AI in Drug Discovery and Development

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

Week-09
Lecture-41

Welcome to the course "AI in Drug Discovery and Development." So, in this session, we will talk about the introduction to generative AI in drug design. So, by the end of this lecture, you will be able to explain the limitations of traditional drug discovery in terms of time, cost, and attrition. Describe the concept of generative AI and how it differs from predictive AI in the context of drug design. And then identify key applications of generative AI in early-stage drug discovery, such as de novo molecule generation or scaffold hopping. As well as discussing the opportunities and challenges of applying generative AI to drug discovery.

So, we have seen in earlier sessions that there are multiple ways to identify hit molecules and conduct virtual screenings. Predictive modeling is one of the tools that are used extensively, but if we talk about drug discovery and the challenges associated with it. So, we can see that the chemical space is really huge, containing at least 10 raised to the power of 60 possible molecules. Identifying a single drug molecule out of, you know, this huge space is a Herculean task.

So, ideally, we just, you know, pick molecules or filter down these huge molecular databases, and then we identify those molecules and optimize them into clinical drugs; however, that is a lengthy process. So, we use virtual screening, which is a set of computational methods that analyze large databases or collections of compounds in order to identify potential hit candidates. This search can be performed in corporate libraries and in virtual libraries. So, VS is faster than experimental screening because experimentally screening those millions of molecules is very difficult, challenging, time-consuming, and costly. So, the virtual screening can test about 10 raised to the power of 8 compounds in a day, while experimental screening may take years to screen those molecules, and it is also much cheaper than experimental screening.

So, what we do is take the chemical library, actually screen it based on filter-based methods or AI/ML-based methods, and then make predictions. If the prediction is good, we will pick those molecules and perform the experimental validation. So, let us talk about virtual screening. It is, you know, restricted to existing chemical libraries. So, the chemical space, as we talked about in the earlier slide, is about 60, which is vast, and VS can cover only a tiny fraction, like searching for treasure in one corner of a huge ocean.

However, there is one advantage: there is no need to synthesize those molecules because you can simply take the 2D or 3D structures of those molecules and predict whether they will be active or not. But there are, you know, if we are using a structure-based method like molecular docking. So, there are associated problems with them, such as the docking and scoring inaccuracies, where the docking algorithm simplifies reality by treating the protein as rigid most of the time. As well as the scoring function, which estimates the binding affinities, these are imperfect and can misrank the ligands. So, they can have a lot of false positives, actually.

And then the limitation, if you look at it, is that it lacks coverage. So, another thing is that the traditional techniques actually work best on handcrafted features. So, if you look at the ease of synthesis, for example, and coverage, these molecules are a little bit easier to synthesize. Why? Getting them from already existing libraries means that most of those molecules already exist and have been synthesized, so we are picking them from that library, cherry-picking them. So, they are easy to synthesize; however, their coverage is not very good.

Actually, they cover a tiny space because it is impossible, or it is highly difficult, to screen all those possible molecules, which can go up to 10 raised to the power of 60. And why those traditional methods are limited in chemical space exploration is due to high throughput screening, as if we wanted to screen them in physical mode using high throughput screening. So, we can only screen the existing chemical substances, such as those that you already know are synthesized and those that are in stock. Because we want to make a combinatorial library of millions or billions of molecules. First, we have to synthesize them, and then we need to test them.

So, it can only explore a small fraction of the possible chemical space. So, if we use combinatorial chemistry, it is still restricted by the initial set of building blocks and predefined reaction paths. We cannot explore highly novel or underexplored regions of chemical space. And the trial-and-error screening, which involves the random generation or modification of compounds and testing their activities. It is inefficient and resource-intensive to explore only a small fraction of all possible compounds.

