AI in Drug Discovery and Development

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

Week-04
Lecture-19

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about protein structure prediction and binding site identification using AI. So, by the end of this lecture, you will be able to understand the principles of protein structure prediction and their role in drug discovery. Compare traditional and AI-driven approaches for structure determination and also analyze different binding site prediction methods. So, if we talk about the protein structure, So, you know the protein has structures at different levels, actually. So, the protein is made up of amino acids, and the sequence of amino acids comprises the 1D structure, which is also known as the primary structure.

where those amino acids are connected to one another through peptide bonds. And then when those amino acids, they fold into either alpha helices or beta sheets, those are known as the secondary structure. These structures are formed by hydrogen bond interactions between the residues. So, you can see here that this sequence adapts to a helical alpha-helical structure.

And then the tertiary structure consists of the three-dimensional folding of a polypeptide. So, where it is taking not only the secondary structure in the form of an alpha helix or a beta sheet, but it is also a combination of, for example, loops, beta sheets, and alpha helices, and then comprises the three-dimensional structure. And then for functionally becoming active. So, the protein then adapts to the quaternary structure, which is the assembly of multiple polypeptides. For example, this complex contains multiple protein chains and a globular or quaternary structure.

So, this is functionally active and plays an important role as, you know, a structural component or, you know, as a functional element to perform various functions such as enzymes or receptors. If we talk about the structure determination methods, So, there are, you know, experimental methods like X-ray crystallography, nuclear magnetic resonance spectroscopy, or cryo-electron microscopy. And then there are computational methods as well, like homology modeling, ab initio modeling, threading, and AI-based prediction, which is largely based on deep learning. So, we will discuss that during this session. So, if we talk about X-ray crystallography.

So, this is a primary technique for determining 3D protein structures, and when we talk about the PDB structures deposited in the Protein Data Bank. So, about 84 percent of the structures are resolved using x-ray crystallography. Because x-ray crystallography has been, you know, a golden standard for determining the 3D structure of proteins. So, how do we determine the structure using X-rays? So, the first step is to purify the protein and then crystallize it. So, once we get those crystals, we need to collect the diffraction data.

So, we shoot them with the X-ray, collect the diffraction data, and then, by analyzing that diffraction data, we determine the phase. And once the phase is determined, we can then map that to the electron density. And then this electron density map is fitted into the model, and then the model is refined. So, finally, we get a three-dimensional structure that satisfies the electron density obtained from the X-ray crystallographic data. So, some of the limitations of X-ray spectroscopy and obtaining 3D structures from it are that it provides static structural information because whenever a molecule is crystallizing into a crystal, so that it contains only a single conformation which is repeatedly arranged in a crystal structure.

So, it limits the insights into protein flexibility, and it is highly costly and time-intensive, thus making large-scale studies challenging. Many of those proteins, for example, the membrane-bound proteins, are very tough to crystallize. So, you cannot obtain well-diffraction-quality crystals for all of those proteins, actually. And then we have the nuclear magnetic resonance spectroscopy technique, which is a powerful technique to determine the 3D structure of biomolecules in solution. So, it is suitable for non-crystallizable proteins and also for dynamic studies because here you can observe the multiple structures and multiple conformations of a protein.

So, how do we do that? Again, you prepare the sample; you purify the protein. And then you collect the NMR data, and then you assign the peaks for all those known atoms in the protein. Then you use the distance restraint extraction, calculate the structure, define the structure, and follow it with validation and analysis. So, some of the limitations of using NMR for structure determination are that the resolution is lower than that of X-ray crystallography. and then it is limited to small and medium-sized macromolecules, mainly less than 50 kilodaltons.

Because if you go beyond that, then it will be difficult for you to obtain the structure you know using NMR spectroscopy. And then another powerful technique that is now used very frequently is cryo-electron microscopy. Which really determines protein structure by rapidly freezing the samples, capturing the multiple projections using an electron beam, and then reconstructing the 3D structure using computational averaging. So, it is like obtaining your own 3D model by taking maybe 100,000 pictures of you from different angles and then merging all of them into a 3D statue of you, you could say. So, the first step is to prepare the sample, and then you simply freeze it using liquid nitrogen; that step is also called vitrification.

And the next step is to collect the data, such as images of the grids on which we have applied your sample and vitrified it. And then you pick the particles, and here "particle" means those single particle images of those molecules. Then, you do a 2D classification, classifying those particles based on their orientation, as we can say, for example. So, then you will get maybe 100 different kinds of clusters, which means that those 100 clusters are pictures of those particles or protein molecules from a certain angle. And then, using the 3D reconstruction with ab initio model generation, you generate the 3D structure of that protein model, followed by refinement.

