AI in Drug Discovery and Development

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

Week-06
Lecture-28

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about AI-driven QSAR modeling. So, by the end of this lecture, you will be able to understand the principles of QSAR modeling and AI-driven approaches. Explore machine learning and deep learning models for QSAR, and design a complete QSAR workflow from data curation to model evaluation. Interpret the model predictions, apply QSAR models across different domains, and recognize the challenges and limitations of AI in QSAR modeling. So, let us have a look at what QSAR modeling is.

So, as we have seen earlier, we can predict the properties of a series of molecules by utilizing their features. For example, in this case, we are using log P to predict the biological activity of this series of data, where log P is correlating well with the biological activity. So that we can predict the biological activity of any new molecule by using this linear correlation. So, basically, QSAR is a mathematical model that can be used to predict the biological activity of compounds, or it can be any other property as well.

Using their physicochemical properties or any other features. So, traditionally in QSAR modeling, we needed to have the compounds from structurally similar series, where we suppose that Those compounds are targeting the same site with the same mechanism of action if we are interested in determining or predicting their biological activity. But lately, we have been using QSAR modeling for various purposes, where we can have data from, you know, multiple sources as well. which can be integrated and the features can be the pattern can be learned by these advanced AI based QSAR modeling tools. And then another thing is that the enzyme inhibition assays may not directly reflect in vivo effects.

So, we also need to take that into account because sometimes compounds that show very good activity in the enzyme inhibition assays may not have a very good effect in the in vivo systems. So, usually what we do is take the historical data, which is, you know, the training data we use to make a predictive algorithm. And then we develop a QSAR model that can be used for making predictions for a new compound. So, let us take a look at some key terms and terminology related to QSAR. So, the early QSAR developed by Corwin Hansch.

So, they gave this term called the substituent hydrophobicity constant. It was a feature that

was used to predict the biological activities of the molecules. So, this pi, which is a substituent hydrophobicity constant. So, it is a quantitative descriptor that is used in QSAR to measure the hydrophobic effect of substituents related to hydrogen.

$$\pi X = \log PX - \log PH$$

So, if we have a benzene ring that is substituted with, for example, chloro, what will be the substituent hydrophobicity constant value for this chloro-substituted benzene ring This is the value of the log P value of the benzene ring alone and then the log P value of this substituted benzene ring subtracted by the log P value of the benzene ring alone with the hydrogen in place of that chloro substitution.

So, that will be the substitution that will be the substituent hydrophobicity constant of the chloro group substitution on the benzene ring. So, this pi value was only, you know, determined experimentally, and it was only for the substitution on the benzene ring. So, a positive value of pi implied that substituents are more hydrophobic than hydrogen, and a negative value of pi implied that the substituents are less hydrophobic than hydrogen. So, likewise, they had this, you know, Hammett substituent constant, which is denoted as sigma x. So, it is a quantitative descriptor used in linear free energy relationships to describe the electronic effects of substituents in organic molecules.

Where it could be calculated, the sigma x could be calculated by determining the values of log kx minus log kh divided by the reaction constant.

$$\sigma_X = \frac{\log K_X - \log K_H}{\rho}$$

So, where kx is the equilibrium constant or reaction rate for a substituted compound with substituent x, and kh is the equilibrium constant or reaction rate for the parent compound without substitution. And, this is a reaction constant, actually, which depends on the reaction type and conditions. So, the sigma, if the sigma is greater than 0, means for electron-withdrawing substituents like nitro, cyano, CF3, and COOH. And the sigma value is, you know, less than 0 for electron-donating substituents like OH, NH2, OCH3, and CH3.

Then, another feature is the Taft Steric Factor. So, which is a quantitative descriptor used to measure the steric or spatial hindrance of a substituent in a molecule.

$$ES = \log kx - \log ko$$

Where Es is equal to log kx minus log ko, where kx represents the rate of hydrolysis of a substituted ester and ko represents the rate of hydrolysis of the parent ester. If the Es value is less than 0, then the substituent increases steric hindrance, making the reactions slower; this means bulky groups like tertiary butyl or isopropyl have, you know, an Es value of less than 0. And if the Es value is greater than 0, the substituents reduce steric hindrance, making the reaction faster. And then if the Es value is equal to 0, then the substituents have little or no steric effect compared to the hydrogen.

