AI in Drug Discovery and Development

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

Week-02
Lecture-06

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about history and evolution of AI in Drug Discovery. So, by the end of this lecture, you will be able to understand the timeline of AI advancements in Drug Discovery. Recognize key technological breakthroughs from 1950 to 2024, explore how AI evolved from rule-based system to deep learning and generative models. And also identify emerging technologies shaping the future of AI in Drug Discovery. So, if you look at AI, so 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, basically artificial intelligence is a kind of a technique where we are trying to develop a machine or an algorithm so that that algorithm can think and act like humans. So, it uses algorithms and vast amount of data to recognize patterns, make decisions and improve their performance over time. So, the early idea of these machines functioning on their own, it dates back to the ancient time with inventors creating mechanical automaton. So, those were called as automaton and it is a Greek word that means acting of one's own will.

So, those automaton were you know the earliest idea where those machines they can act on their own without any human intervention. So, the early examples include a mechanical pigeon from 400 BCE created by a friend of Plato and also the Leonardo da Vinci's automaton knight around 1495. So, these early inventions introduce the idea significant progress towards modern AI. So, these early you know inventions they introduce the idea that we can have such machines which can function on their own, but the actual you know the development or we can say the modern AI that began in 20th century with advancements from engineers and scientists. So, if you look at the brief history of, you know, AI, so around 1950, in 1950s, so the, you know, the Alan Turing, so he proposed the Turing test and that was considered as a, you know, birth of AI actually.

So, that was a test to evaluate if a machine is able to exhibit human-like intelligence through conversation. And then in 1956, John McCarthy, he coined the term artificial intelligence at Dartmouth conference. So, some of the early AI programs, those were, for example, the logic theorist. It was designed in 1955 by Newell and Simon. It was designed to mimic human reasoning and prove mathematical theorems.

And then there was another tool called a general problem solver GPS, which was developed in 1957. It was aimed to solve a wide range of problems using a step-by-step approach like humans. So, the idea of you knows, developing something, some algorithm or machine which can act as humans is like you know, it's always you know, dream to make such a machine or such a tool, but now in the 20-21st century, it has come true actually. So, if you look at the other early concepts of AI, so there was this, you know, in 1964, there was an early AI system named as STUDENT, which was developed by Daniel Bobrow. So, STUDENT was one of the first natural language processing programs.

So it could solve algebraic word problems by interpreting and converting English sentences into equations. And then there was ELIZA in 1966 which was created by Joseph Weizenbaum. So, ELIZA was an early chatbot designed to simulate human conversation. So, one could converse it with ELIZA like, you know, nowadays you converse with, you know, all those AI chatbots. So, you could do that with the ELIZA as well.

Of course, it was not capable as much as the today's chatbots. So, the AI initially, it relied on if and then rules to simulate human decision making, showing that rule-based AI could simulate human-like interactions. And then expert systems around 70s to 80s are like MYCIN, which is like made for medical diagnosis, and DENDRAL, which was for chemical analysis. So, they use predefined rules to solve the domain-specific problems. So there were, you know, the domain-specific tools were also being developed during that time as well.

So, the AI has experienced alternating phases of enthusiasm. So, those are named as like booms and disillusionments as well like bursts due to limitations in technology and funding. So, if you look at the timeline of major boom and bust cycle. So, the history of AI includes significant boom in 1956 that you can see here in this you know diagram as well. So, in 1956 we had a major boom with the development of heuristic search method.

And then in 1980 we had another boom which was due to the expert system development and in 2012 we had another boom because of the development in the deep learning stuff. And then it had like winters as well like in 1974, 1987. So those we will have a quick look at those actually. So, the first boom came in 1956 that was due to the development of heuristic search and logical reasoning. So, the early AI focused on problem solving using logical expression and search algorithms leading to optimism in their potential.

But then it saw a winter, the first AI winter came in 1974 where the early AI research faced setbacks due to limited computing power and unrealistic expectations. So governments and investors they cut funding as AI system failed to deliver on promises. So they were like huge expectations from those early AI systems but then they did not meet the expectation

because of the limited computational power available and the limited development in the algorit hms as well. So, that is why there was a lot of investment cut and then it saw the first AI winter in 1974. But then the resurgence happened in 1980 through the revival through expert system and early neural network led to the AI gain momentum with rule based expert system and the emergence of neural network for pattern recognition.

