Spring - 2021

IE – 6318-001

DATA MINING AND ANALYTICS

Predicting molecular properties for identification of new protein using datamining and machine learning techniques

- **Guided by Dr. Shouyi Wang**

Submitted by -

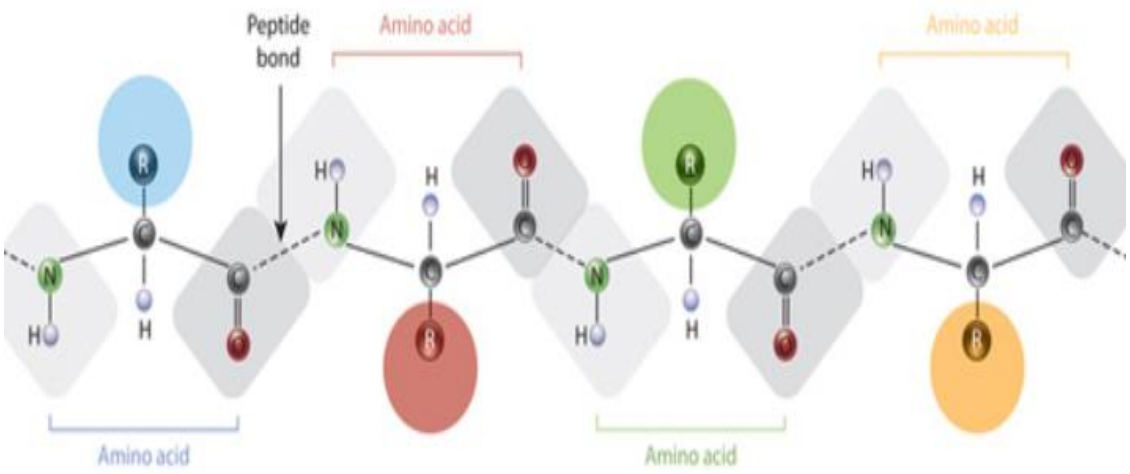
Jayant Madan #1001814817

ABSTRACT

Proteins are large, complex molecules that carry out the tasks of life. Each protein is initially formed as a string of amino acids whose identity and order are dictated by a gene according to the sequence of its DNA bases. It is crucially important to identify the protein which will aid researchers to determine, predict and verify whether an organism is having a genetic disease. Most genetic diseases are kind of caused by the protein mutation which has a composition slightly different from that of the normal protein it replaces. Mass spectrometry is the most widely used and powerful tools present to study protein. The mass spectrometer is an ideal instrument for identifying amino acids the building blocks of proteins and determining the order in which they are arranged. Number of properties like molecular weight, topological polar surface area, lipophilicity. In this analytical study compound's canonical smiles are used to encode their fingerprints and those are then used for predicting the molecular properties.

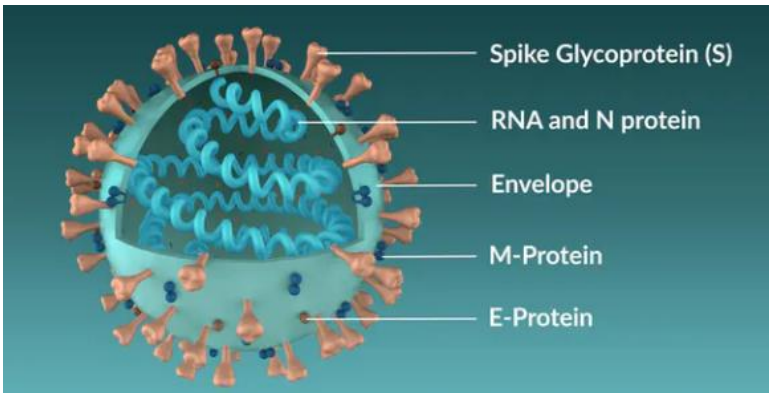
BACKGROUND

Protein is existent in all living being and they are most essential and play many vital roles in our body. Protein consists of hundreds and thousands of smaller units also called as amino acids. There are almost 20 different types of amino acid that contribute to the protein. Mostly proteins are made up of carbon, hydrogen, nitrogen oxygen and sulfur. Protein functions in many ways and have many different forms. Antibodies, a form of protein, it bonds to foreign particles to help prevent and protect. Enzyme, it carries out all the reactions takes place and aid in the formation on new molecule. Messenger proteins transport signals between cells and tissues. Structural protein provides structure and support for the cell. Protein comes in different shapes; different properties and each protein is different.



Proteins are classified based on solubility, biological functions. New innovations in Artificial intelligence for generating protein will help speed up the drug development. Researchers from university of Sweden published a work on capable of generating novel and functionally active protein. Drug based protein are very common, even the insulin used for the diabetic patient is the most basic example. Some of the most expensive and effective medicines for cancer are made because of the protein. Moreover, proteins are also in the research phase for treating corona virus. New technology is advancing and focusing on fast synthesis of the protein. Because most of the time devised in the protein synthesis are captured by the patients. The researchers are working on further improving technology that can assemble 400 long amino acids. Understanding the complexity and the process of evolution can be achieved by the identification of novel protein.

Which are nowadays made possible with the help of artificial intelligence, with fast computing techniques and ability to store large amount of data.



INTRODUCTION

In most recent innovations drug-likeness is a factor on which how druglike a substance is. It is estimated from the molecular structure before even the compound is synthesized as tested. A druglike compound has a properties such as molecular weight, topological polar surface area, lipophilicity, etc. Drug likeness indices are inherently limited tools. Drug likeness can be estimated for any molecule and does not evaluate the actual specific effect that the drug achieves (biological activity). Simple rules are not always accurate and may unnecessarily limit the chemical space to search: many best-selling drugs have features that cause them to score low on various drug likeness indices. The amino-acid composition and properties such as molecular weight produced by a certain type of bacteria which is different from those of protein forming coat of particular virus. The ability to identify protein allow researchers to determine whether an organism has a genetic disease.

Lipophilicity (MolLogP in RDkit)

Lipophilicity is a key physicochemical property that plays a crucial role in determining ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties and the overall suitability of drug candidates. indicates a compound's affinity for nonpolar versus polar environments. It is a component of drug activity and many diverse properties (e.g., pharmacokinetics (PK), toxicity). Examples of these include target binding, solubility, permeability, metabolism, active transport, absorption, and off-target toxicity. A common term for lipophilicity is log Dx, the equilibrium distribution coefficient of a compound between octanol and aqueous buffer at pH x. Structural modifications, such as adding a methylene or removing a polar or hydrogen-bonding group, change the lipophilicity and can simultaneously alter multiple properties of the compound to various extent.

Topological polar surface area (TPSA) (CalcTPSA in RDkit)

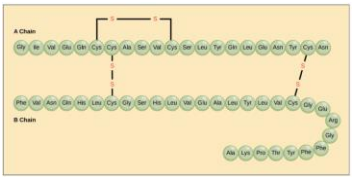
Topological polar surface area is defined as the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms. PSA is a commonly used medicinal chemistry metric for the optimization of a drug's ability to permeate cells. Molecular polar surface area surface belonging to polar atoms, is a descriptor that was shown to correlate well with passive molecular transport through membranes and, therefore, allows prediction of transport properties of drugs.

