**UNIT II**

**DRUG DISCOVERY AND MOLECULAR MODELING**

Introduction - The scope of artificial intelligence in drug discovery - Types of machine learning in artificial intelligence - Molecular modelling and databases in AI for drug molecules - ML methods in molecular modelling- Drug characterization - Drug design for neuroreceptors using ANN techniques - Use of deep learning in drug design

* 1. **Introduction - The scope of artificial intelligence in drug discovery**
* Pharmaceutical drug discovery requires **handling and analysing large databases of chemical compounds** and therefore means to scan over a huge amount of chemical information. This can be achieved by using **machine learning techniques (MLTs)** that can operate and reach their goal in a relatively short time.
* In drug discovery, what matters is not only **processing a large amount of data in a short time** but also an ability to **correlate or associate different data to molecular structure and/or properties**.
* These tasks can be done by MLTs, for example, artificial neural networks (ANNs), that greatly help to develop design of drugs.
* ML is a necessary tool for handling a large amount of chemical data in drug development.
* However, the type of data related to molecular structures and function that is needed for drug development is often highly nonlinear in nature when it comes to assessment and prediction of drug effects.
* This means that a little change in the molecular structure of the drug can change the effect of the drug dramatically.
* This **nonlinearity** is also typical for assessment of what mushrooms that are edible or lethal although having almost identical structure.
* This is opposed to linear incrementation/extrapolation or linear regression methods This is opposed to linear incrementation/extrapolation or linear regression methods, where a little change in the structure is followed by a small effect unlike the effect of being alive or dead.
* The nonlinearity is an ability that can exist in some MLT, for example, in the synapse function of neural networks.
* Still the most obvious reason for making use of ML or any other automated computational method in biotechnology or drug development is the ability to scan large datasets in a short amount of time and pick out a particular number or pattern that is optimal with respect to a certain performance.

**1.1.1 Areas in which machine learning techniques are applied in biotechnology**

* The areas of application of MLT in Biotech are typically

(1) screening of large libraries of ligands

(2) virtual screening or high throughput screening

\*-

(3) structure of potential ligands and their pharmacophoric values as a drug

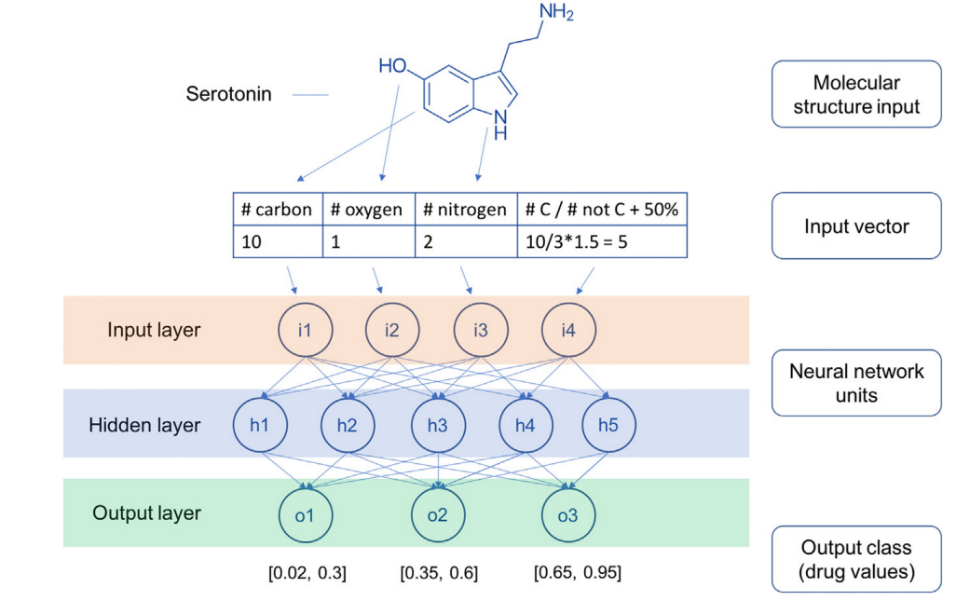
(4) quantitative structure-activity relationship (QSAR) studies

(5) Biological functionality of drug receptors and docking of drugs to them aided by molecular dynamics (MD) simulations

(6) expected toxicity properties of given drugs

In all these areas, especially ANN is a valuable tool often together with MD programs.