And then another feature was, you know, molecular reflectivity, which measures a substituent's volume. So, it has this correlation factor for polarization, and then molecular weight divided by the density defines the volume of a molecule.

$$MR = \underbrace{\frac{(n^2 - 1)}{(n^2 - 2)}}_{\text{Correction factor for polarisation (n=index of refraction)}} \times \underbrace{\frac{\text{mol. wt.}}{\text{density}}}_{\text{Defines volume}}$$

So, in the Hansch equation, what did the Corwin Hansch guy, who did seminal work on QSAR, say? So, what he deduced was that he correlated the biological activity of the compound with its physicochemical properties, such as lipophilicity, electronic effects, and steric factors. So, you can see here, for example, the log 1 by c, which is the biological negative logarithm of biological activity. It was correlating with the log P sigma steric factor, and these k1, k2, k3, and k4 are the constants.

So, it was linear in the case of, you know, the hydrophobicity spreading to a small range;

$$\log\left(\frac{1}{C}\right) = k_1 \log P + k_2\sigma + k_3 E_s + k_4$$

however, if the hydrophobicity values are, you know, spreading to a large range. So, then he observed that the relationship is parabolic in nature.

$$\log\left(\frac{1}{C}\right) = -k_1(\log P)^2 + k_2 \log P + k_3\sigma + k_4 E_s + k_5$$

So then, a Free Wilson approach was another approach that directly correlates structural modifications with biological activity without explicitly considering the physicochemical properties. So, for example, if we do not have any, you know, all those experimentally

calculated parameters like the hydrophobic substituent hydrophobicity coefficient or Taft static parameters. So, what we could do is instead of assigning numerical values to different substituents and determining their contribution to biological activity through statistical regression analysis.

$$\text{Activity} = k_1X_1 + k_2X_2 + \ldots k_nX_n + Z$$

An equation is derived that relates biological activity to the presence or absence of particular substituents. So, what we could have done is like what we can do in free Wilson analysis is have this variable called an indicator variable, which is represented by x_n and given a value of either 0 or 1 depending on whether the substituent is present or not. For example, in this case, we can say it is like K_1X_1. So, this X_1 represents the presence or absence of a chloro group in the fourth position of the benzene ring. So, if it is present, we can keep the value of X_1 as 1, and if the chloro group is not present, then we can keep the value of X_1 as 0.

So, Z is a constant representing the overall activity of the structure studied, and the contribution of each substituent to activity is determined by the value of Kn. So, let us have a look at the overall QSAR workflow, where the first step is data collection, during which we collect the experimental activity data and the chemical structures. The next step is the descriptor calculation or featurization, where we convert these chemical structures into features such as molecular descriptors or fingerprints. The next step is model building and validation. So, we perform the internal validation known as cross-validation when we have built a model.

Then, that model is used for prediction. So, we apply the model to new compounds to predict their properties. So, let us see them one by one, step by step. So, the first step is data selection and curation. So, there are multiple sources from which we can get the data.

So, there are, you know, for example, ChEMBL; ChEMBL is one of the databases where you can get the biological activity associated with the chemical structures. Then you have PubChem, which hosts millions of molecules and their structures. We have the zinc dataset, which again has billions of molecular structures in it. We have the drug bank, which contains the structure and activity details of all the molecules that are either FDA approved or are experimental drugs in clinical trials. And then we have to, you know, ensure the data quality because sometimes what we face is that when we download the data from ChEMBL, for example, it might contain several duplicate compounds.

So, we need to remove the duplicates, the outliers, or the inconsistent data as well. And then we need to balance the data set; we need to avoid overrepresentation of certain classes. The data set, you know, is diverse, and it represents all the structural classes equally. So, it is not like a biased data set; we need to take care of that as well. Downloaded and curated the data.

So, there are other methods as well; sometimes we generate data on our own, like in our lab we synthesize molecules, we determine the activity, and we use that data to make the models. So, the next step is the extraction of molecular representation features. So, as I said, there are multiple kinds of features that can be used. So, the first one is the descriptors, which can be the physicochemical properties like log P, molecular weight, and topological indices, and then we can use the fingerprints as well, like ECFP, MaCCS Key, and Morgan fingerprint, which are being used extensively for molecular similarity. Then we can use the graph-based representation where deep learning methods, like graph neural nets, can be used to obtain the graph-based representation.