Molecular weight (ExactMolWt in RDkit): The smaller the better because diffusion is directly affected. The great majority of drugs on the market have molecular weights between 200 and 600 Daltons, and particularly <500; they belong to the group of small molecules

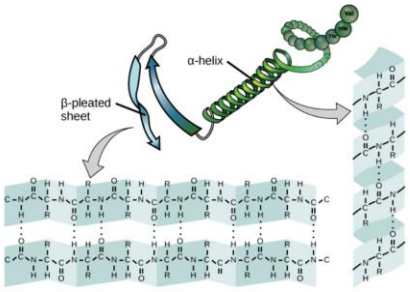
A protein is a linear chain of amino acids connected by peptide bonds. The primary structure of a protein is just the amino acid sequence ordered in the polypeptide chain. Repeated regular conformations on the polypeptide chain are called the secondary structures of proteins. From the secondary structures, a protein can be folded into a stable three-dimensional structure, which is called the tertiary structure of a protein. Although a protein's structure is largely determined by its

amino acid sequence. Proteins are made up of amino acid residues which are bound together by peptide bonds between the amino nitrogen (N) and the carboxyl group. Just 21 distinct amino acids exist, but they can bind together in such a variety of three-dimensional (3D) polypeptide chains that almost limitless potential protein chain sequences exist. This leads to millions of intricate, distinct, potential protein structures. The diversity of protein structures is due to subtle chemical variations which come from differences in the amino acids charge, shape, functional group composition, and size. There are three protein structures describes below.

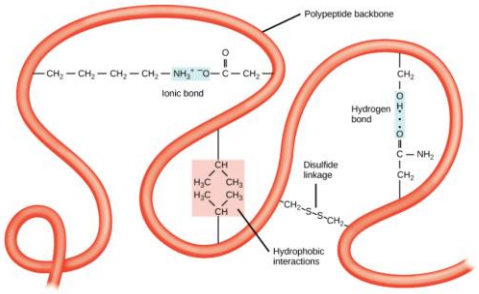
The simplest level of protein structure, **primary structure,** is simply the sequence of amino acids in a polypeptide chain. For example, the hormone insulin has two polypeptide chains, A and B, shown in diagram below. ach chain has its own set of amino acids, assembled in a particular order. A change in the gene's DNA sequence may lead to a change in the amino acid sequence of the protein. Even changing just one amino acid in a protein's sequence can affect the protein's overall structure and function.



The next level of protein structure, secondary structure, refers to local folded structures that form within a polypeptide due to interactions between atoms of the backbone. The most common types of secondary structures are the α helix and the β pleated sheet. Both structures are held in shape by hydrogen bonds, which form between the carbonyl O of one amino acid and the amino H of another.



he overall three-dimensional structure of a polypeptide is called its tertiary structure. The tertiary structure is primarily due to interactions between the R groups of the amino acids that make up the protein.

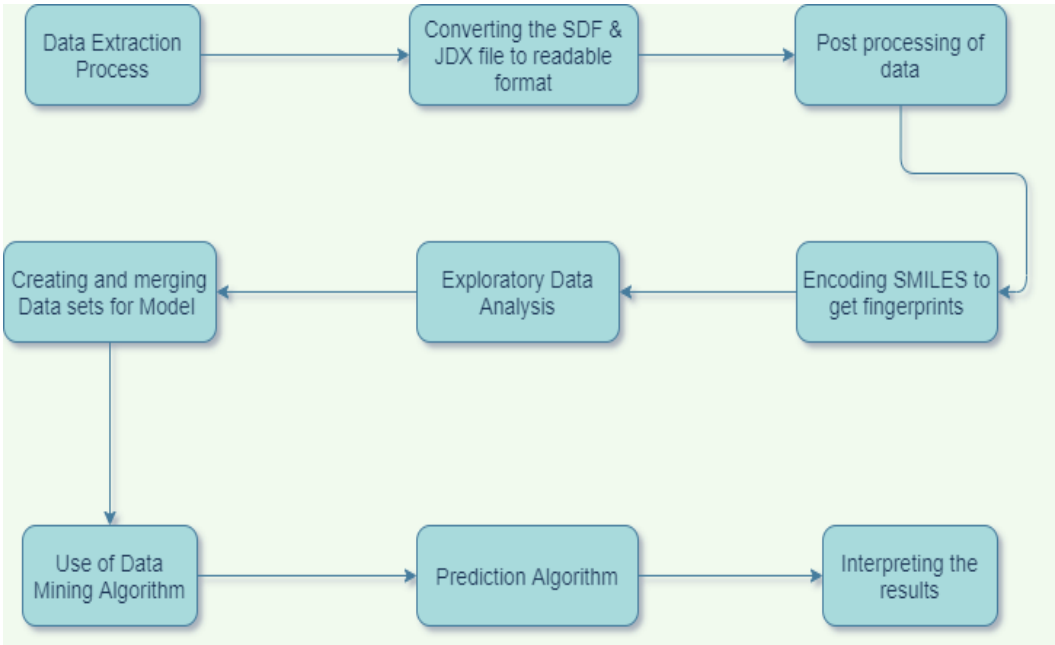


SMILES is currently widely recognized and used as a standard representation of compounds for modern chemical information processing. SMILES provides a linear notation method to represent chemical compounds in a unique way in the form of strings over a fixed alphabet. SMILES uses specific grammar and characters to describe all the atoms and structure of a chemical compound. SMILES can strictly express structural differences including the chirality of compounds.