And then there are post-processing steps such as masking, sharpening, resolution estimation, etc. And finally, you build the model, refine that model, and validate that model to see if it is close to reality or not. So, however, some of the limitations of cryo-electron microscopy are that it requires expensive microscopes. and also, computational resources because analyzing you know millions of those images. Picking those millions of particles and then generating the 3D reconstructed map is a little bit of a time-consuming process.

And then the sample preparation, data collection, and processing also take significant time for highly dynamic or small proteins, which are challenging to image at high resolution. So, it is not suitable for highly dynamic proteins that are changing their shape very frequently. Also, small proteins cannot be used for determining their 3D structure using cryo-electron microscopy. So, it is largely suitable for determining the 3D structure of large complexes, large protein complexes. Now, we have seen that there are some methods that we can use for determining the 3D structure experimentally, but can we predict the structure? So, that is a big question.

Can we predict the three-dimensional structure of proteins using computational methods? So, what we can get if we are determining the 3D structure from computational means is that we need, we do not need to do any experiments; we can save a lot of money, time, and resources. Computationally, 3D structure prediction involves the prediction of a protein's 3D structure from its amino acid sequence. Why is it important for drug discovery and structural biology? So, it can provide the insight into molecular interaction, functional mechanisms and therapeutic targets. It can help us design the drug with greater specificity and efficacy. It can enhance the understanding of disease mechanisms at the atomic level, and we can identify the functions of orphan receptors as well.

We can classify and cluster proteins based on resolved structures, and we can learn new biochemical mechanisms. For example, the active or inactive state structures of those proteins are important. Then we can predict the consequences of genetic variants, such as somatic mutations or others, which lead to functional changes in the protein structures, and we can design new proteins with new functions as well. So, if we are able to determine the structure of a protein, we can do all this, right? And if we can determine the structure of a protein by using computational tools. So that is even better because then we are saving a lot of, you know, resources, time, and money on doing the experiments to determine the 3D structure.

So, let us take a look at the computational methods for determining the 3D structure. So, one of the earliest methods is called homology modeling. So, it predicts protein structure using a homologous template from a known structure based on the assumption that proteins with similar sequences share similar structures due to evolutionary conservation. So, there are several tools like SwissModel, Modeller or iTasser. So, they use this methodology where you use a template structure.

So, the template is a 3D resolved structure obtained using either X-ray crystallography, NMR spectroscopy, or electron microscopy. So, you have that template, and if you can determine whether your target protein has a similarity to the template. So, you can assume that your target protein will have a structure similar to that of the template protein. So, you first identify a template, selecting a template that identifies a known structure with high sequence similarity. And next is the sequence alignment: you align the target sequence with the template.

Then you build the model by constructing the 3D structure based on the alignment and then refining that model by optimizing the structure using energy minimization. And finally, validation by assessing the quality using tools like the Ramachandran plot and DOPE scores. And this is how you get a computationally determined 3D structure of a protein. But the limitation is that it fails when no suitable template is available. For example, if there is no such template structure available that is similar to your target structure, it is difficult to use homology modeling.

Some of the tools that use homology modeling are, for example, SwissModel, which uses homology modeling to determine the 3D structure. So, it's a web server that uses the protein sequences or a template PDB and generates a 3D PDB structure. And it is moderate in speed, so you can get a structure very easily. And then some of the unique features of using this are that it generates reliable structures for highly similar sequences and it has a user-friendly web interface as well. And then it generates multiple models, optimizes the geometry, and also supports the modeling of the loops.

So, this is also one of the areas where this has been used very often to model the loops in a protein structure. And then you have Phery2, which is based on homology modeling and fold recognition again. It is a web-based server that uses a protein sequence in the FASTA format and then generates an output. It gives the predicted 3D PDB structure, and it is actually very fast. And then it is good for low identity templates and provides the confidence score as well, which indicates which of these predictions are good or not good.

And then you have the Robetta, which works on the basis of the principle of comparative modeling. It is again a web-based server that uses the FASTA sequence to generate the 3D PDB structure, which is moderately fast, and it can perform automated domain parsing and structure generation. So, the next technique was, these are not historical, but these are the earlier techniques that were used for determining the 3D structure of proteins. Another technique is ab initio or de novo modeling. So, it predicts the protein structure without using a template, purely from sequence through physics-based methods or AI-driven deep learning models.