After featurizing, the next step is feature selection and pre-processing because many times when we calculate the descriptors, those descriptors can actually be numerous. For example, when we use Moodrad or Paddle those python libraries for descriptor calculations, so we see that they can calculate thousands of descriptors. So, now how do you know to select a handful of features from those thousands of descriptors that is a big question. So, for that, we use some methods like dimensionality reduction methods, such as principal component analysis, which is a dimensionality reduction technique that transforms correlated descriptors into uncorrelated principal components. So, it helps us to retain maximum variance while reducing the number of input variables.

We can also use correlation analysis. So, what we do is we correlate among the descriptors and identify the highly correlated descriptors. And then we remove those redundant descriptors with very high correlation, where those descriptors are showing a Pearson correlation greater than 0.9. And then we avoid them, we filter them out, and then the remaining descriptors we use for building the QSAR model.

We can also use, you know, feature importance-based selection, where we can use, for example, random forest, which uses decision trees to rank descriptors based on their importance. Or we can use the lasso, which is the least absolute shrinkage and selection operator, that penalizes less important features by shrinking their coefficients to zero. By using mutual information, we measure the dependency between variables to select the most relevant descriptors. And this is a crucial step in QSAR modeling to ensure relevant descriptors are used while minimizing the noise, because if we are using 10,000 descriptors in a single QSAR model, that is actually very difficult. So, the next step in model selection

is determining which algorithm and method we are going to use.

So, there are multiple, you know, possibilities ranging from machine learning to deep learning approaches. And then you have the traditional ML approaches like random forest, SVM, XGBoost, partial least squares, or we can use the AI-based methods like deep learning, ANN, CNN, RNN, GNN, or auto QSAR tools. And then we can also use a hybrid approach where we combine the ML with the mechanistic models. So, once we have selected an algorithm for making a model. So, the next step is training and validation, where what we have to do is split the data, as we discussed earlier.

We have to split the data into training, validation, and test sets that usually can contain 70 to 80 percent of the data in the training set, 10 to 15 percent in the validation set, and 10 to 15 percent in the test set. And then we can use it for cross-validation. So, we can use either k-fold cross-validation, where we split the data into k subsets, with k minus 1 for training and 1 for validation. Or we can use a leave-one-out cross-validation method, which is also known as LOOCV, where each sample is used once for testing, and it is usually a good approach for small data sets. And during the training and validation, we need to tune the hyperparameters as well, where we can optimize the learning rates.

The number of hidden layers and the decision tree depth are especially important when we are using advanced ML and AI methods for making the QSAR model. Okay, once we have made a model. So, how do we evaluate that? So, we have some performance evaluation metrics for that; the first thing we use is the R-squared, which is a correlation coefficient that measures how well the predictions match the actual value. And then we have the ROC receiver operating characteristic curve under the area under the curve. So, it assesses the classification performance by balancing sensitivity and specificity.

You have the MAE, mean absolute error, which measures the average absolute deviation from the three values. We have the RMSE, which is the root mean squared error that quantifies the average prediction error. So, based on all these performance evaluation metrics, we evaluate how good a model is using the evaluation metrics. So, let us have a look at the next step, which is to ensure the interpretability and explainability of the QSAR models. So, since AI-driven QSAR models are often black boxes in nature.

So, their explanations are essential. So, we can use this feature importance analysis like SHAP, which stands for Shapley Additive Explanations, that assigns contribution values to the descriptors. Like which of those features or descriptors are contributing to that specific activity, we can use it. And then we have LIME, which stands for Local Interpretable Model-agnostic Explanations, and it creates surrogate models to explain the predictions. And then we can use the permutation importance, which assesses how

shuffling a feature impacts the model's performance. So, by using all these methods, what we ensure is that the model is interpretable, meaning we can know which features are important for the activity, and that is explainable as well.

And in addition to that, we also need to do the applicability domain analysis, like where we can apply those QSAR models. Because suppose we are making a QSAR model with 1000 compounds for predicting solubility and we have a data set of 1000 compounds. So, we cannot use that model for predicting the solubility of every other compound that exists in this universe. Because that model will not be applicable to every other compound, it is called, as you know, the applicability domain. So, that also we need to consider in in while building the QSAR models.