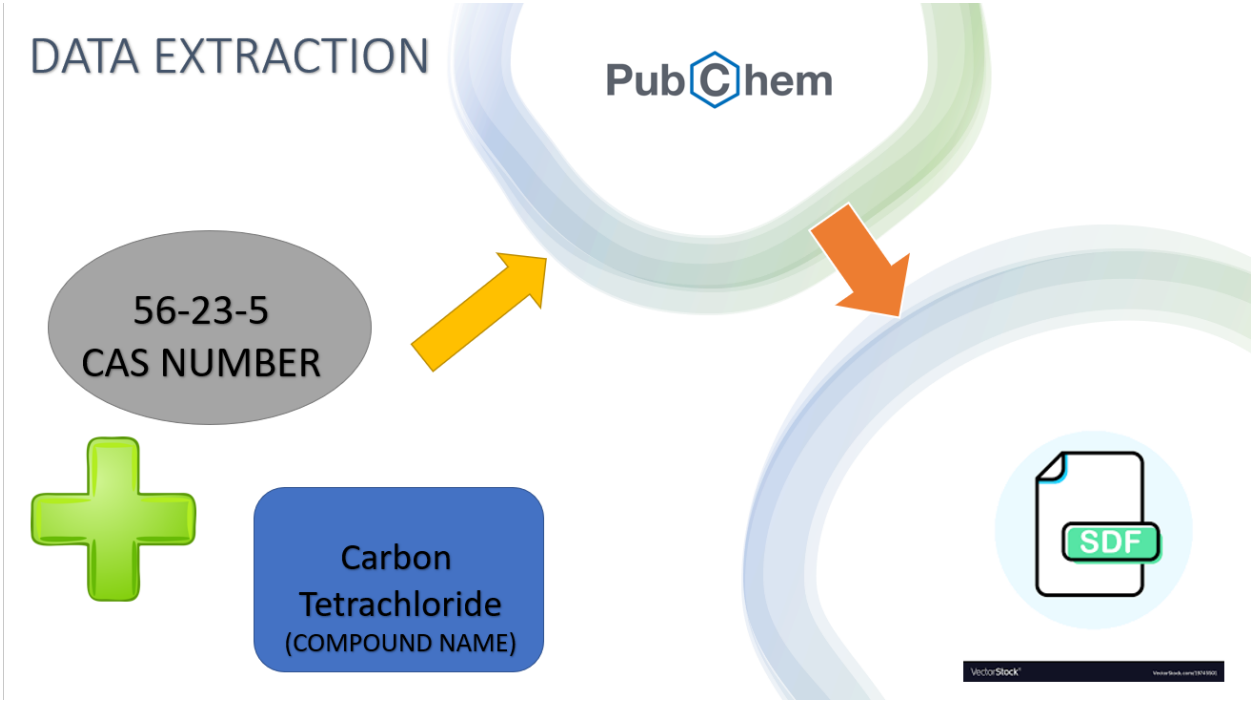
A “fingerprint” is a vector that represents a property of a chemical compound. Many methods for creating fingerprints have been reported. The launch pad we normally use for all fingerprints is 2D fingerprint to indicate what kind of partial structure the compound possesses. In this regard, the most used algorithm is the extended-connectivity fingerprint (ECFP, also known as the circular fingerprint or Morgan fingerprint).

The notion of chemical similarity (or molecular similarity) is one of the most important concepts in cheminformatics. It plays an important role in modern approaches to predicting the properties of chemical compounds, designing chemicals with a predefined set of properties and, especially, in conducting drug design studies by screening large databases containing structures of available (or potentially available) chemicals.

METHODOLOGY

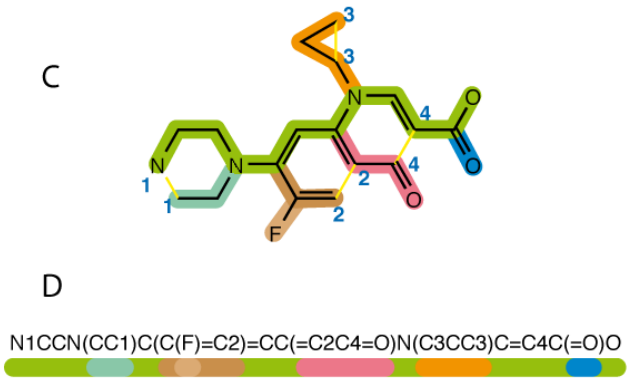


RD-kit was used to read the SDF format into the data frame and JDX format was read with JCAMP-DX. The common files in the respective folders of the aforementioned file formats were extracted so as to prepare the data for a future study on mass spectra prediction as well.



Using RDkit, the SMILES were extracted from the SDF files and beyond that, the chemical fingerprints of each compound was extracted from these SMILES. Each fingerprint is a bit vector of binary data representing the properties of the compound. This is an extremely sparse vector and has a large size of more than 8 million features. Thus, a PCA was performed on the training and testing dataset so as to extract the 5 most important features. This was done to enable the processing of this computationally expensive data. Thus, the desired array for the input data for the machine learning models is obtained.

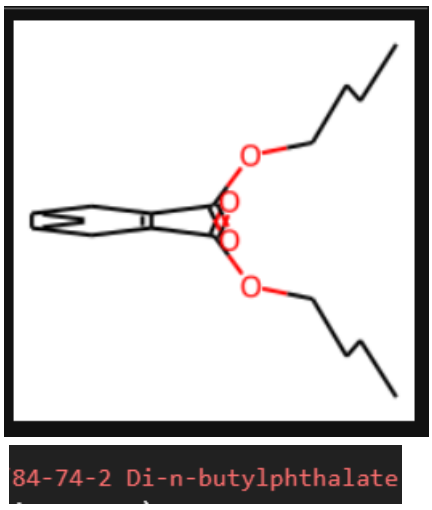
The three properties to be predicted were the exact molecular weight, Lipophilicity and Topological Polar Surface Area (using the in-built functions from RDkit: ExactMolWt (mols), MolLogP (LogP), CalcTPSA(Å^2)).



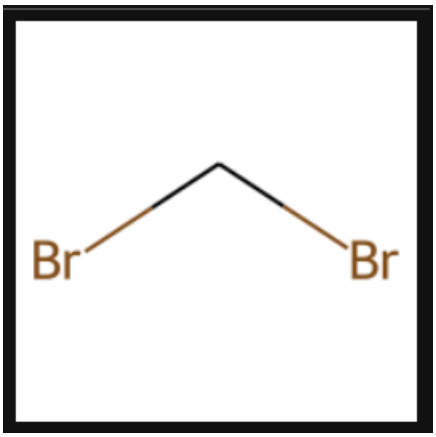
The above figure shows the canonical smiles and structure of the compound. The different colors shows the different chemical bonds with the different compounds and it dorms a long chain.

Parameter Name	Unit	Description
Setting/Input Parameters:		
Fingerprints	Bit vector	It is a vector that represents a property of a chemical compound.
Output Parameters:		
Exact Molecular weight	Mols	The molecular weight of the compound in mols
Lipophilicity	LogP	LogP, this is the partition coefficient of a molecule between an aqueous and lipophilic phase, usually octanol and water.
Topological Polar Surface Area	$Å^2$	The surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.

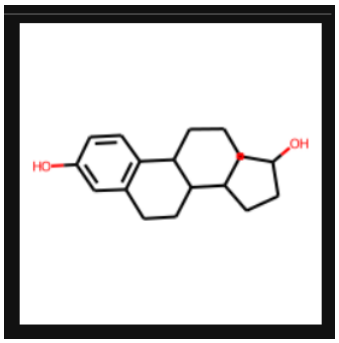
Below figure shows the compound image extracted using the RDKIT.