How it works is that it does the energy minimization using the force fields to find the lowest energy state. and then followed by Monte Carlo simulations which generate multiple conformations to find the best fit Then, using deep learning, which utilizes the database and neural networks for accurate predictions of protein structures. So, some of the tools that are using this ab initio modeling are Rosetta and Quark. One of the limitations of using ab initio is that it determines the structure from scratch; therefore, it is computationally very expensive and sometimes less accurate, especially for large proteins. Another technique is threading, which is known as fold recognition.

So, it predicts the structure by matching the sequence to a library of known protein folds with the assumption that even with low sequence similarity, proteins may adopt similar 3D folds. So, this is similar to, you know, homology modeling, but instead of using the full protein of the template structure. So, instead of what we are interested in, we are interested in smaller sequences of the sequence database that contain, for example, similar sequences that are being folded into similar shapes. So, instead of using a whole structure, it uses the short sequences and their structure to build the model by combining all those sequences into a structure. So, some of the tools are iTasser, Phery2, or Muster.

So, how it does this is by identifying the fold library. So, it searches for the structural motifs in the known databases. And then, for example, 200 amino acids; the first 50 amino acids make one structural motif so that those 50 amino acids can have a similar structure in the database. So, you identified the fold library. So, you search for structural motifs in a known database followed by aligning those sequences to the folds.

Use the scoring functions to find the best match, and then construct the model and build the structure based on the closest fold. and then followed by refinement and validation, where you improve the accuracy with energy minimization. So, some of the examples are Rosetta, which uses ab initio, homology modeling, and threading as well. And then you have iTasser, which is using threading and ab initio. Then you have the Quark, which is based on ab initio or fragment-based methods.

And all of them are web-based servers and take the protein sequence in the FASTA format. And then it generates the 3D PDB structures as output. And then, speed-wise, the Rosetta is a little slow. And then iTasser is relatively faster compared to Rosetta and Quark. And then the unique features of Rosetta are that it can generate flexible loops.

So, it can also perform molecular docking, protein-protein docking, and it has a protein design module as well. And then for iTasser, it predicts the structure and biological functions and ranks the models. And then for Quark, it predicts without a template, and it is good for novel folds. So, the folds that have not been, you know, previously reported. So, it can generate the structures for those folds as well.

So, this is, you know, an excellent exercise called critical assessment of structure prediction (CASP). So, it was established in 1994 and is held every two years to benchmark protein structure prediction methods. The exercise is kind of a competition where the organizers ask for three-dimensional experimentally resolved structures from researchers that have not been published yet. But they have resolved the structure, and then they ask the research community to predict the three-dimensional structure of those structures by using different computational methods. And then, after the submission of the results, they compare the structure with the experimentally determined structures, which have not been released yet.

And then they see which of those model methods they are able to use to predict the structure accurately. So, it started in 1994, but for example, if you compare the CASP 7 competition, which was held in 2006. So, those fragment-based assembly methods, like threading or homology modeling. So, those were, you know, dominating at that time. Because homology modeling, threading, and ab initio modeling were the only methods used for determining and predicting the 3D structure.

So, the GDT (Global Distance Test). So, it determines how efficiently the experimental methods can determine the accurate structure of those proteins. So, by 2014, people were using coevolutionary analysis with deep learning, and this coevolutionary method was, for example, if mutations were happening in one protein during evolution. For keeping that protein active, the partner protein in the complex is also evolving so that it can preserve its function. Otherwise, that mutation can lead to non-functional proteins and can result in the death of the cell, or lead to, you know, non-functional proteins, actually. Can we determine the evolution of those mutations by using computational means? So, all those methods were used during 2014, and then these methods were used in this co-evolutionary analysis.

And also the deep learning, and here deep learning was usually used, as you know, like simple neural networks, like convolutional neural networks. For determining the structure or fitting those known protein structures into the models, actually. A breakthrough happened in 2018 when AlphaFold participated in the CASP competition for the first time and was able to achieve a GDT score of more than 50. And then, in this case, AlphaFold was not using deep learning for the end-to-end process. So, it was just using deep learning to minimize the structure of those proteins, actually, those models.

But the AlphaFold 2, which had a GDT score of more than 80, was impressive. So, the AlphaFold 2 was revealed in 2020. So, it was an end-to-end method where it just used the sequence, and from that sequence, it was able to determine the 3D structure of those proteins by using deep learning methods. So, this was, you know, a kind of revolution, I would say, and after that, several other platforms have emerged, like RosettaFold, OmegaFold, and ESMFold. So, they have been, you know, using all these deep learning technologies to determine the protein structure.