And then the next step is model deployment and application. So, these AI-powered QSAR models have various real-world applications, such as virtual screening, toxicity prediction, and drug repurposing, by deploying these QSAR models. Looking at the traditional QSAR models, these models model the relationship between chemical structures and biological activity using a mathematical equation. They were largely using linear regression, multiple linear regression, partial least squares method, or a 3D QSAR technique called comparative molecular field analysis.

So, let us take a look at the 3D QSAR. So, early QSAR models were based on the 2D structures of the compounds, and they did not take into account the three-dimensional structure, while it is the three-dimensional structure that binds to the drug target and has the biological activity. Can we include that three-dimensional information into the model, and can we improve the predictability of that model? You know the properties of the predictive ability of that model by using this 3D information. So, that was the idea of using 3D QSAR models. So, these 3D QSAR models help analyze the spatial arrangement of molecular features influencing biological activity using key methodologies like Comparative molecular field analysis and comparative molecular similarity indices analysis, which are named COMFA and COMSIA. So, the key features were that the physical properties were measured for the entire molecule instead of using the properties of the substituents.

Like, if you remember in the Hansch analysis, we were talking about the substituent hydrophobicity constant, while here we are measuring the property of the entire molecule. And then, properties are computed using the specialized software. So, we are not relying on the experimentally derived, you know, features like hydrophobicity, substituent, and substituent hydrophobicity constant while we are determining these properties computationally by using the software. And then no reliance on experimental constants, direct measurements, as well as molecular properties is represented as fields instead of their

individual values. And then, in this case, we were using three different fields: one is the steric field, which defines the size and shape of the molecule.

We have the electrostatic field, which highlights electron-rich and electron-poor regions, and the hydrophobic properties, which are considered less significant in most cases. So, some of the advantages of 3D QSAR models are that we do not need any dependence, we do not need any experimental data for the features; actually, it is not the dependent variable. So, it is the features. So, we do not need any experimental feature data, and it is effective for molecules with unusual substituents as well. If you remember, for normal QSAR we needed to have molecules from similar series acting through the same mechanism of action, but this was an advantage of 3D QSAR where we can use molecules from different, you know, series as well.

It can be applied across diverse structural classes, and it has a strong predictive capability for drug activity. So, let us have a look at how the 3D QSAR workflow was done. So, for example, we have this molecule, and we build a three-dimensional model of it. We generate a minimum energy conformation for this molecule, and we believe that this minimum energy conformation is the active conformation. So, now that we have the active conformation, we assume that this molecule binds in this conformation to the target, to the receptor, to the enzyme, to the ion channel, elicits the response, and gives the biological activity.

So, by using this active conformation, we then determine the pharmacophores, as we have seen earlier, that it is nothing. but you know the three-dimensional spatial arrangement of those features that are responsible for the activity. So, now we have obtained a pharmacophore from this active structure, and then we place this pharmacophore into a lattice of grid points. Where you can see that the intersection of these lines is a grid point, and then we place the pharmacophore inside it, and then we align all those molecules on this pharmacophore. So, we position the molecule to match the pharmacophore in this grid space, and then at each lattice point, at each intersection of these lines in this grid space, we place a probe atom.

So, they are like different probe atoms for steric and electrostatic fields. So, we use those probe atoms and then we calculate, you know, the contribution of, for example, the interaction of this probe atom with the atoms of the molecules, actually. For example, if there is any functional group that is hydrophobic in nature and if we have a probe atom that is, you know, determining the hydrophobic contribution. So, this probe atom will get some, you know, data, and then for each lattice point where we have placed the probe atoms. We get them into this, you know, a table where each lattice point has got a column for it, and then we add all those, you know, contributions from all these steric and electrostatic fields.

And then, by using these as features, we create a prediction model using the partial least squares method, and then we formulate a QSAR equation something like this. Where activity correlates with feature 1, feature 2, and then feature n, as well as the constants represented by, you know, for example, z. So, once we get this model, what we can see is that we now have those steric and electrostatic fields, which are correlating with the activity of this molecule. So, we get a map of, you know, steric and electrostatic properties which tells us if we put steric or electrostatic groups in that area, whether it will be good for the activity or not. And this is how we can use the 3D QSAR for many purposes, such as lead optimization or even virtual screening as well.