An exponential growth of chemical space: the number of possible molecular structures increased exponentially with each additional atom or functional group. With just 20 building blocks, the number of unique molecules can grow to billions, making exhaustive screening impractical. Another important limitation, you know, is the synthetic feasibility. Many theoretically viable compounds are challenging or impractical to synthesize in the lab, limiting exploration to compounds that are synthetically accessible. So, if you look at this figure, it actually explains the molecular design approach.

So there are, you know, basically two approaches. One is the direct approach; the other is the inverse approach. And the direct approach is, you know, where we are experimentally determining the properties of each molecule. So here it is, you know, the functional space of the molecules where we have plotted the desired properties, such as solubility, toxicity, or redox potential. So, you can see this is actually a latent space in which we have plotted these properties.

And then we have the chemical space, like we have those drug-like libraries of molecules. So, in the direct approach, what we do is pick one molecule and then experimentally determine its properties. So, what is the solubility? What is the toxicity? We experimentally determine these properties. And then, in this case, if I want to identify a molecule with optimal solubility and no toxicity, I can. So, what I have to do is test all these compounds, and this is how you know the high-throughput screening works: it means a direct approach where we pick up each molecule and test those molecules.

So, testing all those molecules means it will take a lot of time; you know it will be costly, and it will also be manpower-intensive. However, if we talk about the inverse approach, what happens is that we already know the properties of those molecules. And then we know that these molecules exist; it is like the kind of high-throughput virtual screening where we have, you know, the solubility and toxicity data associated with them. And then we use this information to identify a molecule that is beneficial for us with the desired properties. And then this is called the inverse approach, and another inverse approach is where we use generative algorithms, or it is known as generative AI for de novo drug design.

In this case, what we do is start with, you know, one molecule, and then, for example, we design one molecule or pick one molecule, and then we determine the properties. And then we see that, okay, this molecule is lying here, and our objective is to reach this space, which has optimal properties according to our needs, okay? So now what we will do is generate a derivative of this molecule using generative AI, then determine the properties, and then we will do that again. So, in a minimal number of steps, we will try to reach this optimal space, and this is, you know, what the power of generative AI is and how we actually use it. So, generative AI is a broader category of AI that focuses on producing new content, whether it is text, images, music, or any other form of media. So, ChatGPT is a very specific example, a very specific type of generative AI: a large language model that is trained to engage in human-like conversations.

So, it uses natural language processing to understand the user input and generate relevant and coherent text responses. So, for example, here I am just asking if you can suggest some journals in the field of AI and drug discovery. So, it gave me a nice list of journals that are associated with, or are, you know, publishing research papers in this area. So, however, if I ask the generative AI, it is as if we ask ChatGPT to generate molecules. So, can you generate structures of new drug-like molecules that are permeable to the blood-brain barrier? So, it is not able to do that because this model is specific to text generation.

So, you know, nowadays it has the capability to generate images too. And maybe in the future, it will be kind of a general agent that can generate all types of content. But at that time, it was not able to generate the structures of the molecules. But there are tools that exist, and we will talk about those tools in this unit. So, ChatGPT can discuss, suggest, or analyze chemical ideas, but it cannot directly design valid molecules or simulate chemistry.

So, it talks about molecules; real molecule design needs domain-specific generative AI, and that is what we will discuss in this unit. And this is called, as you know, de novo drug design. It is the process of creating new drug molecules from scratch. De novo, from the beginning, involves designing entirely new chemical entities. So, it uses models like variational autoencoders, generative adversarial networks.

So, these are nothing but neural networks, actually, and they are used in reinforcement learning for molecule creation. So, we can design novel compounds with the desired properties, and we can tailor molecules to specific biological targets as well. So, we can also optimize multiple parameters, such as potency, selectivity, and ADMET properties, knowing that traditional drug discovery can take over 10 years and billions of dollars. So, the AI-powered de novo drug design can generate candidates in months, and we have seen some examples of this, such as those in silico medicines or BenevolentAI. So those companies have been using these generative models to generate drugs in a very short period of time.