So deep learning has transformed protein structure prediction by utilizing neural nets and large-scale data, significantly improving the accuracy, efficiency, and accessibility of complex structures. So, it utilizes multiple sequence alignments (MSA) to identify coevolutionary relationships between the residues. And extract the structural constraints from homologous sequences and predict the 3D structure using deep neural nets and attention mechanisms. So, it is highly accurate for proteins with homologous sequences and incorporates evolutionary constraints to improve predictions.

The example, as I said, is like AlphaFold to RosettaFold. And some of the limitations of these tools are that they are less effective for orphan proteins. or for the proteins where the structure of those, the structure of a similar protein has not been resolved earlier for example, for the intrinsically disordered proteins, such as the human tau protein. So, if you look at how Alphafold works, it starts with the input sequence and then searches the genetic database, prepares the multiple sequence alignment, and aligns all those sequences that are similar to the input sequence. And simultaneously, it searches the structure database where it identifies templates, and then, from the MSA representation and multiple sequence representation, the templates from the structural database. It feeds this data into the Evoformer, which is a deep learning-based model, and then it generates the structure through the structure module.

So, you get a structure, and then the refinement happens multiple times, and finally, you get a refined 3D structure that shows good confidence, and you also get this score as well. The residue by score shows you which of those you like, so you get the residue score as well, which we will discuss later in this course. Okay, and then you have the protein-language models. So, which learns directly from sequence data without relying on multiple sequence alignment. So, it uses transformer-based architectures like natural language processing models to understand sequence patterns based on the first structural features and the contextual embedding of amino acids.

And it is highly effective for proteins with no evolutionary information, and it is also faster compared to the alignment-based methods. So, examples are ESM fold and omega fold, and some of the limitations of these protein language models are that they have slightly lower accuracy than the MSA-based models for well-characterized proteins. So, this is a comparison between those transformer-based deep learning models like AlphaFold2, RosettaFold, OmegaFold, ESMFold, and ProteinBERT. So, all of them are transformer-based deep learning models, and they are largely written in Python; most of them have this cloud web server. They can be used to take the FASTA sequence as input and then generate the PDB structure as output.

And then most of them are relatively fast, but this AlphaFold has a moderate speed. And then some of the unique features, for example, AlphaFold, are that it is highly accurate. And it also supports multiple multimer modeling, where you can see how two protein chains interact with each other; then it can predict the side chain as well. And then for Rosetta Fold, it is a multitask network that can predict structure and contacts as well. And then the Omega fold works without MSA, as I said; it is using just the sequence and directly generating the sequence structure from the sequence only.

And then the ESM fold is lightweight, and it's very, very fast as well. And then ProteinBERT focuses on function prediction alongside the structure as well. So, then there are hybrid models which combine template-based modeling and deep learning predictions. So, it uses both non-structural templates and AI-based structural refinement to enhance accuracy.

So it balances accuracy and generalizability. It's effective for both homologous and novel proteins. Some examples are RaptorX and DeepFold. However, the limitation of these hybrid models is that they require high-quality templates for optimal performance. So, that was about, you know, protein structure determination. Next, coming to the binding site prediction, because if we did now, we have determined the binding site.

So, can we determine the binding site, actually? So, if we can determine the binding site of a protein or a receptor or a target. So, that will be highly useful for us to design drugs using the structure-based drug design methods. So, a binding site is a specific region on a biomolecule, typically a protein or an enzyme, where a ligand, such as a drug, substrate, or inhibitor, binds. So, this interaction is usually mediated by non-covalent forces such as hydrogen bonding, van der Waals interactions, ionic bonds, and hydrophobic interactions. And this is an example where an enzyme has been shown with a, you know, co-crystallized ligand bound here, and this is the binding site where this ligand has been bound.

So, then there are multiple types of binding sites, like one being the active site, which is found in enzymes where the substrate binds and undergoes a chemical reaction. It is also known as the catalytic site, and then you can have an allosteric site, which is a secondary site where a molecule binds to regulate protein activity. So, those allosteric sites are different from the catalytic site or the active site. And then you have the orthosteric sites. So, the primary binding site for endogenous ligands, such as neurotransmitters and hormones, exists.

So these are usually the case with receptors. So, there are some challenges in traditional binding site prediction. So, there are multiple binding site prediction methods, like grid-based methods, energy-based methods, and evolutionary conservation approaches. For example, the grid-based methods use a 3D grid around the protein to identify potential binding pockets. So, it requires predefined grid parameters, thus limiting the flexibility. So, it may miss dynamic or allosteric sites due to rigid assumptions, and it is also not ideal for proteins with unstructured or highly flexible regions.