Okay, so let us see what is important. For making QSAR models, one of the biggest things is that we have to distinguish correlation from causation. and knowing when we have enough training examples to generate a model that makes accurate prediction for a new cases. In this example, you can see that dry, hot, sunny summer weather leads to sunburn, and this also leads to an increase in the sales of ice cream. So, there is a direct causation that an increase in the temperature due to this dry, hot, sunny summer weather leads to an increase in ice cream sales and also an increase in sunburn. So, these are, you know, causation; there is a direct causation between these things, and they are correlating well, these events.

However, there is a very strong correlation between the increase in cases of sunburn and the increase in ice cream sales. There is a strong correlation between both of these events; however, there is no causation. So, we need to, for making a QSAR model, have a look at causation, not correlation, because mere correlation will not help us in designing or improving those compounds' biological activity. So, then talking about the AI approaches for QSAR modeling. So, there are machine learning-based QSAR approaches that use descriptors to establish a relationship between molecular structure and biological activity.

The key regression models we use are random forest, support vector machine, or XGBoost. And then these are some of the tools that are being used widely for making QSAR models, like scikit-learn, which is a Python library. So, it's a widely used ML library in Python, offering algorithms for regression, classification, clustering, and feature selection. So, it supports key ML techniques like Random Forest, SVM, and XGBoost for QSAR modeling and requires integration with cheminformatics tools like RDKit to compute molecular descriptors, fingerprints and structural features. So, RDKit is another tool that is used extensively for cheminformatics analysis, especially for, you know, featurization of molecules, where you can calculate a lot of descriptors by using this tool.

And then there is another tool called Weka, which is an open-source machine learning

platform with a graphical user interface for easy model development. It offers built-in ML algorithms for classification, regression, clustering, and feature selection, making it useful for QSAR analysis. And it also supports cross validation and statistical evaluation of QSAR models. And then there is a commercial tool, Schrodinger, which has a tool named AutoQSAR, a proprietary tool designed by Schrodinger, especially for automated QSAR model generation. So, it automates descriptor calculation, model selection, training and validation, optimizing the QSAR workflow.

And it uses machine learning techniques such as random forest, support vector regression, and partial least squares. And then there are deep learning tools, unlike traditional machine learning, which relies on handcrafted molecular descriptors. These deep learning methods automatically extract features from the molecular structures, improving prediction accuracy. So, some of these architectures being used in QSAR are convolutional neural networks, recurrent neural networks, and autoencoders. So, where CNNs, specialized for spatial feature extraction, have been adapted for molecular graph-based QSAR and in graph convolutional neural networks.

So, the CNNs process molecular graphs, capturing substructure patterns and atomic relationships. The CNNs can also process 2D molecular fingerprints and image-based representations of molecules that can be used for molecular docking purposes as well. And then we have the RNN, which excels at handling sequential molecular data, particularly SMILES strings and variants like long short-term memory potential. And gated recurrent units address vanishing gradient issues and improve long-range sequence learning. They are useful for predicting bioactivity and designing novel molecules by generating optimized SMILES strings.

And then the autoencoders, which are unsupervised neural networks, encode high-dimensional molecular descriptors into a lower-dimensional latent space. And then you have the VAEs (variational autoencoders) and DAEs (denoising autoencoders); these are used for QSAR tasks like feature extraction, data denoising, and molecular generation. And then we have VAEs, which enable generative QSAR modeling as well, facilitating de novo design with the desired biological properties. So, these are some of the tools like DeepChem, which supports multiple deep learning methods, and it offers pre-trained QSAR models, molecular featurization tools, and custom model training. and it integrates seamlessly with the Tensorflow and Pytorch making it a flexible framework for ligand based virtual screening, ADMET prediction and activity clip modeling.

And then we have ChemProp, which is another PyTorch-based deep learning framework specialized in molecular property prediction and QSAR modeling. It implements message-passing neural networks, a class of GNN that can learn molecular structure representation

directly from graphs. And it captures in intricate molecular relationships without relying on predefined descriptors improving predictive performance. And it supports multitask learning, enhancing generalization across multiple QSAR tasks: solubility, toxicity, and biological bioactivity prediction. And then we have DeepMol, which is an ML and DL pipeline designed for QSAR modeling, cheminformatics, and molecular property prediction.

