AI in Drug Discovery and Development

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

Week-01
Lecture-03

Welcome to the course AI in Drug Discovery and Development. Today, we will be talking about drug design strategies. By the end of this lecture, you will be able to understand structure-based drug design, ligand-based drug design, and fragment-based drug design approaches. You will also be able to learn more about different techniques like molecular docking, pharmacophore modeling, and QSAR for hit identification and lead optimization. And we will also learn about bio-isosteric replacement for drug optimization. So, we talk about the computational methods for drug design.

So, there are basically several possibilities where we can use computational methods in drug design, but it depends on whether the structure of the receptor and the ligands are available or not. So, if the structure of the receptor which is either an enzyme or a receptor or it can be the DNA or RNA. So, if it is if the structure is in our hand and the structure three-dimensional structure or two-dimensional structure of the small molecules is also in our hand, then we can use structure-based drug design and ligand-based drug design methods. So, which are you can see on the top right hand corner quadrant.

But if the structure of the receptor is only available and there is no ligand information available, so we can use only structure-based methods, which we will talk about different structure-based method later in the course. And if we have the structure of ligands only in our hand, we cannot use structure-based method, we can only use ligand-based methods. And if there is no structure information available for either receptor or the ligand, we cannot use any of those methods. In that case, we have to determine the structure of the receptor either using homology modeling or using the 3D structure prediction tools like alpha fold, etc. So, let us talk about structure-based drug design methods.

So, the structure-based drug design methods, they rely on three-dimensional structure of a biological target that is again a protein, enzyme or receptor to design the drugs that fit precisely into the active site. So, we need the three-dimensional structure of biological targets and those structures we can determine using X-ray crystallography, NMR spectroscopy and cryo-EM. Now, we can even use the deep learning methods to predict the three-dimensional structure of proteins or other biological targets. So, here you can see an example of the three-dimensional structure of protein which is called as acetylcholinesterase which has been deposited as a PDB ID 4EY7 in the protein data bank. So, you can see the ligand bound in the binding pocket and then the protein is shown as a

cartoon.

So, those structure-based drug design methods they also help in optimizing drug target interaction for increased potency and selectivity. So, they come very handy when we are optimizing those lead compounds or we are converting a hit compound into a lead compound. So, molecular docking is one of the structure-based method which is a computational method that predicts the preferred binding orientation of a small molecule which is ligand within a protein's active site. So, it involves two steps, the step one is a pose generation. So, we take the receptor structure or the protein structure or the biological target we can say and then we take the structure of the small molecule.

And, after obtaining these protein structure which can be from either from the X-ray crystallography or from NMR or from cryo-EM or it can come from the predictive modelling as well like alpha fold. So, then you prepare the ligand and protein structures. So, there are several steps which we have to follow to prepare the receptor and the protein structures well as well. So, after preparing the structures, we perform the molecular docking. So, there are several tools for performing molecular docking.

So, some of them are for example, auto dock, glide, gold. So, what those molecular docking tools are doing is they first generate the poses of these ligands into the binding pocket of the receptor. And in the second step, those generative poses are being ranked. So, the ranking is based on some scoring. So, these terms are called as docking score or it can be termed also termed as binding affinity, which is in the form of the docking score.

So, using this docking score, you rank those poses and then you prioritize some poses which you believe that those are close to the natural binding pose of that ligand into the receptor. So, there are multiple types of docking methods available in the field. So, most preferably we perform rigid docking where the protein remains fixed, where the ligand is kept flexible because of if we keep the protein flexible, then there will be a lot of degrees of freedom which will be difficult to control. And in the flexible docking, we keep both the ligand and protein flexible. That is a better way to perform the docking because then you are also taking into account the flexibility of protein side chain residue or the amino acid residues.

However, it will be computationally highly intensive. So, these are some of the limitations of molecular docking. So, we talk about the first limitation which is related to structural and flexibility issues. So, the protein and ligand flexibility which is often oversimplified in the molecular docking because as I said like we are using rigid protein and the flexible ligand. However, in reality it is not true because in reality the protein is also flexible at physiological conditions.