Coming to the energy-based methods, it employs force fields to calculate interaction energies between the protein and small molecules. So, computationally it is very expensive, especially for large systems and it struggles with highly flexible binding sites and then it can generate many false positives which requires additional filtering. And then the evolutionary conservation approaches identify conserved residues across homologous proteins, assuming that they play key functional roles. So, it is limited to well-characterized protein families, making it ineffective for novel proteins. And you cannot predict new binding sites in orphan or poorly studied proteins in this case by using evolutionary conservation approaches.

So, let us have a look at various methods that use AI for binding site prediction. So, here is a comparison. So, for example, you can use DeepSite, P2rank, Kalasanty, Siteseq, DeepBind BC, GraphSite, FPPred, and BindUP. So, most of them are using the input structure as the 3D structure of the protein, while some of them can use only the sequence as well. like they themselves can generate the 3D structure and then identify the binding pocket, like Siteseq, FPPred, or BindUP, actually.

And then they are using DeepSite like CNN, Convolutional Neural Net. P2Rank is using machine learning, and Kalashanty is using 3D CNN. And then most of them are actually using the deep learning models. So, the key features of deepsite are, for example, that it analyzes spatial features of proteins, and the strength is that it is highly accurate on known structures, while the limitation is that it needs accurate protein structures. If the protein structure is not accurate, then the prediction of the binding pocket will not be accurate.

And then the P2 rank uses an atomic feature-based scoring method. It is highly fast and structurally independent. So, it has limited performance on the flexible proteins. Then you have the Kalasanty, which is trained on known pockets, and it is robust on diverse structures. However, it's computationally intensive, so it requires a GPU for the best performance.

And then you have the SiteSeq, which combines sequence motifs with deep learning. And then it also works without the structural data. However, it has lower precision compared to the structure-based method because it is, you know, using the sequence to determine the structure and then to determine the binding pocket. So, it adds, you know, those additional errors into it, and then you have the DeepBind B C, which uses the protein surface modeled as a graph. And then it captures the geometric features.

However, it is computationally intensive again. Then you have the GraphSite, which learns from the spatial protein topology and captures complex biological sites; however, it needs high-quality input structure. And then you have the FP-PRED, which uses physicochemical properties. It is good for functional site predictions. Then, however, it is limited to certain protein types. And then you have the BindUP, which is only a sequence-only predictor; it is effective when the structure is unknown; however, it is less accurate for novel proteins.

So, largely this is a comparison between different AI-based tools that are being used for binding site prediction. So, DeepSite, if we go into a little bit of detail on DeepSite. So, it uses a 3D CNN model trained on scPDB for protein binding site detection and utilizes Autodock 4 atom types as input, voxelizing the protein structure into a 3D grid. Then, predict binding site coordinates with confidence scores as well. So, it captures special features better than classical methods; it is web accessible, so no complex setup is required.

However, it is limited to the scPDB dataset and struggles with novel proteins, and it has lower performance than P2Rank on benchmarks as well. And the P2Rank, which is another highly used AI-based tool for binding site prediction. So, it is a template-free machine learning tool that uses a random forest model and analyzes solvent-accessible surface points with chemical and geometric descriptors. and predicts the binding site without relying on homologous structures or pre-processing, and it assigns a likability score to detected pockets. The advantage is that it has higher accuracy than DeepSite and SiteHound, and it requires no template or supplementary data set because it works only with the protein sequences and input.

So, coming to the summary, this AI-driven protein structure prediction is transforming drug discovery by improving accuracy and efficiency. Because if we can generate an accurate protein structure, a 3D structure of the protein, by using artificial intelligence, then we do not need to use those expensive and time-consuming wet lab methods like X-ray crystallography or single-particle cryo-electron microscopy. So, then we can use those protein structures for, you know, structure-based drug design, and then design novel drugs and develop drugs for the diseases. And then deep learning models such as AlphaFold 2 and ESMFold have surpassed the traditional methods, enabling rapid and precise structural determination. And then the binding site prediction has evolved from classical approaches to AI-based approaches like DeepSite and P2Rank, improving drug target identification.

So, in the end, I have an activity for you. So, you shall get the sequence of any protein of your choice from the UniProt data set and predict its 3D structure using any of the protein structure prediction servers that we have discussed. And these are some of the references to which you can refer. So, to get more information on this topic. And with that, thank you very much.