It provides advanced molecular featurization techniques, including fingerprint-based, descriptor-based, graph-based, and embedding-based features. It supports scikit-learn, DeepChem, and TensorFlow, enabling flexible model training and validation, and also facilitates high-throughput virtual screening, lead identification, and bioactivity prediction. So, this is, you know, our DeepMol workflow. How does it work? So, we have this AutoML, where this optimization framework is present and the configuration space exists. So, the model is being trained on the features and then it is being optimized through AutoML optimization.

So where the training data set is being used for the model training. And then further, this generated model is being used for predicting the properties of the new compounds or the test set, which can be used for virtual screening and for making new predictions. And then we have the graph neural networks for QSAR. So, this is a powerful approach for effectively capturing the topology and connectivity of molecular structures. So, instead of treating molecules as linear fingerprints or sequences, GNN represents them as a graph where the atoms are represented as nodes and the chemical bonds are represented as edges. So, the GNN learns molecular features through a message-passing framework where information propagates between atoms and bonds, enabling a more context-aware molecular representation.

So some of the key tools that use the GNN are MolGraph, which is a Python-based framework for building graph-based molecular machine learning models and supports custom GNN architecture for QSAR modeling. And then we have ChemProp as well, which is a deep learning framework optimized for message-passing neural nets and excels in property prediction, toxicity assessment, and ADMET profiling. So, this is an overall architecture of, you know, ChemProp, where you can see that the molecules are from the SMILES. They are being converted into graphs, and then those graphs are being used to make models that can predict the properties of a molecule.

And then we have the transformer-based models for doing QSAR analysis. So, these transformer models originally developed for natural language processing have been adapted for QSAR modeling by treating molecular representations like SMILES or selfies as sequences, enabling models to learn atomic dependencies. So, these models they use

self attention mechanism to extract molecular features eliminating the need of predefined descriptors. So, in this case, you do not need to calculate the features of the molecules. So, these transformer-based models directly featurize the molecules based on their SMILES string.

So, what we need is only the smile string of those molecules. Some of the tools for transformer-based QSAR are like Camberta. So, which is a model trained on smile sequences for predicting bioactivity, toxicity, and ADMET properties. So, it uses self supervised learning to generate context aware molecular embedding improving the QSAR predictions and it demonstrates superior performance over the traditional fingerprint-based ML models. Okay, so now once we have built all those QSAR models using either the classical QSAR techniques, ML-based techniques, or advanced transformer-based models. So, the next step is the benchmarking, actually, because benchmarking is essential for evaluating the performance of AI-driven QSAR models, ensuring generalizability and reliability across different chemical data sets.

So, as we discussed earlier, we need to make sure that the models we have developed are general, so they can be used for predicting the properties of any other molecule, and that they are reliable. So, there are standard benchmark datasets that provide diverse molecular structures, well-characterized bioactivity data, and standardized evaluation matrices. So, we have the key benchmark data sets, such as ChEMBL, which is a large-scale bioactivity database with curated experimental data from Medichem literature and is widely used for training QSAR and deep learning models. And then we have TOCS21, which is a dataset from the Toxicology in the 21st Century initiative and contains toxicity profiles for thousands of compounds tested across nuclear receptor and stress response pathways. And then we have MoleculeNet, which is a benchmark suite for evaluating ML and DL models in molecular property prediction, covering diverse QSAR tasks, including toxicity, solubility, and bioactivity.

So, why is benchmarking important in QSAR? It ensures model reproducibility and comparability across studies. It also provides a real-world chemical diversity for robust model training, as it enables fair performance assessment across different AI architectures. Okay, so if we just have a look at the traditional and AI-driven QSAR models. So, if we talk about the modeling approach which they have been using. So traditionally, the linear regression and partial least squares methods were used for making models, and then in AI-driven QSAR modeling, we have been using neural networks, deep learning, or even transformer models.

And then descriptor selection in traditional QSAR was either manual or expert-driven, or we had to, you know, provide the descriptors or features to the model. While in AI-driven

modeling it is highly automated, and sometimes, especially in the transformer-based models, it learns the features directly from the smile string of the chemical structure. We do not need to explicitly calculate those features and provide them to the model. And then we talk about the data requirements.