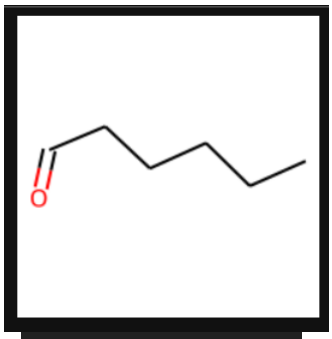
84-74-2 Di-n-butylphthalate



74-95-3 Dibromomethane



50-28-2 Estradiol



66-25-1 Hexaldehyde

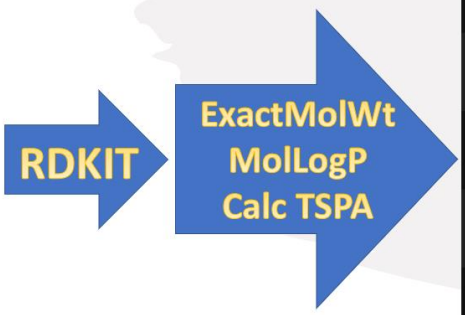
The complete data extraction and data processing is depicted below.

STEP I (using RD_KIT to open SDF files)

index	SMILES
100-41-4	CCc1ccccc1
101-81-5	c1ccc(Cc2ccccc2)cc1
101-84-8	c1ccc(Oc2ccccc2)cc1
1016-05-3	O=S1(=O)c2ccccc2-c2ccccc21
1031-07-8	O=S1(=O)OCC2C(CO1)C1(Cl)C(Cl)=C(Cl)C2(Cl)C1(Cl)Cl
...	...
97-53-0	C=CCc1ccc(O)c(OC)c1
97-63-2	C=C(C)C(=O)OCC
97143-65-2	CN(C)CCOC(C)(c1ccccc1)c1cccc[n+]1[O-]
98-01-1	O=Cc1ccco1
99-85-4	CC1=CCC(C(C)C)=CC1

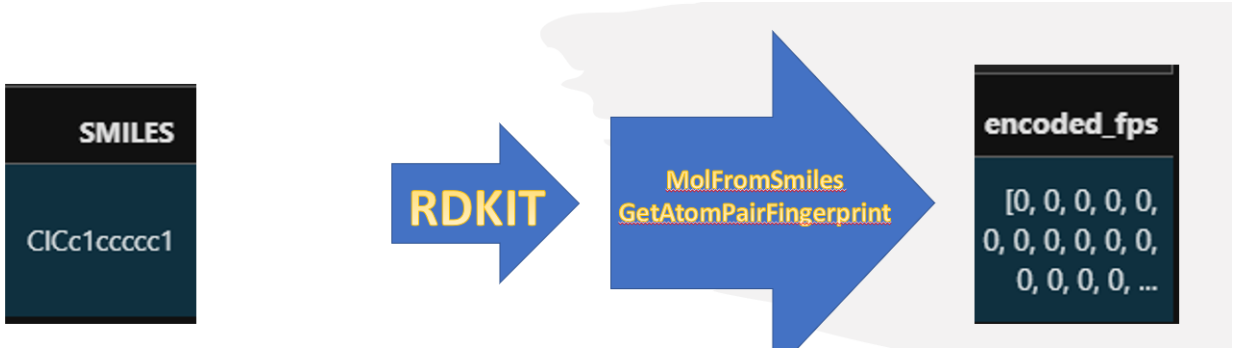
STEP II (extracting molecular properties)

index	SMILES
100-41-4	CCc1ccccc1
101-81-5	c1ccc(Cc2ccccc2)cc1
101-84-8	c1ccc(Oc2ccccc2)cc1
1016-05-3	O=S1(=O)c2ccccc2-c2ccccc21
1031-07-8	O=S1(=O)OCC2C(CO1)C1(Cl)C(Cl)=C(Cl)C2(Cl)C1(Cl)Cl
...	...
97-53-0	C=CCc1ccc(O)c(OC)c1
97-63-2	C=C(C)C(=O)OCC
97143-65-2	CN(C)CCOC(C)(c1ccccc1)c1cccc[n+]1[O-]
98-01-1	O=Cc1ccco1
99-85-4	CC1=CCC(C(C)C)=CC1



Mol_Wt	Mol_log_P	Calc_TPSA
126.023628	2.42540	0.00
107.073499	1.14530	26.02
103.042199	1.55828	23.79
106.041865	1.49910	17.07
108.057515	1.69520	9.23

STEP III (converting to fingerprints)

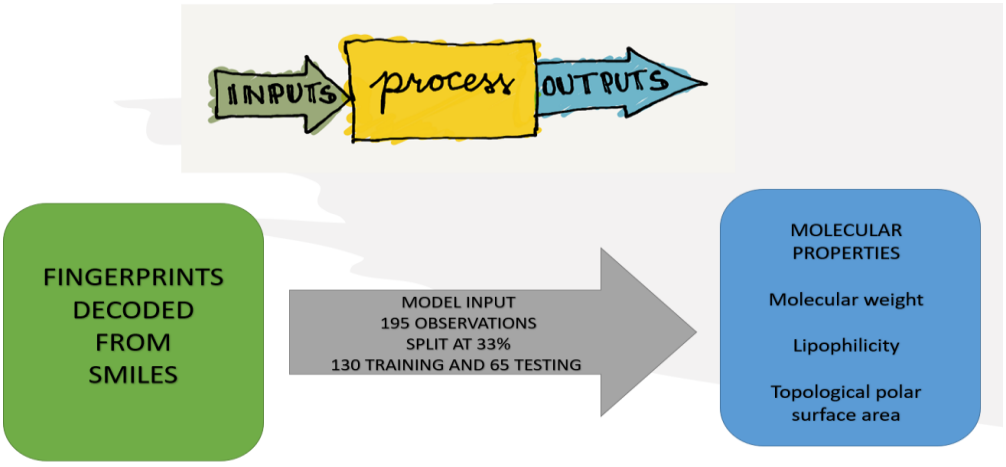


MACHINE LEARNING MODEL SELECTION

Regression is a supervised Machine learning technique Which is used to predict Continuous values. As in our case we are predicting Molecular properties Which is a continuous variable The goal is to plot the bestfit line between the data It help us to predict output Variables using dependent input variables.

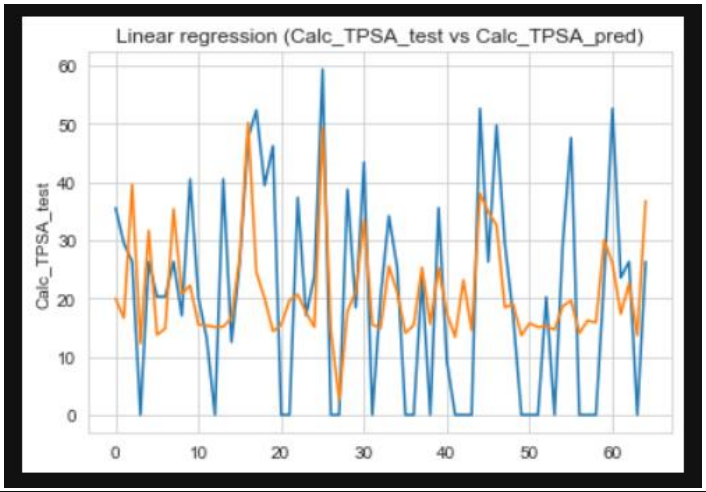
Algorithms	Description
K-nearest Neighbors Regression	KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. In K-Nearest Neighbors Regression the output is the property value for the object. There is no model other than the raw training dataset and the only computation performed is the querying of the training dataset when a prediction is requested. It is a simple algorithm, but one that does not assume very much about the problem other than that the distance between data instances is meaningful in making predictions. As such, it often achieves very good performance. When making predictions on regression problems, KNN will take the mean of the k most similar instances in the training dataset.
Multi Output Regression	Multioutput regression are regression problems that involve predicting two or more numerical values given an input example. An example might be to predict a coordinate given an input, for example predicting x and y values. Another example would be multi-step time series forecasting that involves predicting multiple future time series of a given variable. Many machine learning algorithms are designed for predicting a single numeric value, referred to simply as regression. Some algorithms do support multioutput regression inherently, such as linear regression and decision trees. There are also special workaround models that can be used to wrap and use those algorithms that do not natively support predicting multiple outputs.