And then companies they adopted AI for industrial and commercial use leading to renewed funding and then it saw a boom in that time. So, the second winter was witnessed in 1987. So, that was you know due to the challenges in scaling AI. So, the limitations of expert system including their rigidity and cost caused another decline in AI research and investment. So, those AI models they struggled with real world adaptability and funding cuts from government and corporations further slowed down the AI research and advancements.

So, the limitation including the rigidity and cost caused another decline in AI research and investment and that was you know the reason behind the second AI winter in 1987. So, here you can see the overall summary of the AI revolutions. On the x-axis you can see the year-wise development in the AI technologies and on the y-axis, you see the magnitude of model parameters. So, you can see that those early AI methods like in 50s, 60s, 70s. So those were largely ruled by robotics, rule-based systems, search algorithms and symbolic reasoning.

And those magnitude of model parameter was not as much high actually as of now. So, then these in around 1980s, so there were developments happened in the ML field where the Bayesian network, gradient boosting, random forest, regression, support vector machines, so those all those technologies were developed. And then later on the deep learning saw a surge around after 2010. So the autoencoders, convolutional neural network, generative adversarial network, graph neural network and recurrent neural network. So those technologies like those artificial neural nets, they were developed and they were emerging during this era.

The model parameters were at the scale of 10 to the power 4 actually. And then around 2017, we saw a lot of developments in the transformers like BERT, GPT-T5 and VIT. And then around 2020, so we saw the development of large language models. And we saw the development of ChatGPT, Cloud, Gemini, Lama. And you could see that the magnitudes of model parameters were around 10 raised to the power 12 actually.

So those models have been trained on tremendous amount of data actually. So, if we see the evolution of AI in drug discovery around somewhere around 50s to 80s. So, around 50s the computational chemistry models were emerging actually and then in 60s it was largely

dominated by the quantitative structure activity relationship models. And in 70s the molecular mechanics and energy minimization techniques were developed and in 80s rule based expert system for drug property prediction were used largely. So, we look at the quantitative structure activity relationship.

So, the original idea of quantitative structure activity relationship was that can we correlate the structure of molecule with some kind of bioactivity or the or some property of these those molecules. And by correlating the structure with the bioactivity, can we use that correlation model, can we use that developed model to predict the bioactivity of new molecules by just calculating their features or calculating their descriptors. So, the key pioneers were, you know, the Corwin Hansch and P. Yates. So, they significantly contributed to QSAR advancements through their research and methodologies.

And the initial QSAR applications were focusing on predicting the pharmacological effect of chemical compounds in medicine and agriculture. So, we can see here for example, in 60s, so the Hansch developed this very famous Hansch equation. So, where property biological activity we can say with the parameters like pi which is which is called which is indicating lipophilicity and sigma which is indicating the electronic parameter. And then in 70s, there was integration of steric factors as well, the Taft's steric parameter. And most of the times, the multiple linear regression was used for, you know, predicting all those parameters.

$$\log(1/C)=a\pi+b\sigma+c$$

So, here we have the log 1 by C. So, where 1 by C is the C is the biological activity. And then this C can be correlated with the lipophilicity and electronic parameters by using this linear regression. Between 70s and 80s there were a lot of expansion and application of QSAR models happened. So, like the incorporation of steric factors like a Taft's steric parameter or introduction of three-dimensional QSAR, comparative molecular field analysis, comparative molecular similarity analysis.

So those were, you know, developed and then development of even non-linear QSAR models using the neural networks or principal component analysis that was also being explored actually. And then side by side, the data handling was also improved. So, there was enhanced computational power for data analysis and adoption of multiple linear regression was used as standard tool. And those early use of machine learning in QSAR prediction was also adopted during that time. So the 1990s saw a rise of the computational chemistry because during that time, all those machine learning models were available like SVM decision trees.

And those were then implemented in bioinformatics or cheminformatics or computational drug discovery, we can say. As well as the development of docking algorithms for structure-based drug design also happened. And the first hybrid model combining physics-based and statistical approaches was also being developed during that time. And the high throughput screening data, it was becoming widely available. And with the help of this high throughput screening data, one could develop all those fancy QSAR models on large training data sets.

And then they can use that for predicting the properties of new molecules. So further, the machine learning technique was applied to the chemical library profiling. And also the early predictive models for ADMET properties were also being developed during that time. Okay, so if we go back to the developments in the technology for say, so the early machine learning models were developed during the period of 2000 to 2010. Where it saw the development of SVM which was largely used in drug discovery for classification of bioactivity data.