We can, you know, work with the small to moderate data sets in traditional QSAR. However, if we have a large data set, such as millions of compounds, we can use AI-driven methods, and we also have diverse data sets for optimal performance in the case of AI-driven QSAR modeling. So, based on the data used to train the model, the applicability domain for traditional QSAR is narrow. Because you know that if you are making a QSAR model on 100 or 1000 compounds, it cannot be used for predicting the properties of any other large number of molecules, like millions of compounds, you cannot use that. But in the case of AI-driven models, since we are using a large training data set, the applicability domain is wider; however, the applicability domain must still be well defined even for the AI-based QSAR models. And regarding speed and efficiency, those traditional QSAR models are slower, especially with the high-dimensional data.

However, the AI-driven QSAR models are pretty fast. And they can handle large and complex datasets as well. And the interpretability for traditional QSAR models was high because we could directly see that if logP is contributing to the activity. Means lipophilicity is the major contributor to the activity; increasing lipophilicity in a molecule will lead to increased activity. However, in the case of AI-based models, it is often low because, as I said, AI works like a black box. However, we have some explainable AI methods that, you know, are currently evolving, and in the future, those will be available.

Where we can use those explainable AI models to mark which of those features are important for the activity. Also, there are some regulatory considerations for AI-based QSAR models, and these are the basic five principles of OECD guidelines for QSAR modeling. So, the first thing is that we should have a defined endpoint. So, models must predict a well characterized measurable effect. So, it should not be an ambiguous, you know, biological activity, and then the second thing is unambiguous algorithms.

So, we need to have a clear, reproducible modeling procedure. So that the data can be reproduced and used for, you know, predictive modeling. And then the third thing is the applicability domain, where we need to specify the chemical space for valid predictions again. As we cannot use a model that is trained on a very small data set to predict the properties of a large number of molecules. And then those models must have appropriate measures of goodness of fit, robustness, and predictivity. So, we need to use internal and external validation to assess the model quality and then provide mechanistic interpretation if possible.

So, it should explain the biological or chemical rationale behind the predictions; for example, it should indicate if hydrophobicity is the key parameter that correlates well with the biological activity. So, how can hydrophobicity be explained as a key contributor to the activity? And then we already talked about it, but some of the applications of QSAR modeling are in drug design and optimization, in predicting the ADMET properties. In determining the environmental risk of the chemical substances and the chemical risk and safety evaluation of the new compounds. So, if we talk about the QSAR-assisted drug design pipeline, you can see here we start with the data collection, curation, and integration.

So that data can come from either electronic databases, lab notebooks, or literature data. after getting all this data curating this data cleaning this data. So, this goes into the data repository. So, a predictive QSAR model can be formed using various AI and ML-based techniques, which can then be used for virtual screening and molecular design experiments. And, once we have identified some hit molecules, those shall be experimentally validated, and those experimentally validated hit compounds will contribute to the buildup of this data again. So, it is a kind of every time we run it, we contribute to the enrichment of this database by adding more labeled compounds to it.

And then in organic synthesis in synthetic organic chemistry, we can actually use it for making decisions. For example, we have a target molecule to synthesize in the lab. So, we can use QSAR models to determine synthetic feasibility; is it easy to synthesize? We can determine that. And we can also use them for, you know, synthetic planning as well as how to make those models, and then once we have done that. So, we start with the starting material, and then we generate the target molecule. Again, we can decide how to proceed with the optimal reaction conditions; we can optimize the synthetic scheme, synthetic process, or synthesis procedure as well.

And then we can also optimize the reaction outcome, kinetics, and thermodynamic parameters by using all those QSAR models. And then we use the AI-driven QSAR for ADMET predictions. So, for example, there are several tools. One of them is ADMETLAB, PKCSM.

And then we have this DruMap as well, developed by Professor Kenji Mizuguchi at Osaka University in Japan. So, in the ADMETLAB, you can predict all those ADMET properties. And then there are multiple properties that can be predicted using those. And all these predictions are made by the QSAR modeling, you know. Okay, coming to the challenges and limitations of AI in QSAR modelling, the first challenge is data heterogeneity, where the biological data varies by species or protocols, and the physicochemical data depends on conditions like solubility at specific pH.