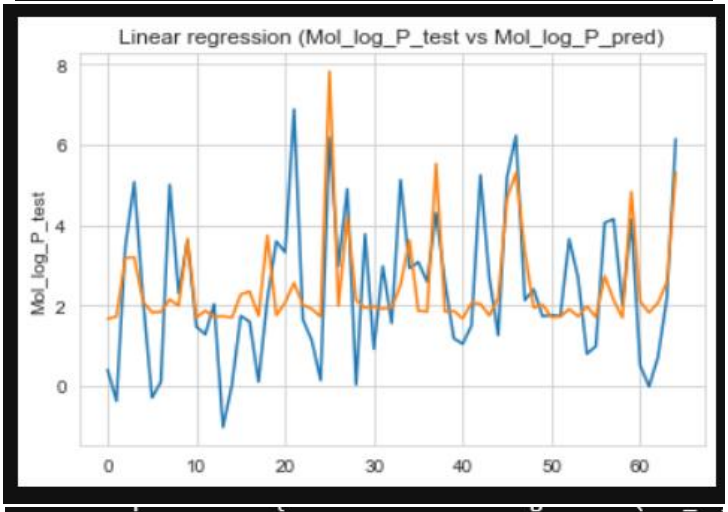
Random Forest Multioutput Regression	Random Forest regression refers to ensembles of regression trees where a set of n tree un-pruned regression trees are generated based on bootstrap sampling from the original training data. For each node, the optimal feature for node splitting is selected from a random set of m feature from the total N features. The selection of the feature for node splitting from a random set of features decreases the correlation between different trees and thus the average prediction of multiple regression trees is expected to have lower variance than individual regression trees. A random forest regressor is used, which supports multi-output regression natively, so the results can be compared. The random forest regressor will only ever predict values within the range of observations or closer to zero for each of the targets. As a result, the predictions are biased towards the center of the circle. Using a single underlying feature, the model learns both the x and y coordinate as output.
Gradient Boosting for Regression	GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. Gradient Boosting Regressor supports a number of different loss functions for regression which can be specified via the argument loss; the default loss function for regression is least squares ('ls').



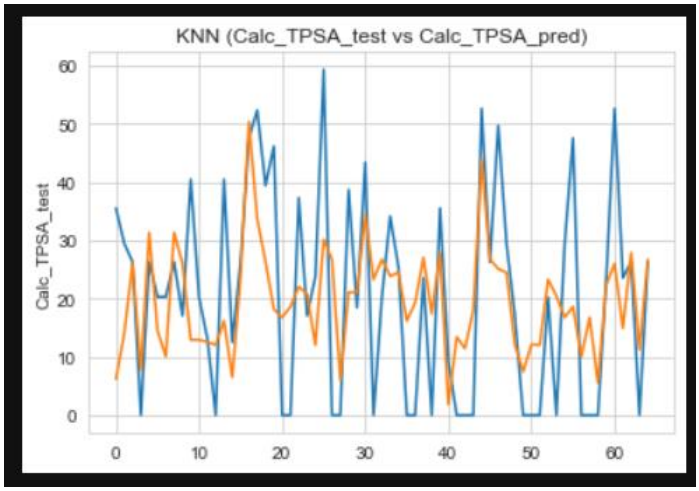
LINEAR REGRESSION MODEL OUTPUT

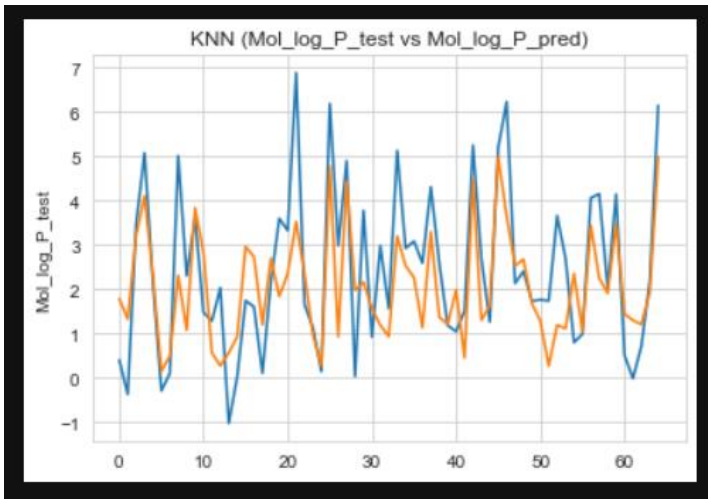
The blue line represents the actual values and the red line represents the predicted values.