So it is also known as generative chemistry, reflecting the rise of generative models in AI. The most important thing is that if we talk about the advantages, it is the coverage, because in virtual screening, what we get is very limited coverage. So, our molecules, which we identify from virtual screening, might not be novel in nature; in fact, someone else might have obtained IP rights on those molecules. But for the de novo drug design, these molecules are novel and new. So, you are moving into uncharted path.

So, it is like a completely new domain and a new molecular structure. So, you have a very strong advantage in obtaining those intellectual property rights for those molecules. And that is why generative AI, especially in drug discovery, is actually booming in this field. So, it is working in this case; in GenAI, the AI works as a molecular architect. So, generative AI in drug discovery learns from existing drug-like molecules to design novel structures with desired properties.

So, it explores a wide range of chemical space guided by drug-likeness rules and ADMET criteria by integrating property predictions, such as QSAR models. It focuses on generating molecules with optimized pharmacological profiles that go beyond traditional compound libraries. So, it is like ChatGPT has been trained on text. It can generate any kind of text based on the user prompt; likewise, we can develop a tool or a model. Which can learn the chemistry or the grammar of chemistry, such as how molecules are made, what their properties are, and which properties are related to what kind of structure.

So then we can have a generative AI tool that generates desired structures as well. So, what makes generative AI suitable for molecular generation? So, it explores vast chemical space efficiently. It can generate novel structures beyond the existing datasets, covering unexplored regions of chemical space. It can also learn the underlying chemical patterns. It can capture the structural relationships, chemical rules, and drug likeness from the training data.

It can be, you know, again, used for de novo molecule design, which means it can create entirely new molecules from scratch without relying solely on databases or fragment libraries. And then we can do the conditional generation as well, which can be directed to produce molecules with specific properties, like target binding, solubility, and toxicity. And the fact that it can optimize multiple objectives simultaneously means we can use it for multiple-parameter optimization as well. This is needed, especially during the lead optimization stage, as we need to optimize the solubility, potency, permeability, and toxicity. So, if you wanted to optimize all four of these parameters simultaneously, GEN-AI can be of good help for this.

And then it can, you know, accelerate lead discovery by reducing time and cost, and then we can do the integration with the predictive models. It can be combined with property predictors, docking scores, or synthetic models to filter and refine the output. And we will talk when we discuss those tools, you know, which are being used for Gen AI, okay? Talking about different stages, for example, GenAI accelerates drug discovery by creating, optimizing, and repurposing molecules, which makes the process faster. So, we can use it for de novo molecule design, where it generates novel molecular structures from scratch, focusing on ensuring drug likeness, potency, and structural novelty. An example is the AI designing antibiotics and kinase inhibitors.

We will talk about this later in the successful case studies, and then it can be used for lead optimization, as it can fine-tune existing drug candidates to improve efficacy, solubility, and safety. Where it uses AI-driven property prediction and multi-objective optimization.

The example is optimizing kinase inhibitors to reduce toxicity. It can be used for drug repurposing; it can identify new therapeutic uses for approved or existing drugs, which can reduce development time and costs compared to de novo drug design. And then an example is repurposing existing drugs for COVID-19 treatment using Gen AI.

And then it can be used for personalized medicine, where it can design molecules tailored to individual genetic profiles and leverage omics data for developing targeted patient-specific therapies. An example is AI-driven precision oncology for cancer treatment. Okay, but before talking about those models or studying those tools that can be used for generative AI, let us talk about molecular representation because it is important. If we want to teach the model how molecules are made, we need to understand how those molecules are represented. So, there are multiple methods of representing molecules.

So we can say that the most popular are the string representation and the line representation. And then you have the chemical table representation, feature-based representation, and computer-learned representations. So, in the string and line-based representation, you have the registry systems like the CAS. And then, in the structure-based string line representation, you actually have the InChI keys or the SMILES. And SMILES is becoming more and more popular, especially for these chemical language models where it is being used to generate new molecules.

