**Proposal for Bayesian Docking-Score Predictor**

Created: 10/03/2023 by Tom Lever

Updated: 10/14/2023 by Tom Lever

According to Dr. Bill Basener, “the goal of the project is for you to apply Bayesian machine learning to a real dataset in an advanced way. This means that the project should show you can apply probabilistic reasoning to a nontrivial problem of your choosing.”

I propose to work with Naomi Ohashi to develop Bayesian Docking-Score Predictor. Bayesian Docking-Score Predictor will receive the Simplified Molecular Input Line Entry System (SMILES) of a ligand and provide a docking score of how well that ligand binds to a protein.

Our predictor will receive the SMILES of a ligand. A SMILES is a string representing the chemical structure of a ligand. A ligand is a molecule that binds to another molecule that is usually larger. There is one SMILES per ligand. There is one ligand per SMILES.

A SMILES of a ligand can be converted into a numerical representation of the chemical structure of the ligand. Such numerical representations include a vector of the number of occurrences of certain common substructures of molecules or a fingerprint. One substructure is a structure of carbon atoms. A Morgan-2 vector of counts is a related to a radius of . Other radii include and . A fingerprint is a vector of ’s and ’s. A vector of counts that is sparse can be folded into a vector of 1024 or 2048 counts.

A docking score is a measure of how well a ligand binds to another molecule. A docking score is a change in the molar Gibbs free energy of a compound. A docking score is measured in kilocalories per mole. According to Dr. Ryan Weil, is amazing and is okay. See Exploring Change In Molar Gibbs Free Energy below.

I propose that Data\_Frame\_Of\_SMILESs\_Docking\_Scores\_And\_Other\_Data.csv (DFSDSAOD) be our data set. DFSDSAOD is associated with one protein and with one site of that protein at which ligands bind. Each row in DFSDSAOD encapsulates data relating to a ligand docking to the protein. There is one ligand per row. There is one row per ligand. Each row has an Entry ID. There is one Entry ID per ligand. There is one ligand per Entry ID.

A feature matrix could be constructed with each row being a vector of counts. Alternately, the columns of our feature matrix could be columns of chemical properties of ligands (e.g., ten columns including molecular weight, LogP, volume, number of hydrogen-bond donors, number of hydrogen-bond acceptors).

We may train and predict with a Bayesian Additive Decision Trees (BART) model based on Python package *openbt*. Interestingly, in the original paper *BART: Bayesian Additive Regression Trees*, “BART’s many features are illustrated with a bake-off against competing methods of 42 different datasets, with a simulation experiment and on a drug discover classification problem.” According to *Statistical Learning: 8.6 Bayesian Additive Regression Trees*, “It turns out that the BART method can be viewed as a Bayesian approach to fitting an ensemble of trees: each time we randomly perturb a tree in order to fit the residuals, we are in fact drawing a new tree from posterior distribution. Furthermore, the BART algorithm can be viewed as a Markov Chain Monte Carlo procedure for fitting the BART model. We typically choose large values for [number of iterations] and [number of trees per iteration] , and a moderate value for [number of burn-in iterations] : for instance, , , and are reasonable choices. BART has been shown to have impressive out-of-box performance – that is, it performs well with minimal tuning.”

We may employ a histogram of predictions of Random-Forest models. We may train and predict with a Bayesian Linear Regression model. We may calculate a posterior probability density distribution for the probability of a docking score given a vector of counts based on a likelihood density distribution for the likelihood of a vector of counts given a docking score and a prior probability (e.g., ). Such a likelihood density distribution might be a Gaussian distribution with number of dimensions equal to number of counts. I am uncertain how to condition such a distribution on a continuous docking score.

A trained Bayesian Docking-Score Predictor could be used for developing a training data set of SMILES’s and docking scores of ligands for another predictor.

Bayesian Docking-Score Predictor could be trained on (e.g., ) pairs of SMILES and docking score and used to predict docking scores of ligands based on many (e.g., ) other SMILES’s. Rows of data representing docking of ligands could be sorted in descending order by predicted docking score. The top rows of data could be added to the training data set.

Bayesian Docking-Score Predictor may be used to calculate Bayesian predictions, estimate uncertainty, or in optimization of hyperparameters of Bayesian models.

*Exploring Change In Molar Gibbs Free Energy*

The enthalpy of a compound is the sum of the compound’s internal energy and the pressure energy of the compound. The pressure energy is the energy required to establish the compound’s physical dimensions. The pressure energy is the product of the pressure on the compound by its surroundings and the volume of the compound.

The molar internal energy of a compound is the internal energy of the compound per mole.

The molar volume is the volume of the compound per mole.

The molar pressure energy of a compound is the pressure energy of the compound per mole.

The molar enthalpy of a compound is the enthalpy of the compound per mole. The molar enthalpy of a compound is the sum of the compound’s molar internal energy and molar pressure energy.

The change in enthalpy of a reaction is the difference between the sum of the enthalpies of the products of the reaction and the sum of the enthalpies of the reactants.

The change in molar enthalpy of a reaction is the difference between the sum of the molar enthalpies of the products and the sum of the molar enthalpies of the reactants.

The entropy of a compound is a measure of uncertainty, disorder, or mixedupness of the compound. The entropy measures the degree to which the probability of the compound being in a particular microstate is spread out over different microstates. A microstate specifies all molecular details about the system including the position and velocity of every molecule. The more such states are available to the compound with appreciable probability, the greater the entropy.

is the probability that the compound is in the th state according to the Boltzmann distribution.

is the number of microstates whose energy equals the compound’s energy.

For an isolated system, .

The entropy of a compound is a quantity that satisfies “an infinitesimal change in entropy is equal to the ratio of an infinitesimal quantity of heat in a reversible reaction and the temperature of the compound”.

The molar entropy of a compound is the entropy of the compound per mole.

The change in entropy of a reaction is the difference between the sum of the entropies of the products of the reaction and the sum of the entropies of the reactants.

The change in molar entropy of a reaction is the difference between the sum of the molar entropies of the products of the reaction and the sum of the molar entropies of the reactants.

The Gibbs free energy of a compound at some time is the difference between the enthalpy of the compound and the product of the temperature and entropy of the compound.

The molar Gibbs free energy is the difference between the molar enthalpy of the compound and the product of the temperature and the molar entropy of the compound.

The change in Gibbs free energy of a reaction is the maximum amount of free and useful energy available to do non-volume-expansion work that can be extracted from the reactants at fixed temperature and pressure, which can be attained only in a completely reversible process. The change in Gibbs free energy of a reaction is the difference between the change in enthalpy of the reaction and the change in the product of the temperature and entropy of the compound.

The change in molar Gibbs free energy of a compound is the difference between the change in molar enthalpy of the compound and the change in the product of the temperature and molar entropy of the compound.