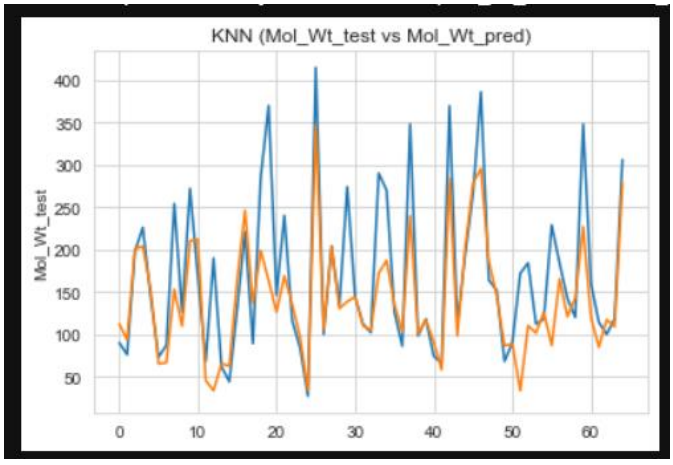




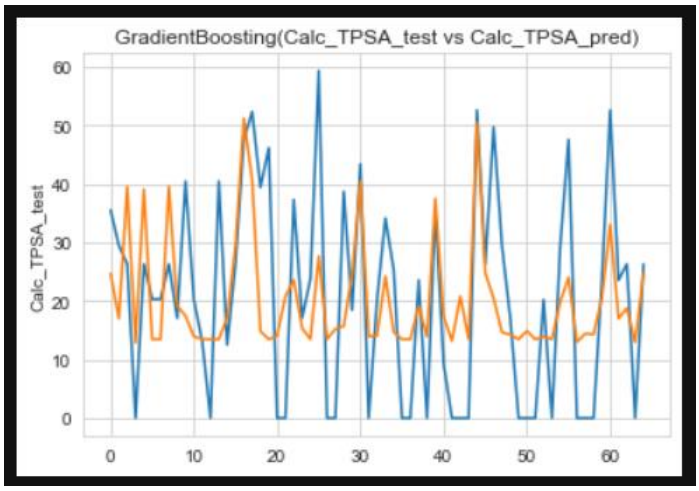
KNN REGRESSION MODEL OUTPUT

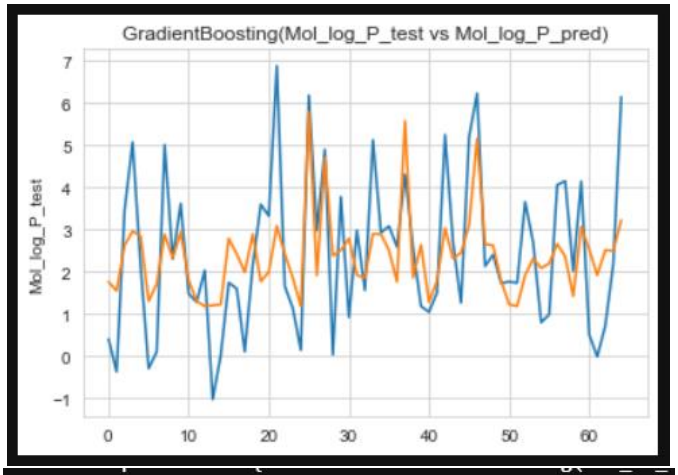


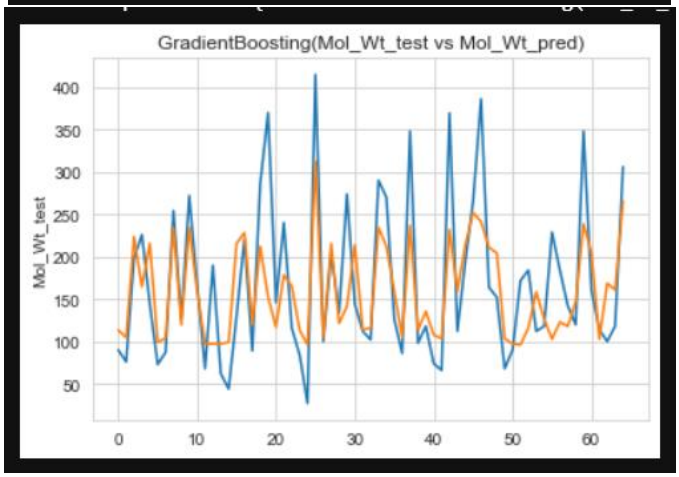




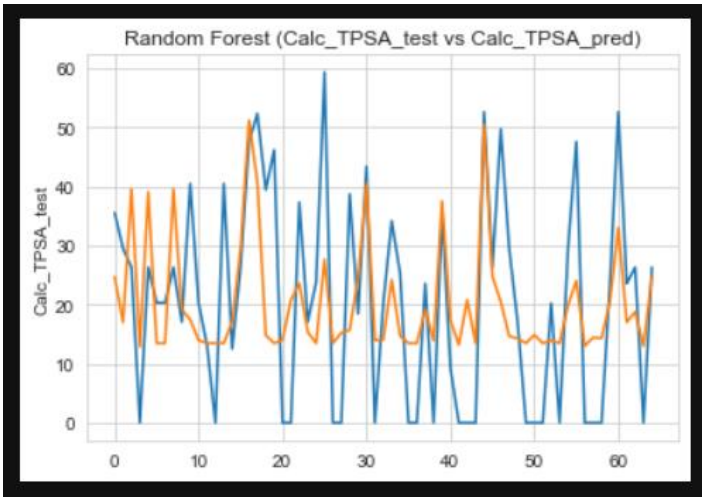
GRADIENT BOOSTING MODEL OUTPUT

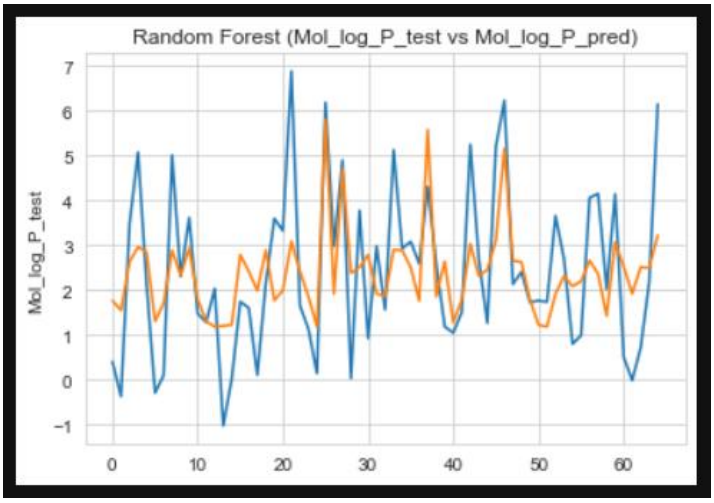


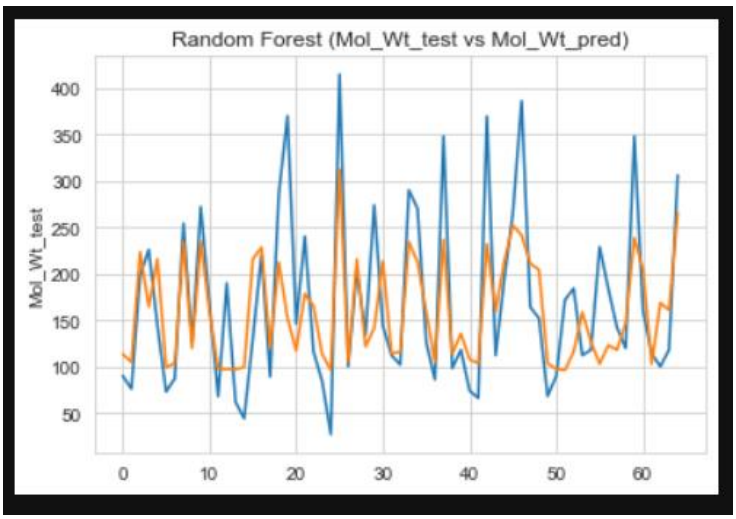




RANDOM FOREST REGRESSION MODEL OUTPUT


Random Forest (Calc_TPSA_test vs Calc_TPSA_pred)
Calc_TPSA_test
0
10
20
30
40
50
60


Random Forest (Mol_log_P_test vs Mol_log_P_pred)
Mol_log_P_test
−1
0
1
2
3
4
5
6
7
0
10
20
30
40
50
60


Random Forest (Mol_Wt_test vs Mol_Wt_pred)
Mol_Wt_test
50
100
150
200
250
300
350
400
0
10
20
30
40
50
60

These are the results we got from the random forest regressor. We did the comparison between three molecular properties with the actual and predicted data. On x axis we have the index and y axis the respected molecular values. The blue line represent the actual values and the orange line represents the predicted values. In case of random forest regressor both the line graphs are deviated a little bit. In case of KNN regression we got the same results but better as sompared to previous ones. Linear regression haven't performed well as it is apparent by the graph. Among all the models gradient boosting have performed the best.In all the four model evaluvated on the metrics it is apparent that the gradient boosting regression model performed best.R2 is the highest in gradient boosting as same as the case with the r2 adjusted. Rmse is lower in case of gradient boosting and so as the case with mae

PERFORMANCE METRICS

Metric are the measure of Performance on how well the Model performs. Machine learning model were evaluated on 4 metrics.