Okay, and then in the chemical table, you have the MDL MOL file. Feature-based, you know, the eccentric connectivity fingerprints; these are the kinds of feature-based methods of representing a molecule. And then there are, you know, computers that learn, like those advanced neural networks such as VAEs, RNNs, and GANs. So, they can learn the features of the molecules, and then they can be represented by those features, you know. So, I'm talking about the string representation; it is more compact and easier for humans to read and write than any other representations.

So it uses the American Standard Code for Information Interchange (ASCII) character encoding standard. So, you can see here, this is the structure of dicyclovirine hydrochloride. So, the molecular formula is like this, and then the IUPAC name is again a feature, you know, a string representation method, and then you have the CASRN, Chemical Abstract Service registration number. So, this is the CASRN for it, and then you have the canonical SMILES. So, this is, you know, the SMILES representation, and then you have the InChI key, and then you have the WLN.

So, all these are, you know, string-based representation methods that represent molecular structures. So, talking about IUPAC, it is a set of rules and guidelines developed by IUPAC for naming chemical compounds systematically. So, it provides a uniform and unambiguous way to name chemical substances, ensuring that chemists worldwide can communicate effectively about the compounds they encounter. As an example, the molecule CH4 has the IUPAC name methane, but it would simply be referred to as C in the SMILES, and we will see how SMILES actually work. So, then again, I said that the most popular method of chemical representation is SMILES, which is the Simplified Molecular Input Line Entry System.

So, it enables easy exchange of chemical information and is essential in various applications including virtual screening, similarity searching and chemical database storage. So, it represents molecules with a string of ASCII characters, and atoms are represented by their elemental symbols. So, hydrogens are usually implicit, meaning they are assumed to be attached to atoms to satisfy their valency unless explicitly stated. Okay, but now the issue is that it can have, you know, a kind of randomized representation that can have multiple smiles for a single molecule, as you can start from anywhere. Like in this case, you are starting from here, so then C, C, and then C, okay? And then the like C and then C.

So, it is like first representing this benzene ring, then moving to this carboxylic functional group, and finally moving to this functional group. However, it will be somewhat of a problem because you can start from here. So, you will have multiple strings for the same molecule. So, that is why the canonical smile came into the picture, where they made a kind of set of rules that stated, "Okay, this will be a unique smile for each molecule." And then you can see here, for example, all these input molecules.

So these are input SMILES for the same molecule, and these are the unique SMILES in canonical form for these molecules. Likewise, this molecule can also be represented by these three SMILES. However, the canonical SMILES will actually be like this. And this makes it easy to have a unique smile for each of the molecules. And then you have this International Chemical Identifier, commonly referred to as the InChI key.

So it is an open-source string representation developed by IUPAC in 2005. So, unlike SMILES, which is a linear representation, InChI represents a 2D chemical structure as a fixed-length text string designed to be easily generated from a molecular structure. So, it allows for straightforward conversion to and from other chemical formats, making it a valuable tool in cheminformatics and chemical databases. So, it has several layers, actually. It has the main layer, which represents the co-parent structure; then, you have the charge layer.

Followed by the stereochemical layer, followed by the isotopic layer, followed by the fixed hydrogen layer, and the reconnect layer. So, you can see here that this is an example of a molecule, and this is the InChI Key for this molecule. And then coming to the representation of the chemical table. So, a chemical table lists the x, y, and z coordinates of each atom in a connectivity table and how they are bonded to each other in a molecule. So, it is typically used for representing molecules within databases or programs, and the most widely used format is the MDL MOL file, which exists in two versions: V2000 and V3000.

So, the MDL MOL file consists of three main sections: the first is the header block. So, you can see here that this is the MDL MOL file representation of leucine. So, the header contains the title, timestamp, and optional comment, and then you have the chemical table, which consists of the number of sections. So, you have the chemical table that shows the coordinates of x, y, and z, and then this block is actually showing the connectivity between the atoms, and in the end, this line ends with the m end. So, you have the atom block, then you have the bond block, and then you have the end, and this whole thing is called a chemical table, okay.