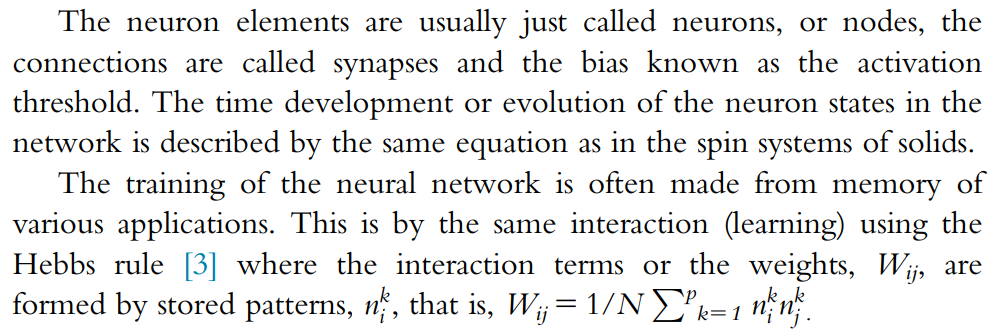
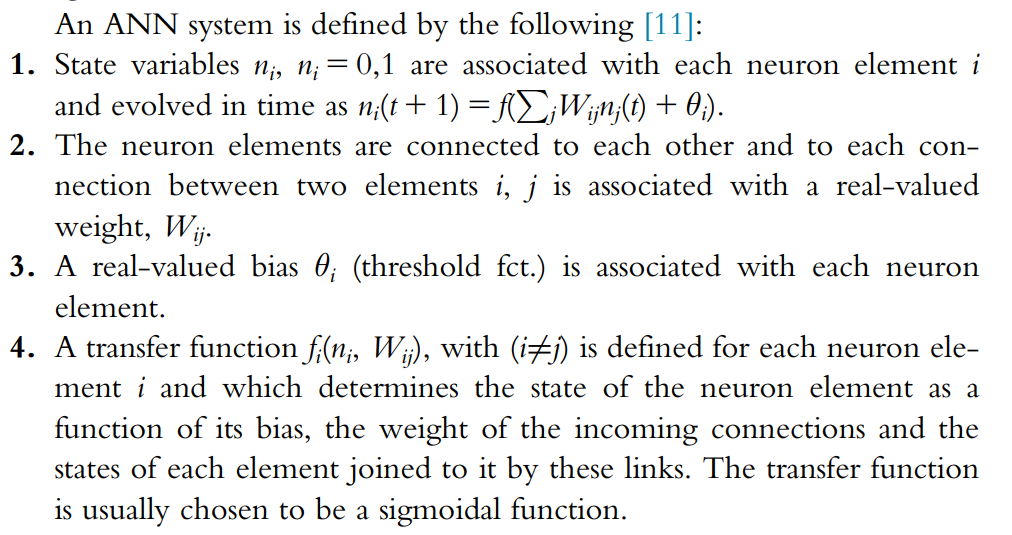
* An example of about docking of a drug molecule:

The main aim of drug docking to a receptor molecule is to be able to see how well a given drug molecule fits into a specific receptor. In terms of a quantitative measure, the aim is to be able to predict the binding affinity of that drug to a given receptor. In such a search, it is implicit that the most effective drug for a given receptor is the one that binds most strongly to the receptor. However, there are cases where a drug that binds poorly to a receptor can have a strong effect as a medication and vice versa.