And then the random forest which was improving the handling of chemical data sets. And then Bayesian networks were also developed where they could they could model the model the drug target interactions. And then hidden Markov models were also developed where the early sequence analysis of protein structure prediction could be done by using those models. So, the 2012 so, you know a breakthrough that was the deep learning breakthrough and here you can see that this was a result from the competition. Where the AlexNet it wins the ImageNet competition showing better performance than traditional methods.

So, the outcome was that so, there was like a lot of reignited interest in neural networks. And it was also leading to the you know the strengthening the fact that GPUs could accelerate the model training. and it also marked the shift from feature engineering to representational learning. So, before that all those things were converted into features but then in this case neural nets were used. And then so during the time 2010 to 2015 the big data era started so where a lot of biological data like genomics data proteomics data and cheminformatics data was you know exploding and it was becoming available from large scale experiments And then the deep neural nets was also developed and it was, you know, it could learn from the large chemical libraries because it can handle even millions of compounds.

And then the CNN, Convolutional Neural Networks, it could analyze the images, help with the image-based analysis of molecular structures. And then we had the development of RNN as well, so which was used for analyzing time series biological data. And then we saw, you know, the development of DeepChem in 2015, which is an open-source deep learning library used largely for cheminformatics purpose. Where you can build, you know,

predictive models for predicting any kind of property, or you can do a lot of other stuff as well. And then in 2014 to 2015, the GANs and the Generative Adversarial Network and reinforcement learning was developed.

So in 2014, the GAN was introduced where the machine can learn to create realistic data. And then in 2015, the DeepQ network by DeepMind it was developed which beats human performance in Atari games. And then the AI started generating media and mastering complex strategies autonomously. So, in the year 2016, so another you know breakthrough where the DeepMind's AlphaGo it defeated world Go champion Lee Sedol. So, the Go game is a you know a very complex game at least for the traditional algorithms actually.

So in this case, what they use was they use reinforcement learning plus Monte Carlo Tree research method. And also, they also worked on the policy and value network, which predicted moves and outcomes. So, for the first time in 2016, the AI surpasses human intuition in strategic thinking. And then the reinforcement learning, it starts being adopted for the drug design. And then now you can see a lot of applications of reinforcement learning in in drug discovery and in drug design.

And then the Q-learning models, they were also being used or being explored for the de novo molecular generation. Okay, then in 2017, another breakthrough happened. And that was, you know, the transformer era. And then the Google introduced transformers, which revolutionized language processing. So, they published this paper, which was, you know, titled as attention is all you need.

And it was, you know, authored by eight Google researchers. And out of those eight Google researchers, six of them were born outside of US actually. So, the idea was that the advantages of the Transformers were that self attention mechanism allows parallel processing. And the models, they could scale, they scale better than RNN and CNN on, you know, on language processing. So, after 2017, we saw a lot of development in natural language processing.

So, the breakthrough models were like BERT, GPT, T5, they all they emerged after 2017. So, where we saw a lot of development in this field. So, in 2017, so we also saw the generative models as well. So, the introduction of variational autoencoders for molecular design was happening during that time. And then again were also used for designing novel molecules in the de novo drug design.

And then reinforcement learning, it was combined with generative models. So, by combining reinforcement learning with generative model, so you could produce, you could generate new chemical structures with desired properties like desired solubility, desired

bioactivity, desired permeability. So, all those things were the advantages of using reinforcement learning with the generative model. So then in 2018 to 2019, so AI goes multilingual and multitasking. So, in 2018, the BERT pre-trained enables contextual understanding of the language.

And in 2019, GPT-2, it generates human-like text, but OpenAI initially withhold it due to the misuse concern and then they released it later as GPT-3. And then the new capabilities were the translation, summarization, question and answer tasks all improved dramatically because it was a big challenge to use these models for natural language processing until the transformer technology was developed by Google. So, after that, we saw now, for example, we have this ChatGPT. So, it can excellently communicate with us without any problem because it has been trained on billions of words from the literature. And then in 2018, we saw another breakthrough where the alpha fold, which is again developed by DeepMind, which is a Google subsidiary company.