Then there are feature-based representations that describe molecules based on their features or substructural patterns. So, these representations focus on identifying and encoding specific features, motifs, or other meaningful characteristics that are important for understanding the properties and behaviors of molecules. So, these feature-based representations are valuable in various applications, including drug discovery, chemical informatics, and molecular similarity analyses. Then there are graph representations; this is another, you know, really important representation that is used extensively in graph-based generative modeling. So, these are, you know, the molecular graphs; they map atoms and bonds to nodes and edges, respectively, although alternative mappings are possible.

So, nodes in the molecular graphs are typically represented by atom symbols or points where the bonds meet. And the molecular graph representation is a two-dimensional object. They convey 3D information, such as atomic coordinates or bond angles. So, there are multiple tools like ChemDraw, Mercury, Avogadro, Vesta, Pymol, VMD, which facilitates visualization of both 2D and 3D representation of molecular graphs, aiding in understanding and analysis. So, you can see here, for example, this is the molecular structure of acetic acid, and we wanted to represent it in an adjacency matrix.

So, this is how we can represent these atoms in molecular graphs based on their connectivity with each other. And then there are extended connectivity fingerprints. So, these ECFPs are a type of molecular fingerprint method used in cheminformatics to represent chemical structures. So, there are, you know, different variations of ECFP, such as ECFP4, which has a radius of 4, and ECFP6, which has a radius of 6, that differ in the radius of the circular neighborhood considered. Smaller radii capture local environments, while larger radii capture more extended structural features.

The ECFP fingerprints are widely used in cheminformatics software and databases due to their simplicity, efficiency, and ability to handle a large number of molecules effectively. So, they are among the most popular fingerprinting methods for molecular representation and similarity-based tasks in chemical informatics and drug discovery. This is a kind of summary where you can see that a molecule can be represented as fingerprints, SMILES, potentials, graphs, and, you know, 3D properties. Okay, so all these are, you know, representations of a molecule that is being used extensively in cheminformatics and in drug design and discovery.

Okay, so let's have a quick look at those generative drug-design models. So, there are, you know, basically, we have two types of models: one is the chemical language model, which treats molecules as sequences of characters, like the SMILES string. So, it uses techniques from natural language processing and learns the grammar of chemistry, and based on that, it can generate new molecules. Okay, so the examples are Smiles, Variational Autoencoders, and REINVENT, which is a reinforcement learning model based on Smiles. And then Camberta, which is a transformer model based on SMILES, and then you have the SMILES recurrent neural networks and long short-term memory potentials.

LSTM networks, and then you have the graph-based models. This is another type of, you know, generative model that treats molecules as graphs where the atoms are considered as nodes and bonds are considered as edges. So, it uses graph neural networks or other graph-based techniques, and examples include MOLGAN, graph VAE, graphAF, or JT-VAE. These are some of the deep learning models that are being used, such as variational autoencoders, generative adversarial networks, and reinforcement learning. And then recurrent neural nets, transformer-based models, and diffusion models; we will go through them in detail in the coming sessions. So, coming to the summary, generative AI is revolutionizing drug discovery by accelerating de novo drug design, optimizing molecular properties, and significantly reducing the development timelines.

So the de novo drug design with AI shifts the focus from molecule discovery to the creation of entirely new compounds. And that is how it is, you know, increasing the chances of IP and IP assets, intellectual property assets as well. And then AI-driven drug discovery is inherently interdisciplinary, requiring knowledge in machine learning, chemistry, bioinformatics, and computational biology.

I have a suggestion for further reading. So, you can go through these articles. So, if you want to learn more about the introduction to generative modeling for drug discovery and development. And with that, thank you.