Evaluation metrics and distance/similarity measures:

Similarity measure:

Jaccard Coefficient: The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient T. Two structures are usually considered similar if T > 0.85 (for Daylight fingerprints). [5]. The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets.

Jaccard Coefficients

J = number of 11 matches / number of not-both-zero attributes' values = (M11) / (M01 + M10 + M11)

From the Morgan Fingerprint function, the chemical similarity was found for two chemical SMILES: 'O=[N+]([O-])c1ccc(O)cc1' and 'Cc1cc(-c2ccc(N)c(C)c2)ccc1N' was found to be:

```
morgan score:     0.1639
```

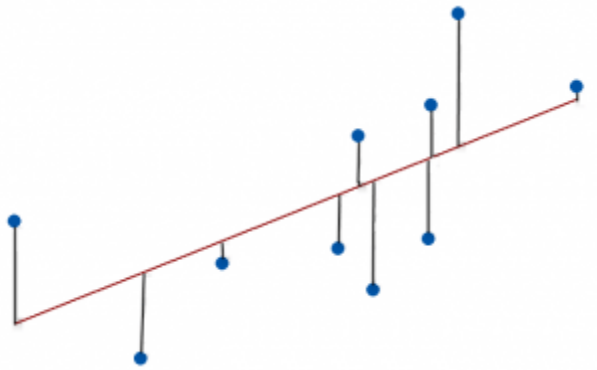
This is another way of comparing the similarity of two compounds than the bit vector representation from the Jaccard coefficients. This similarity is used to compare the likeness of two compounds in drug discovery as well.

R^2 Score

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain

collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. After fitting a linear regression model, you need to determine how well the model fits the data. Does it do a good job of explaining changes in the dependent variable? There are several key goodness-of-fit statistics for regression analysis. In this post, we'll examine R-squared (R^2), highlight some of its limitations, and discover some surprises. For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good.

Assessing Goodness-of-Fit in a Regression Model- Residuals are the distance between the observed value and the fitted value. Linear regression identifies the equation that produces the smallest difference between all the observed values and their fitted values. To be precise, linear regression finds the smallest sum of squared residuals that is possible for the dataset.



R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values. R-squared is the percentage of the dependent variable variation that a linear model explains.

R-squared is always between 0 and 100%. 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model. 100% represents a model that explains all the variation in the response variable around its mean. Usually, the larger the R^2, the better the regression model fits your observations. However, this guideline has important caveats that I'll discuss in both this post and the next post.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

Adjusted R-squared

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. Suppose you compare a five-predictor model with a higher R-squared to a one-predictor model. Does the five-predictor model have a higher R-squared because it's better? Or is the R-squared higher because it has more predictors? Simply compare the adjusted R-squared values to find out! The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it is usually not. It is always lower than the R-squared.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

