pathways based on stability rewards. Clinical trial optimization where it can simulate patient responses to design efficient trial protocols. And personalized medicine where it can adapt the treatment strategies based on patient specific data.

Then we have the GenAI, we generate new realistic data. It can be text, image, molecules, etcetera, by learning from the existing data. And in drug discovery and development, we can use it for de novo drug design, like the tool REINVENT, I discussed earlier which proposes new chemical structures with desired properties. And then protein sequence generation which can be it can design synthetic proteins for therapeutic targets. And then we can also predict the adverse effects by simulating molecule interactions to anticipate side effects. And molecular docking enhancements we can generate improved ligand conformations for docking studies.

Then we have the expert systems which are basically rule-based system that mimic human expert decision making in specific domains. So if, we can use it for drug repurposing where it can recommend alternative uses for existing drugs based on mechanistic rules. Or clinical decision support which where it can assist doctors by suggesting diagnosis and treatment or toxicity prediction it can applies biochemical rules to flag potential toxic compounds. And formulation optimization where it can recommend excipients and composition for drug delivery systems. And then we have the knowledge representation and reasoning where the structures restructures the data in a way that machine can understand, draw inference, and make logical decision.
So, in drug discovery and development, we can use it for like biological knowledge graph where we can connect the diseases, genes, and drugs for hypothesis generation. We can do the pathway analysis where we can map the metabolic and signaling pathways for drug mechanism insights. We can do the synthesis route planning, design optimal synthesis pathways for novel compounds. We can do the understand the disease mechanism by integrating multiomics data to model the disease progression.

And then the autonomous system. So, these are a little bit futuristic. So, these are self operating system that perceive, decide and act independently in dynamic environments. So, it is like the self driving cars actually. So, the key functions it can play in drug discovery and development are the robotic laboratory automation where we can conduct high throughput synthesis and screening autonomously. And as I said like InSilico Medicine is one of the companies which is extensively using it. Of course, there are other companies as well which have been using automated robotic laboratories.

And then you have the adaptive drug formulation system which adjusts drug composition based on real-time feedback. And then you can use it for pharmaceutical supply chain optimization, where you can enhance the logistics for raw material and drugs. And then

you can use it for personalized drug delivery devices, which can tailor dosing regimen using real-time patient data. Okay, so coming to the summary, so AI, it encompasses technologies like machine learning, deep learning and NLP that mimic human intelligence. And the core components, it includes data, algorithm and computing power because without all these three pillars, it is, you know, we cannot use AI for, you know, our benefit actually.

And then a neural network, they form the backbone of many advanced applications like image recognition and autonomous systems. Okay, in the end, I have, you know, open question for you. So, what makes AI intelligent? Is it the data, the algorithm or the computing power or a combination of all three and why it is so? And you can go through these, you know, articles or websites to get more information about this topic. And with that, thank you.