Talktorial 1:

Question 1:

ChEMBL has many bioactivity measurements of many compounds and protein targets. All data that is stored for the compounds aids in drug discovery. In ChEMBL data can be easily searched and filtered in order to look at binding affinities, selectivity, efficacy and ADMET properties.

Question 2:

EC50 is the concentration of a drug that gives the half-maximum response.

IC50 is the concentration of an inhibitor where the response or binding is reduced by half

Question 3:

The data extracted from ChEMBL can be used in drug discovery. All properties of compounds that are stored in ChEMBL can be usefull in determining which compounds are potential candidates as drugs for a certain target

Talktorial 2:

Question 1:

Lipinski's rule of 5 consists of:

- Molecular weight (MWT) <= 500 Da

- Number of hydrogen bond acceptors (HBAs) <= 10

- Number of hydrogen bond donors (HBD) <= 5

- Calculated LogP (octanol-water coefficient) <= 5

A high molecular weight can lead to a decreased absorption because the molecule is for example unable to pass through a membrane

In order for a drug to be metabolized, it has to bind to molecules in the body. few hydrogen bond acceptors and few hydrogen bond donors lead to less likely binding to other compounds

A LogP greater than 5 means a drug is very lipophillic. this affects absorption and distribution of the drug.

Question 2:

Clotrimazole

Question 3:

By adding the compound to the dataset

Talktorial 3:

Question1:

PAINS are compounds that give false positive results in screening. These compounds react with many targets and not necessairily with the desired target. These compounds will seem as potential drugs to use but are not usefull because they often do not target what you desire.

Question 2:

PAINS are known to react nonspecifically. If you have a target compound but not a specific site, these PAINS could potentially be usefull compounds. Some compounds could also potentially get flagged as PAINS while they are actually active. This would result in a loss of potential drug candidates because they were filtered out of the dataset.

Question 3:

The substructures are encoded with SMILES

Talktorial 4:

Question 1:

Activity Cliffs are similar compounds with great differences in potency. In order to avoid activity cliffs machine learning could be used, since there already are large amounts of data of ACs.

Question 2:

Morgan fingerprints are not limited to a specific length

MACCS keys are defined as a binary list, making it easy to compare 2 compounds

Question 3:

Different fingerprints use different measurements to define compounds. This results in different definitions and therefore differences in quantifying how similar compounds are to eachother.

Talktorial 5:

Question 1:

Clustering is grouping similar compounds. This creates clusters of drugs that share the same properties and allows you to investigate what properties are relevant in making similar compounds. This gives a better understanding of what is relevant and allows to improve further based on a cluster.

Question 2:

Jarvis Patric algorithm and Butina clustering

Jarvis Patric uses K and Kmin to determine if compounds belong in the same cluster. Similarity in the parameters determines that compounds are similar or not and determine their position.

Butina clustering creates fingerprints and compares these based on Tanimoto similarity matrixes. These create clusters with centroids that form the best compound and similar compounds to that compound are clustered around that.

Question 3:

Gaussian mixture model algorithm, Birch algorithm, Optics algorithm

Talktorial 6:

Question 1:

MCS is usefull for similarity search, hierarchical clustering, aligning molecules, and automated reaction mapping.

Question 2:

A structure is selected as query, and all other substructures are set as targets. The query substructure is converted into a SMARTS string that is based on the properties that are desired. If the SMARTS pattern is in all targets then the substructure is common. These are used to determine how common a substructure is and what the maximum common is.

Question 3:

Moving the slider to 100% shows the typical fragment of an active EGFR compound. CH3-N-CH2-CH2-CH2-CH2-CH2-CH3

Talktorial 7 (and CBR teaching 2):

Question 1:

ML can be used by creating an algorithm that learns from the data. The ML Algorithm will get better at predicting outcomes after the algorithm gets better and more data. This will result in a better and better model that can eventually be used for virtual screening since the model can predict what will work.

Question 2:

Linear regression, logistic regression, decision trees, random forest

Question 3:

A large dataset is first needed. Next to this the molecules in the dataset need to have their own unique code, and a label. Then a ML algorithm can be used in order to train a model.

Talktorial 14

Question 1:

Traditionally, one can use various fluorescence based-assays to experimentally identify pockets on protein surface, but as this talktorial has demonstrated, machine learning algorithms can be used instead

Question 2:

One advantage is that it's rapid and allows detection of binding pocket without prior knowledge about the protein. One disadvantage is that the results are heavily dependent on the training datasets.

Question 3:

Our target protein is not a kinase. As such, we are unable to check the overlap between the two.

Talktorial 15 *questions require us running the simulation, which has been established in class to not be working. So, we are moving forward to talktorial 16.*

Talktorial 16:

Question 1:

Amongst all non-covalent interaction, both hydrophobic interactions and hydrogen bond seems to be the most important ones based on frequency of interactions alone. However, given the aqueous condition that most of these protein-protein or protein-ligand interaction occurs, hydrogen bond may be more important compared to hydrophobic interaction, as presence of water molecule can further stabilize or destablize a given interaction

Question 2:

Both of them are non-covalent interactions that could transiently occur or form for a longer period of time. However, hydrogen bonds are formed between hydrogen and oxygen or nitrogen of an OH or NH2 group. Hydrophobic interactions are formed based on interactions of non-polar R groups of amino acids.

Question 3:

KLIFS is a kinase specific database, which may limit the number of dataset available.

Talktorial 19:

Question 1:

Forces between two atom (affects bond length), bond angle energy, torsion angle forces, intermolecular non-bonded forces (e.g. van der Waals forces) and electrostatic interactions (partial charges)

Question 2:

Not the exact same.

However, the ergodic postulate still holds that the time average can still be used to determine the ensemble average, which in turns determines average property of interaction

(Question 3 asks for simply trying out different simulation)

Talktorial 20

Question 1:

B. The distance is the shortest, and the angle is relatively close to 180 degree

Question 2 (*answer this question is hypothetical as we are unable to run the code*):

That means the ligand does not move and only the protein move (especially the residues on the binding pocket). As RMSF plot represents distance between two point along time, it should not change as the distance between ligand and protein would remain the same, despite having moved the focus to ligand instead of protein

(Question 3 calls for analysis of another atom upon running the simulation)

Question 4:

RMSF uses Cpptraj