So, we are somehow restricting the flexibility of protein and that leads to oversimplification of this binding problem. So, another challenge is related to the induced fit effect which is not fully captured leading to inaccurate binding predictions. Another limitation is with the water molecules and solvation effects in the binding sites which are poorly modelled in the molecular docking tools. So, there are some limitations related to the scoring functions as well. So, one of them is that scoring functions they prioritise shape complementarity, but may fail to predict actual binding affinity.

Another challenge is that the poor handling of enthalpy-entropy compensation effect ranking of the ligands. Another challenge is that we get a lot of false positive and false negatives due to inadequate energy calculations by the scoring functions. Another area where the molecular docking tools are having problems is that experimental relevance. So, the molecular docking tools they have experimental relevance issues as well such as the docking does not account for physiological conditions for example, pH, ionic strength etcetera. And sometimes the proteins are membrane bound and then the allosteric sites are there harder to model accurately using these available methods.

And lack of explicit solvent dynamics can also lead to misleading results. So, after structure-based design now let us see what are the ligand-based drug design methods. So, the ligand-based drug design methods as I said like they use the structure of the ligand. So, they do not require the 3D structure of the target protein. So, they utilize the ligand structure information.

So, it is a computational approach that relies on structural and chemical similarities to non-bioactive compound. So, the key principle is that structurally related molecules they exhibit similar biological activity due to shared pharmacophoric features. So, you can take this analogy that an aeroplane is structurally very similar to a bird. And that is how what we can assume is that if a molecule is very similar to a known drug, so that should exert a similar kind of bioactivity. So, some of the ligand-based drug design methods are like pharmacophore modeling, molecular similarity searches and quantitative structure activity relationship.

Let us see what is pharmacophore modeling. So, a pharmacophore is a three-dimensional spatial arrangement of atoms, groups or functionalities in a small molecule that are required for specific interaction with its biological target and its activity. So, it is a ligand-based drug design approach that identifies the essential structural and chemical features required for a molecule to interact with the biological target and elicit a biological response. So, what we can do is we can also use structurally diverse compounds that can share chemical binding interactions if they have a similar pharmacophore. So, how do we extract a

pharmacophore is like we start with the active ligand, we obtain the least energy 3D structural conformation and then we extract those features.

And then we develop a pharmacophore model which is the three-dimensional spatial arrangement of all those important features which are required for the activity. So, then we can use that pharmacophore model for performing virtual screening and then that can lead us to identify hit compounds. So, some of the pharmacophoric features, for example, are hydrogen bond donor, so which can donate a hydrogen bond to the receptor amino acid residues. And examples of groups which can act as a hydrogen bond donor are OH and NH_2. And then there are hydrogen bond acceptor.

Carbonyl group can act as a hydrogen bond acceptor. So, then another pharmacophoric feature is hydrophobic centre. So, this is a hydrophobic region which interacts with non-polar amino acid residues. For example, aromatic rings can act as hydrophobic centres. And then we have the aromatic ring also as a pharmacophoric feature, which shows pi-pi staking with receptor residues.

So, benzene can be an aromatic ring pharmacophoric feature in small molecules. So, additionally we have those positive, negative or ionizable groups as well as pharmacophoric features. So, what they do is they have this possibility to interact with the amino acid residues of the protein binding pocket through electrostatic interactions. This could be cationic or anionic. So, quaternary ammonium can be a pharmacophoric feature.

So, quaternary ammonium functional group can be a positive or negatively ionizable group as a pharmacophoric feature. So, here you can see an example of a pharmacophore model where green, orange and white spheres they represent hydrophobic, hydrogen acceptor and hydrogen donor groups respectively. So, then we come to the molecular similarity search as I said like if two molecules, they are already similar in structure. So, we can assume that they will have similar kind of activity as well. So, what we do here is we compare an unknown molecule to a reference compound based on structural and chemical properties and we assume that structurally similar molecules have similar biological activities.