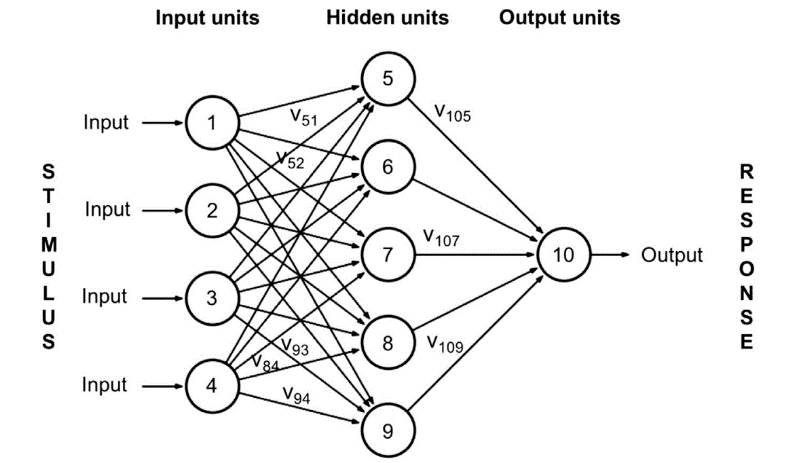
* The data for this are in the first row distributed into categories and assembled as an input vector. This is then coded in the input layer of a neural network shown with neurons in three layers where the hidden layer is next and with the output layer in the bottom.
* There is also the case of the drug Der morphine, which binds 10,000 times stronger than ordinary morphine but with no beneficial effect for the consumer.
* Besides methods in drug docking, there are ML applications for other drug discovery areas such as in High-Throughput Screening, Virtual Screening, QSAR studies, and Docked Complexes analysis that all are connected to drug docking.
  1. **Types of machine learning in artificial intelligence**
* ML is here a matter of computational methods that also implemented in computer hardware to achieve high speed data processing for that particular purpose.
* The type of tasks where ML plays a large role is a task that contains complex problems such as biosystem involving a large set of data and described by a large set of variables.
* There are basically four types of ML described by learning processes

1. **Supervised learning:** Supervision often done by a person handling labelled data arranged in training set of inputs (descriptors) to be withhold with correct output values.
2. **Unsupervised learning:** Here, the data are unlabelled and there is no supervisor or teacher with a known prescription but the machine has to guess a prescription from the data patterns.
3. **Semisupervised learning:** A smaller set of labeled data are within a larger set of unlabeled data, where a pattern has to be guessed from the few data and being less costly.
4. **Reinforced learning:** This is a system with a continuous learning and aimed at using the interaction with the environment and to a given set of data.

* Common models of **algorithms** for
* Supervised learning - Bayes systems, linear regression, and neural networks, among few others. They are the most popular for handling biochemistry data.
* Unsupervised learning - Associated memory model of Kohonen, the so-called Kohonen feature map
* Semisupervised learning - similar to the ones for supervised learning have been used.
* Reinforced learning - Q-learning and deep learning (DL) networks.
* ML methods applied to drug discovery will often involve large datasets and ML is necessary for handling and surveying a large amount of data and these data are a necessity for drug development. On the other hand, MLTs can also be useful in the opposite situation of sparse datasets where machine network methods are employed.
* Here, auxiliary data with perhaps even a few data points were proposed as a variant for one-shot learning. In this case, the human brain has been an inspiration for overcoming the problem with very little data by, for example, similarity measures with multiple neural networks to infer more data. As for most cases, MLTs are strong and needed when large data are to be surveyed and processed.
  + 1. **Artificial neural networks as tools in drug discovery**
* ANNs are one of the most used methods of ML. They are modeled after the biological brain and were developed as a technique in the 1940s almost 80 years ago.
* The Associated Memory neural networks are analogous, isomorphic, and contain procedures for doing cognitive tasks such as learning.
* The basic use of ANN in the AI technology is that of **classification and generalization**.

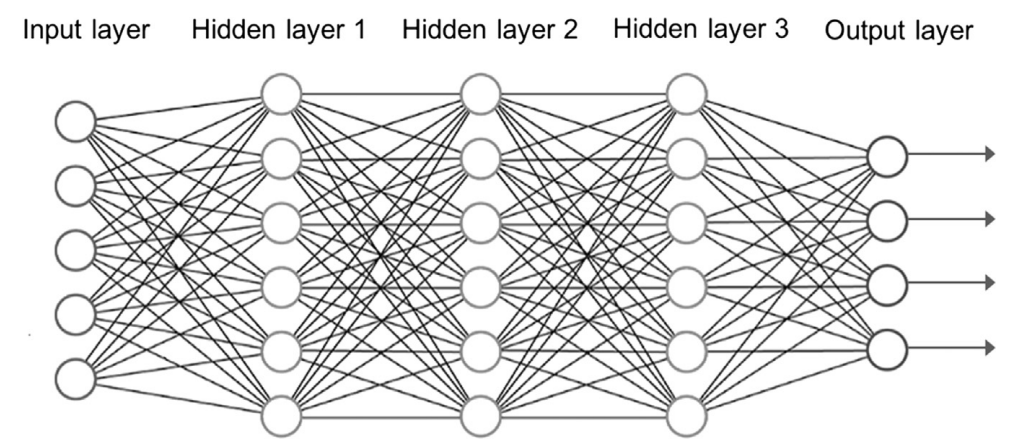


* + 1. **Architecture of artificial neural network for drug discovery applications**
* The various types of artificial neural networks (ANN) commonly used in drug discovery applications are:

1. **Fully Connected Associated Memory Network (Random Neural Network)**: All neurons are interconnected without specific input neurons, often initialized with random values. Used for modeling brain features.
2. **Hopfield Network**: Single-layer network with neurons equal to input signals, capable of autoassociations for image regeneration from corrupted data, similar to 2-D Ising spin models.
3. **Perceptron Network**: Distinct layers of neurons with inputs connected to output neurons, typically used with only two layers (input and output) and weighted connections.
4. **Multilayer Perceptron Networks**: Popular ANN with hidden layers connected to upper layers, capable of recall and extrapolation for logical problems.
5. **Recurrent Neural Networks**: Include feedback loops allowing information flow backward, optimized with fixed points but harder to train than standard perceptrons.
6. **Jury Networks**: Utilize multiple independent ANNs, with the majority result determining the output, enhancing reliability.
7. **Kohonen Feature Map Network**: Unsupervised network without a unique information stream, useful for molecular reaction classification in chemistry.

* **Multilayer perceptron networks, recurrent neural networks, jury networks, and Kohonen feature map networks** are the most employed in drug discovery.
* Training often involves **backpropagation, adjusting weight parameters to minimize error, and Hebbian learning**, a type of reinforced learning.
* ANN has been utilized in gene search within large genetic databases like GenBank.

**1.2.3 Artificial neural network methods for structure prediction in proteins from their sequence**

* One of the earliest and widely used applications of artificial neural networks (ANN) in biotechnology is training them on protein structures and sequences to predict novel structures from known sequences.
* In this approach, the **input** to the network is a **sequence** of amino acids, typically represented as a binary code, while the **output** is **structural information** about the protein molecule, such as the presence or absence of specific secondary structures like helices.
* For instance, the input sequence might be encoded using a 20-letter amino acid code, with each amino acid represented as a binary vector. The output could be a binary indicator (1 or 0) representing the presence or absence of a particular secondary structure, such as a helix.
* The most popular architecture for this application is the multilayer perceptron network, where neurons process input data using weighted sums and nonlinear functions, often sigmoidal functions. These networks are trained on known protein structures to predict novel structures from known sequences.
* Performance is evaluated based on the network's ability to correctly predict structures, with scores indicating the degree of correlation between input sequences and predicted structures. Studies have reported accuracies of up to 70%, particularly when the training and test sets contain low sequence homology.
  + 1. **Artificial neural network methods for spectroscopy in biomedicine**
* The application involves using Artificial Neural Networks (ANN) in spectroscopy, particularly for classifying tissue types using infrared images derived from infrared spectra.
* These ANNs, part of CytoSpec NeuroDeveloper 2.5, employ multilayered feedforward perceptron neural networks.
* The training process involves presenting nonhomologous input data and adjusting synaptic weights via backpropagation to minimize errors and improve prediction accuracy.
* By correlating input data with output data through training, the network can classify new infrared spectra, generating corresponding infrared images.
* To enhance classification accuracy, Hierarchical Cluster Analysis is performed on spectral ranges.
* The precision of classification for input absorption spectra is approximately 10%, while for output tissue images, it's also around 10%, corresponding to the number of peak assignments and distinct image classes, respectively. 