And then we have inadequate or incorrect data. So, there are errors in, you know, chemical names, cast numbers, or structures. For example, chloroxylenol has 18 isomers, and 4-chlorocresol has 2. So, which of those isomers has been tested, evaluated, or named? So, all this, you know, we need to correct that incorrect data as well. And then there is a replication issue, like improper desalting, that can cause duplicate entries.

For example, we have a molecule with salt. So, if we are building a QSAR model, we have to remove that salt. So, if we have removed that salt, those two molecules will now be different molecules. So, if we are keeping both of them, that will lead to a duplicate molecule that will cause problems in the QSAR model building. And then narrow the endpoint range. So, sometimes, you know, we are using the biological activity, which is very narrow in range, very narrow in range.

So, that is a problem as well. And confounded descriptors where we can have the collinear descriptors that add redundant information and complicate the interpretation of those models. And non-interpretable descriptors are another challenge where we have to use the descriptors that explain the mechanism of action only. And then, there might be descriptor errors; for example, there may be measured or calculated values that vary.

For example, 4-nitrophenol has been reported to have logP values ranging from 0.76 to 2.8, which are measured, and then from 1.35 to 1.93, which are calculated. So, which of them to consider that is again a you know a challenge in building a QSAR model. And the lack of auto scaling, where auto scaling helps assess each descriptor's contribution to the model, leads to overfitting, which is another issue if we are using a 5 to 1 ratio of training compounds to descriptors.

So, that is the optimal, you know, the ideal situation, and for non-linear methods like ANN, they have a kind of very high risk of overfitting the models. Then, using excessive descriptors, use the smallest number of descriptors for a robust model where, for example, the molecular length outperformed COMFA for fiber affinity. So, excessive descriptors, for example, we have like a thousand descriptors. So, how do you make sure which of those descriptors are reliable and which ones we can use? And then, missing or misused statistics, like we need to report the R, R squared, adjusted R squared, Q squared, and standard error, are sometimes actually missing, so how can we ensure that those models are reliable? And then we have the calculation errors as well, such as large-scale incorrect calculations that have been reported in the literature.

So, how to get rid of those is again a challenge, ignoring residual distribution. So those random errors can be minimized through careful property selection. And then we need to

plot a residual plot that reveals the systematic error if the residuals are not, you know, equally distributed in the plot. So, then there are some errors which we shall evaluate and rectify. And then, poor QSAR transferability, where the lateral validation ensures model reproducibility and reliable predictions, means that we can only use that model for predicting the properties of new compounds. Undefined applicability domains are another challenge where the models must predict reliably within the chemical or response space of the training data.

And then inadequate training on oblique test set data is, you know, another challenge where we need to have at least a meaningful handful of data for making these models. An incomplete model validation where proper internal and external validation is essential for reliability, and a lack of mechanistic interpretation where the descriptor should align with known mechanisms supported by literature references. And the computational cost, especially for those deep learning models, is computationally very expensive. So, training those complex models with large data sets can be computationally intensive and resource-demanding due to the black box nature of AI models.

So, those many AI models lack interpretability, making it difficult to understand the underlying molecular mechanisms. So, these are after, you know, the challenges. So, these are some of the future directions. So, I think in the future we will see explainable AI at work in QSAR modeling and also multimodal learning, where it can integrate, you know, multi-omics data into QSAR modeling. And then active learning, where we can prioritize the most informative compounds for testing. and the diffusion models that can simulate molecular transformers for more accurate predictions, and the federated learning in which we can train models across institutions without sharing sensitive data.

And this is especially useful when we are working in industry-academia collaboration because industries are usually very, you know, restrictive about sharing their data. So, federated learning can come into play in this case. Okay, let's move on to the summary. So, the QSAR models predict biological activity using molecular descriptors like logP, steric factors, and electronic effects, and AI empowers QSAR accuracy through machine learning and deep learning.

And techniques like SHAP and LIME help interpret models and understand the feature contributions. And we need to address challenges like data quality, overfitting, and limited applicability with future solutions in explainable AI, multimodal learning, and federated learning to expand QSAR's potential for drug discovery and development. And with that, I have an open question for you: what if we could build a QSAR model that evolves like a living system, which is learning and adapting from real-time experimental feedback? Just

think about it; I have some suggested readings that you can look at to learn more about this topic. And with that, thank you.