So, we can use several methods. So, one of them could be Tanimoto coefficient fingerprint-based similarity method. So, where what we do is where we measure the overlap of molecular feature using the fingerprints. So, the range of this similarity could be 0 which indicates no similarity to 1 which is showing that the molecule is identical to the reference molecule. Another method is the 2D structural patterns where we can use again use the molecular fingerprints. So, what we do here is we encode the functional groups atom connectivity and substructures as binary pattern.

Then some of the common fingerprint types are MACCS key which is having 166 predefined substructures and ECFP extended connectivity fingerprint for bioactivity prediction. So, all these fingerprint-based methods we will be going to use during this course actually. And then we can also use 3D shape-based similarity methods, which uses the three-dimensional molecular conformation of the molecules. So, what we do here is we compare molecules in a 3D space based on steric overlap, shape matching, electrostatic potential distribution. And some of the 3D similarity methods are, for example, ROCS, which is rapid overlay of chemical structure, which aligns molecule based on the shape and pharmacophore features.

And Gaussian shape overlap which measures the volume similarity between molecules. Another ligand-based drug design approach is quantitative structure activity relationship. So, these QSAR models they establish mathematical relationship between chemical structures and biological activity helping predict drug potency, selectivity and ADMET properties. So, there are multiple QSAR methods available in the field. So, for example, there are 2D QSAR methods called as classical QSAR which uses physicochemical descriptors to correlate molecular properties with biological activity.

So, the common descriptors which can be used are logP which indicates lipophilicity and it can influence the membrane permeability of molecule. And then we have molecular weight which affects drug solubility and absorption and then we have topological indices which represents molecular shape and connectivity. Then, there are three-dimensional QSAR methods. One of the examples is comparative molecular field analysis or COMFA analysis. So, which analyzes steric and electrostatic interactions of molecules in 3D space and compare a set of aligned molecules to build a predictive model and it can tell us whether a steric or electrostatic groups at certain position in a molecule will be beneficial or harmful for the activity.

And then there are AI driven QSAR methods, machine learning and deep learning-based models. So, which uses complex machine learning models like random forest, neural network, deep learning for high dimensional SAR modelling. So, these we will be going to study later during the course. And the some of the advantages which they show are they had they can handle non-linear relationship in molecular data. And we can learn from large data set improving predictive accuracy and also they can detect hidden molecular features that traditional QSAR methods cannot.

So, if you summarize the workflow of structure based drug design and ligand based drug design methods are that we first start with target identification and after identifying the target. So, after target identification we can go for structure-based methods where we need to have the 3D structure of the target or if we do not have the structure of the target in our

hand. So, we can predict the structure using either homology modelling or the deep learning methods like Alpha fold or Rosetta fold. And then we can perform high throughput docking and we can also use the de novo ligand design methods. On the other hand side, we can use the ligand based methods which where the information of the ligand structures are needed and then we can go for the QSAR, we can go for the pharmacophore mapping or we can perform the similarity based searches and identifies the hit compounds.

So, once we have hit compounds in our hands those can be converted into lead and then those lead molecules can be optimized into the drug candidate actually followed by. So, those lead compounds can then further evaluated and then can be moved on for converting them into a drug candidate. So, another technique which is frequently used in drug design is a fragment-based drug design. So, which involves designing potent small molecule ligands from low molecular mass fragment molecules. Usually, the drug molecules they have molecular weight up to 500.

But if we take a molecule which is having molecular mass of less than 300 Dalton. So, usually they do not possess very potent activity towards the biological targets, but they can be initial starting point for converting them into a potent drug molecule. So, those molecules are called as fragments and then they follow this fragment rule of three where the molecular mass is always less than 300 Dalton. They have less than or equal to 3 hydrogen bond donor groups and they have less than or equal to 3 hydrogen bond acceptor groups and the calculated logP which is ClogP is always less than or equal to 3. So, some of the stages in FB-DDR, we assemble a library of diverse fragment molecule that is called as fragment library design.