**1.3 Molecular modelling and databases in AI for drug molecules**

* Molecular modeling of drug molecules is usually done using MD calculations based on classical mechanics and where the objective is to find the structures of the conformational states of a drug that have optimal effects. Often there is a drug that has to fit to a given drug receptor of known structure and in that such modeling or computations are appropriately based on classical MD.
* However, there's a growing interest in obtaining electronic structures of molecules to deepen the understanding of their biochemical functions. Quantum mechanical (QM) techniques are necessary for such tasks, allowing for the calculation of electronic structures and effects of molecular orbitals. While QM calculations are challenging for larger molecular structures like drug receptors due to their electronic complexity, they are suitable for peptides and smaller oligonucleotides, providing valuable insights for drug discovery.
* Both classical MD and QM techniques are considered as machine learning (ML) tools, involving extensive computer processing akin to artificial neural networks (ANNs). These computational methods rely on system packages and programs to carry out simulations and analyses. Section 3.4 of the document likely delves deeper into the application and significance of MD and QM in molecular modeling and drug discovery.
  + 1. **Databases for the training sets in drug discovery**
* Neural networks play a significant role in biomolecular technology, leveraging **data extraction from databases** for various applications.
* In drug molecule design targeting receptors, molecular binding data are crucial for training ML systems, often sourced from databases like bindingdb.org.
* In these databases, the **chemical descriptor for ligand molecules** is typically **represented** by **molecular images**, correlating with binding strength to specific receptor molecules.
* Thus, the **input to ML systems** comprises structural **data of ligand molecules**, while the **output** data relates to **binding strength**.
  + 1. **Database mining**
* Data mining (DM) is integral to AI technology, particularly in drug development, facilitated by software tools like WEKA for analyzing large databases.
* ML and DM methods basically includes encompassing decision trees (e.g., C4.5), instance-based learning, classification and association rules, support vector machines (SVM), Bayes classifiers, K-means clustering, prediction techniques like linear regression trees, and addressing data-cleansing, boosting, or stacking challenges.
* **C4.5** employs decision trees to classify data based on object attributes, beneficial for drug molecules' bonding structures.
* **K-means** conducts cluster analysis to group similar objects, determining cluster centers based on typical members. A crucial aspect is employing a suitable metric to measure differences within clusters.
* **Support Vector Machines**, akin to C4.5 but devoid of decision trees, classify data into classes (e.g., different-colored balls) using a separating hyperplane, often in higher dimensions.
* In drug development, DM is indispensable for identifying potential drug leads for specific receptor molecules. Its impact was initially limited by sparse and small drug databases, but advancements in modern computer techniques and resources have led to expansive drug libraries, enhancing the efficacy of DM methods.

**1.4 ML methods in molecular modelling**

* ML can also involve and actually constitute computer simulation of molecular systems and in that sense molecular modeling can be considered as an ML program package.
  + 1. **Molecular dynamics simulation for drug development**
* MD methods involve solving classical equations of motion for all atoms within a molecular system, including solvents. These methods require substantial computational resources and rely on sophisticated computer program packages such as NAMD, CHARMM, and What-if.
* Comparing MD methods to artificial neural networks (ANNs), training an MD system involves adjusting the force field parameters based on structural chemistry data.
* The input consists of atomic composition of the molecules and output is then the computed molecular structures.
* The MD simulation techniques basically consist of solving Newtonian equations of motion and integration of observables in time developments.
  + 1. **Computations with quantum mechanical techniques for drug development**
* In drug development at the molecular level, machine learning (ML) algorithms incorporating quantum mechanics (QM) equations play a pivotal role in understanding and optimizing electronic properties of biomolecules such as proteins, DNA, and drug molecules.
* A typical workflow involves:

**Chemistry construction**: Utilizing programs like Spartan, molecules are constructed from sequence data.

**Quantum mechanics optimization**: Quantum mechanics equations, particularly in the Hartree-Fock (HF) scheme, are employed to optimize electronic structures and properties of atoms and drug molecules. This process involves optimizing an energy function with respect to structural parameters and electrostatic energy configurations.

**Electronic property analysis**: QM calculations within DFT or HF optimization systems yield information on molecular orbitals, electrostatic potential surfaces, and electron donor/acceptor sites. This is achieved by optimizing the energy functional based on electronic density and separating electron and nuclear motions under the Born-Oppenheimer approximation.

* For oligonucleotides like small DNA/RNA segments, quantum calculations are feasible, allowing convergence for systems of many hundreds of atoms within a few hundred CPU hours. QM HF optimization employing Gaussian wave functions is advantageous for obtaining electronic properties due to the constrained force field of oligonucleotides compared to peptides.
* A case study illustrates the significance of QM-based ML approaches in predicting molecular behavior. Specifically, the chiral structures of phosphorothioate oligonucleotides, represented as RSRRRRSR and having a sequence of nucleotides CACACTCC, are analyzed. Quantum calculations reveal distinct structural differences between all-R and all-S configurations, highlighting the importance of chiral symmetry in molecular function. Such insights, enabled by QM-based ML systems, aid in optimizing drug design for improved efficacy.
  1. **Drug characterization**
* Isopotential surfaces, crucial for analyzing molecular structure and function, are constructed based on equipotential surfaces where points possess the same electrostatic potential relative to molecular charges. These surfaces are determined numerically using an isovalue of electrostatic potential generated by a positive test charge outside the molecule.
* The process involves:

**Mapping electrostatic potential to electron density**: The electrostatic potential value is mapped to an electron density surface, aiding in understanding bonding types like H-bonding, typically with values around a few electronvolts (eV).