So, the DeepMind developed alpha fold. So, it predicted the protein 3D structure with unprecedented accuracy. So, this we will see in later sessions. And then this marked shift towards AI driven structure biology. And this was the paper which was published by Jay Jumper who later got the Nobel Prize as well in 2024. So highly accurate protein structure prediction with alpha fold.

So, the protein structure prediction problem is a very long-standing problem where the idea was can we solve the 3D structure of a protein with the help of computational tools, and just from the sequence actually. So, can we convert a protein sequence into the 3D structure? So that was the original idea. And then AlphaFold did the job excellently, and it could solve the structure of proteins with high accuracy. And then they developed AlphaFold2. And that is a highly used tool for predicting the protein structure.

And I think they have solved millions of protein structures by using AlphaFold2 now. So then in 2019, the generative models were also being developed. So, where a tool, for example, REINVENT, which is developed by AstraZeneca molecular AI team. So, it is a reinforcement-based learning tool for designing drug-like molecule. And then InSilico medicine, so which discovered a novel fibrosis drug in just 46 days, so which was a speed milestone, because discovering drugs takes like years, actually, several years.

And then atom wise it also it predicted protein ligand binding for small molecule discovery using the generative AI. And then the overall impact of generative models was that these models they prove their potential in early stage drug design. And in 2023, GPT-3 was released. So, OpenAI launched GPT-3, which was trained on 175 billion parameters.

So, it was the largest language model ever at that time. And the impact of, you know, development of GPT-3 was because it was working on the concept of few short learning models performs tasks without specific training. then a lot of developments like chatbots, content generation code, writing so it was developed so this GPT-3 was used in all these all these tasks actually. And then so the idea was that if we have bigger model that will lead to better performance but of course at the huge computational cost. And that is why a lot of you know, development happened in the hardware area as well, where, for example, Nvidia, it became a number one world largest company beating Amazon. And also, every other company they started investing in, in the hardware development.

So in 2021-2022, we saw the rise of multimodal AI where the tools like DALL-E, CLIP, they combine text and images which were enabling cross-domain understanding. And the PaLM which was developed by again Google, It showcases reasoning abilities even in mathematics as well. And then the models they learn from diverse data, text, images, audio simultaneously. So, we could see a kind of a general AI which can do the multiple tasks. And then in 2022, the large language models, they were developed like Galactica from Meta, which is an LLM trained on scientific literature, which can support hypothesis generation.

We had the BioGPT developed by Microsoft specialized biomedical LLM aid literature analysis and target prediction. IBM Rxn for chemistry, which predict chemical reactions and synthesis routes. And then the overall impact of and then the overall impact of all these LLMs was they start assisting chemists and researchers in knowledge heavy industry. Okay, so the year 2023 is saw the rise of autonomous AI chemists, like ChemCrow is one of them, which was, you know, which is an autonomous agent system combining LLM with 17 specialized tools for designing, synthesizing and analyzing molecules.

So, it can be considered as a kind of a robotic medicinal chemist. And then EvoDiff developed in 2023, which is an AI model which generates mobile proteins without existing templates and opening biological drug design frontiers. And Charm Therapeutics, it used deep learning for protein ligand structure prediction in oncology drug development and the autonomous multi-tool AI system which handles the end-to-end drug discovery tasks. And year 2023-2024, it saw a rise in the open-source AI and specialized models. So, there was like several open models they were developed like LLaMA, Mistral, Falcon.

So, they were democratizing the large-scale AI development. And then the AlphaFold2, it solved the protein folding problem and it was revolutionizing the structural biology. And then Gemini and Cloud, they pushed conversation to the new levels. And the focus was on efficiency, personalization, and domain-specific application. And we have recently seen the development of DeepSeek as well, which is claimed to be using a very less

computational power as compared to the other LLM models.

And efficiency-wise, it is nowhere less than others actually. So if we summarize this session, So, the AI evolved from rule-based system to data-driven machine learning models enabling pattern recognition and prediction and then the deep learning revolutionized drug discovery by modelling complex biological data accelerating target identification and molecular design The generative models like GAN, VAEs, they enabled AI to create novel molecules with drug-like properties, transforming lead discovery. And then LLMs like ChemCrow, they integrate multimodal data to assist in chemical synthesis planning, property prediction, and hypothesis generation. And then for more details, you can go through these articles or the links where you will get detailed information about this topic as well. And in the end, I have an open question for you.

So, imagine it is the year 2035. What new and mind-blowing things do you think AI could do that it cannot do today? And with that, thank you.