Followed by fragment elaboration where we expand the validated hits into lead compounds through synthesis, structural insights and computational design. And then we can go for fragment screening where we use biophysical techniques to detect fragment binding to the target. So, now we will have look at different fragment optimization strategies. So, one of them is evolving fragment which is one of the key techniques in FBDD. where initial fragment hits are systematically modified by adding functional groups to enhance binding affinity and specificity.

So, in this approach, so this approach ensures that fragment retains its original binding mode while gaining additional interactions ultimately evolving into a potent lead compound. Here we can see an example of fragment evolution of 6-propyl isocytosine leading to a potent non-peptidic BACE-1 inhibitor, where the compound 1 was initially identified as a fragment which was having 28 percent inhibition at 1 millimolar concentration using surface plasmon resonance which is a biophysical method for performing the binding active binding affinity. which is a biophysical method to determine

the binding affinity. So, this fragment was further evolved into compound 2, which showed an IC_{50} value of 130 μM, which was determined, which was determined using FRET based assay. And then later on this compound was developed into compound 3 which was showing an IC_{50} value of 0.08 μM using the FRET assay. Another technique for getting lead molecules from the fragments is combining fragments. So, where we can if we have two fragments binding to distinct non-overlapping regions of a protein. So, they can be merged into a single molecule. So, in this example, you can see that fragment linking for the BCL2 family of protein leading to the identification of compound 11 as a highly potent inhibitor of BCL2 family of protein. So, here the compound 8 and compound 9, they were independently identified as fragments with having 400, with having Kd value of 400 for compound 8 and Kd value of 2000 μM for compound 9.

So, when these two fragments were merged together, so the IC_{50} value got improved. So, it became 6.9 μM and further it was optimized into compound 11, which was having very high potent activity, which was having potent inhibition activity against the BCL2 family of proteins. Another method is in-situ fragment linking. So, which is a strategy where fragments are linked directly within the active site of protein using in-situ chemical reactions.

So, usually we use a very mild reaction conditions for example, aqueous environment or room temperature. So, some of the common methods they include are dynamic combinatorial chemistry and click chemistry to generate inhibitors. That in situ click chemistry was used to identify sub nanomolar inhibitor of carbonic anhydrase. Which is a zinc containing metalloprotein with a compound 15 having Kd value of 0.037 μM was converted into a lead compound 18 by using the click chemistry by synthesizing this triazole derivative.

And then developed into a potent sub-nanomolar inhibitor of carbonic anhydrase. Another approach is called as fragment tethering. So, which is a covalent fragment-based approach where reactive fragments they form a disulfide bond with a cysteine residue in the target protein. So, how does it work is the chemically reactive fragments they interact with the cysteine near the binding site and then if we have a strongest binder, the strongest binder will form the most stable disulfide bond. So, here we can see in this example that researchers they identified caspase 1 inhibitor using a fragment tethering approach.

So, then there are some computational approaches as well for fragment-based drug design. So, we can generate a fragment library. So, fragments can be generated using computational deconstruction of drugs. And then we can perform the virtual screening and docking as well. So, the structure based computational method they can predict fragment binding modes and assist in lead optimization.

So, the molecular docking it refines fragment orientation and suggest substitutions for better interactions. So, you can see here in this example the identification of potent DPP-4 inhibitors using in silico fragment-based screening. So, where the compound 27 was identified as an initial fragment using in silico methods which was having an IC_{50} value 40 μM. Which was further developed into the compound 28 which showed an IC_{50} value of 0.023 μM. So, these are some of the successful examples of fragment-based drug design where the clinical drugs like pexidartinib, vemurafenib, venetoclax, erdafitinib. So, these all have been developed into successful clinical drugs by using the fragment-based drug design approaches where the starting point of those drugs were the fragments obtained from different screening methods and those fragments were later developed into these clinical drugs. So, some of the advantages of fragment-based drug design are that it is an efficient way to explore the chemical space. So, the traditional high throughput screening libraries, they cover only a small fraction of possible drug like molecules. However, fragment-based libraries, they cover a broader chemical space of possible molecules.