**Quantum computations for wave function**: Large quantum computations are required to obtain the wave function for the molecule, facilitating the determination of electric charges and electron densities.

* Isopotential surfaces play a pivotal role in various studies:

**Drug design**: Isopotential surfaces aid in understanding the electrostatic surface potential of molecules, influencing ligand docking to protein molecules and facilitating therapeutic studies. Quantum mechanics techniques are vital in constructing these surfaces and electronic structures.

**Chiral properties study**: In the case of locked nucleic acid modification of oligonucleotide molecules, isopotential surfaces help analyze chiral properties, impacting ligand docking and therapeutic investigations.

**Physical properties derivation**: For duplex DNA molecules, isopotential surfaces contribute to deriving physical properties, furthering understanding of molecular behavior.

* Additionally, isopotential surfaces are constructed for small peptides like opioid molecules, providing valuable insights into their electrostatic potential and aiding in studying particle attraction mechanisms.
  1. **Drug design for neuroreceptors using ANN techniques**
* Neuroreceptors present significant areas of application for artificial neural networks (ANNs) and other machine learning (ML) methods at the molecular level, particularly concerning ligands, often small molecules comprising fewer than 100 atoms. However, a major challenge lies in acquiring robust chemical descriptors for the ligand molecules that can effectively train ANNs.
* Studies focus on five receptors: Dopamine, Glutamate, Serotonin, Opioid, and Gaba, crucial targets for the pharmaceutical industry. For instance, the dopamine D3 receptor, implicated in various neurological disorders, serves as a template for guiding drug design due to its pharmacotherapeutic relevance.
* Successful applications of neural networks include:

**Deep Belief Networks**: Applied to drug molecules targeting neuroreceptors, achieving high correctness scores, such as 92% for serotonin Kd prediction.

**Kohonen Networks**: Predict chemical activity sites on receptor molecules, aiding in optimizing molecular substructures in drug design. They utilize unsupervised learning to analyze input data, providing insights into the chemical interaction space and aiding drug receptor relation understanding and drug design.

* While Kohonen feature maps offer informative insights into chemical interaction space, they may lack accuracy compared to quantum structure prediction in drug design. However, they remain valuable tools for developing drug design strategies tailored to specific receptors.
  1. **Use of deep learning in drug design**
* Deep learning (DL) is a subset of machine learning (ML) techniques that harnesses artificial neural networks (ANNs) with multiple fully connected hidden layers of processing units. DL algorithms have seen success in various applications due to improvements in training and avoiding overtraining and overfitting.
* Key advancements in DL include:

**Algorithms to prevent overfitting**: Techniques like dropout, dropconnect, and pruning algorithms help manage large neural networks and improve performance by preventing overtraining.

**Specialized ANN architectures**: Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) enhance DL capabilities, especially for tasks like image recognition and sequential data analysis.

* In pharmaceutical research, DL plays a crucial role in drug design, where patterns are inferred from incomplete information about molecular structures. Soft-Max regression systems are used to benchmark different ANN architectures, including feedforward and DL networks, for classification tasks.
* In drug design, DL networks process input vectors describing the chemistry of ligand molecules to predict drug effects, such as binding constants (Kd), IC50, or other relevant properties. DL networks with carefully pruned hidden neurons and Soft-Max regression are particularly effective in predicting these properties for new molecules.
* Results from DL networks, especially for predicting Kd values, outperform other ANN architectures, which compares various drug molecules' predictions for neuroreceptor binding. Notably, predicting Kd is simpler than IC50 values, which depend on multiple factors.
* Two-category networks achieve the highest correctness percentages for ligands compared to other receptor types, demonstrating the effectiveness of DL in drug design tasks.
  + 1. **Other applications of machine learning in drug development**
* to handle large libraries of chemical data to achieve properties of particular drug candidates by operating with virtual screening to obtain knowledge of chemical activity for a given molecular structure
* in medicinal chemistry molecular properties can be extracted for a given molecular compound and related to the physio-chemical bioactivity level so as to target certain properties of a molecule such as its pharmacokinetics properties, for example, absorption, excretion, and toxicity.