The ligand efficiency which is determined by -ΔG/number of heavy atoms. So, all those fragments, they are usually very smaller in structure. So, that is how they have very high ligand efficiency. So, every method has its limitations as well. So, some of the limitations of fragment-based drug design are that it does not work for all drug targets due to low fragment potency.

And it has limitations in whole cell screening and kinetic assays as well and some of the proteins are highly charged. So, they may interfere with reliable fragment binding detection as well. And then of course, there are chances of false positive results as well, which is one of the limitations of almost every screening method. Another drug design strategy is de novo drug design, which is a computational approach that aims to generate novel bioactive compounds from scratch without relying on existing chemical libraries. So, unlike traditional drug discovery methods that screen existing molecules, in de novo drug design approach, we are focusing on constructing molecules that are optimally fit, that optimally fit a given target, thereby improving specificity, binding affinity and drug-like properties.

So, there are two strategies basically for de novo molecular design like holistic generation which construct entire molecule from scratch and it is useful for early discovery phases and wrote and exploring the broad chemical space. Another way is the iterative generation where build we build a molecule step by step. So, it is suitable for refining structures and tailoring them for specific purposes. So, some of the challenges with the dean over drug design are that one of the biggest challenges is the synthetic feasibility because here we are designing a molecule from scratch and we do not know whether the molecule will be synthesizable or not. So, some of the generated molecules may be difficult or impossible

to synthesize practically.

And then we have another limitation is the biological relevance where the computationally designed molecules they must possess drug like properties and bioavailability. And then there is scalability issues as well, where generating and validating large libraries of novel compounds, they require extensive computational and experimental resources and target flexibility where the proteins are dynamic. So, rigidly designed molecules, they might not fit well under physiological condition. And another big limitation is the limited data availability, because if we are using those AI driven approaches, they require large data to train the model on. So, we need high quality data set which are often scarce in the initial therapeutic areas such as those rare diseases.

But despite the challenges, advancements in machine learning, structural biology and cheminformatics are continuously improving its efficiency and applicability. Another technique is bio-isosteric replacement where bio-isosteres are the chemical groups or molecules that have similar physiochemical properties and produce comparable biological effect when substituted in a drug molecule. So, this bioisosteric design involves replacing part of a molecule by another part which has a similar property. And the key goals of bio-isosteric replacements are to improve the binding affinity of the target, enhance pharmacokinetic properties, reduce off-target toxicity and increase the chemical stability.

So, we have two kinds of bio-isosteres. One is the classical bio-isosteres where the atoms or groups have similar valency and electronic properties. For example, hydrogen can be replaced with a fluoro group which modifies the metabolic stability or a hydroxyl group can be replaced with an amino group which alters the hydrogen bonding. or in carboxylic group can be replaced with the SO_2NH_2 group which modifies the acidity and binding. So, another group of bio-isosteres is the known classical bio-isosteres. So, this group that do not have identical valency, but mimic steric electronic and hydrophobic properties.

For example, phenyl can be replaced with pyridine which modifies polarity and metabolism. Amide can be replaced with a tetrazole which improves the bioavailability and carboxylate can be replaced with a sulfonamide which reduces the enzymatic degradation. So, coming to the summary. So, the drug design leverages computational and experimental strategies to develop innovative therapeutics and the computational tools like molecular docking, pharmacophore modelling and QSAR aid in hit identification and lead optimization. Bio-isosteric replacement helps improve drug properties like solubility, stability, and bioavailability.

And the fragment-based drug design and de novo drug design methods, these are another important strategy to design novel drugs. So, I have an activity for you here. So, you shall

choose a successful drug and trace back its design strategy. How might this process have been different with today's AI technologies? So, these are some of the articles which you can refer to for further improvising your knowledge in this area. With that, thank you so much.