Root Mean Squared Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

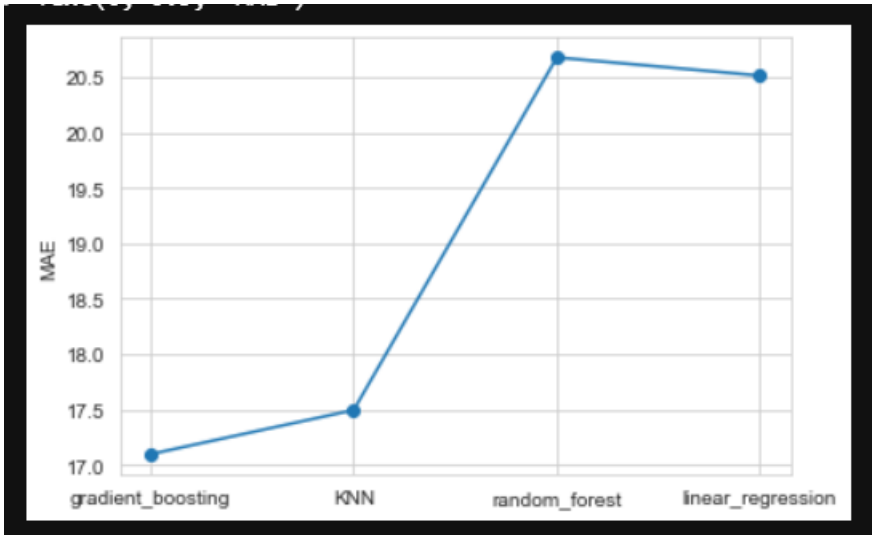
$$RMSE = \sqrt{(f-o)^2}$$

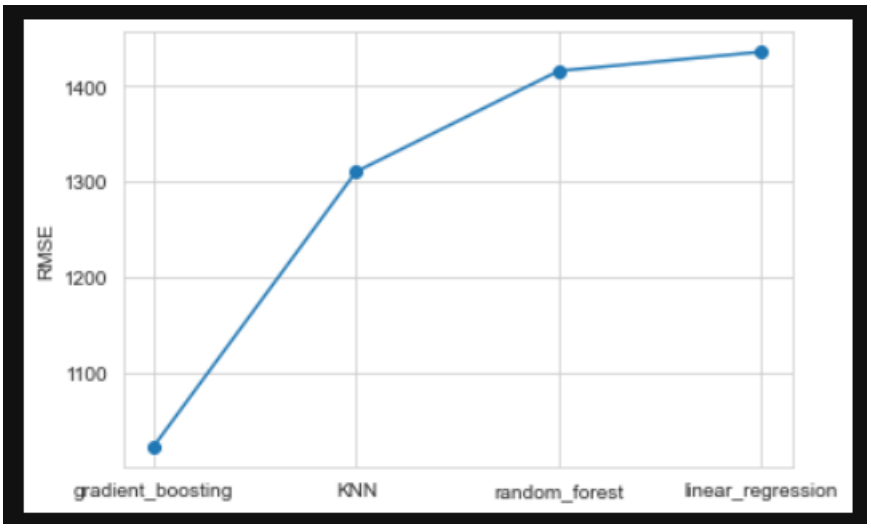
Mean Absolute Error

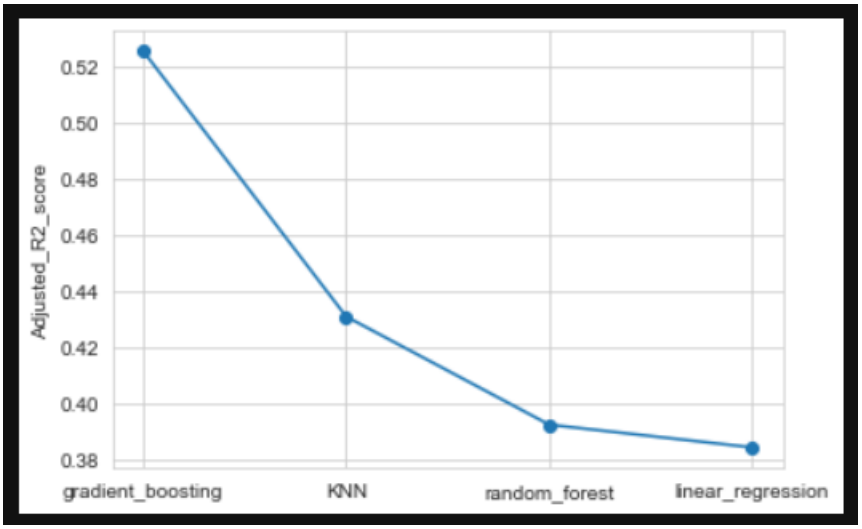
The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures *accuracy* for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

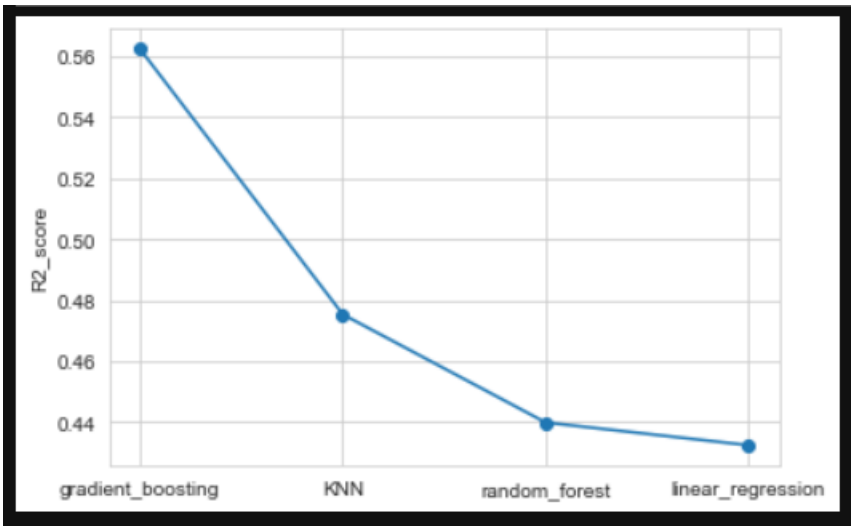
$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

The MAE is also the most intuitive of the metrics since we are just looking at the absolute difference between the data and the model's predictions. Because we use the absolute value of the residual, the MAE does not indicate **underperformance** or **overperformance** of the model (whether the model under or overshoots actual data). Each residual contributes proportionally to the total amount of error, meaning that larger errors will contribute linearly to the overall error. Like we have said above, a small MAE suggests the model is great at prediction, while a large MAE suggests that your model may have trouble in certain areas. A MAE of 0 means that your model is a **perfect** predictor of the outputs (but this will almost never happen). While the MAE is easily interpretable, using the absolute value of the residual often is not as desirable as **squaring** this difference. Depending on how you want your model to treat **outliers**, or extreme values, in your data, you may want to bring more attention to these outliers or downplay them. The issue of outliers can play a major role in which error metric you use.




RMSE
1400
1300
1200
1100
gradient_boosting
KNN
random_forest
linear_regression


Adjusted_R2_score
0.52
0.50
0.48
0.46
0.44
0.42
0.40
0.38
gradient_boosting
KNN
random_forest
linear_regression


R2_score
0.56
0.54
0.52
0.50
0.48
0.46
0.44
gradient_boosting
KNN
random_forest
linear_regression

The best metric to evaluate the models is the MAE as its value tells us that the model is predicting a particular amount more or less on average than the actual value.

Thus, from these metrics, the best model can be seen to be Gradient Boosting, with the lowest Root Mean Squared Error and the lowest Mean Absolute Error. The less the value of MAE the better the performance of the model.

	R2_score	Adjusted_R2_score	RMSE	MAE
gradient_boosting	0.562501	0.525425	1022.223762	17.097531
KNN	0.475377	0.430917	1310.189996	17.495701
random_forest	0.439903	0.392437	1415.724366	20.676971
linear_regression	0.432448	0.384350	1435.685836	20.513405

Conclusion:

Protein synthesis and more generally, drug discovery is an extremely challenging but crucial field for the future of the survival of our species. Gradient Boosting gave us promising results (for the appropriate metrics MAE and RMSE), given the small size of this dataset (due to computational limitations). The next best models were Random Forest and Linear Regression which performed better on MAE and RMSE, respectively. However, the simpler model i.e., Linear Regression might be a better choice to run for a larger dataset of thousands of molecules.

Future Work:

The next step would be to have the computational capacity to add more molecules for training. Beyond this project, there are many other important properties to predict that aid drug discovery based on their proximity with existing molecules. They include but are not limited to electron ionized mass spectra, solubility, the number of H-bond donors for a molecule, and the number of H-bond acceptors for a molecule.

Acknowledgement:

We would like to thank Dr. Wang and the Teaching Assistant, Linh Ho Manh for their unwavering encouragement and readily available support during the trying aspects of this project.

REFERENCES

- https://www.eurekalert.org/features/doe/2001-06/drnl-pib061902.php#:~:text=Each%20protein%20is%20initially%20formed,sequence%20of%20its%20DNA%20bases.&text=The%20ability%20to%20identify%20proteins,organism%20has%20a%20genetic%20disease.
- https://www.britannica.com/science/protein/General-structure-and-properties-of-proteins
- https://medlineplus.gov/genetics/understanding/howgeneswork/protein/
- https://www.livescience.com/53044-protein.html#:~:text=Chemically%2C%20protein%20is%20composed%20of,Institutes%20of%20Health%20(NIH).
- https://courses.lumenlearning.com/wm-biology1/chapter/reading-function-of-proteins/
- https://www.eurekalert.org/pub_releases/2021-03/cuot-uam032921.php
- https://www.eurekalert.org/pub_releases/2021-03/cuot-uam032921.php
- https://news.mit.edu/2020/faster-protein-synthesis-0528
- https://en.wikipedia.org/wiki/Druglikeness
- https://www.sciencedirect.com/science/article/pii/B9780128010761000058
- https://pubs.acs.org/doi/10.1021/jm000942e
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2280-5
- https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure
- https://pubs.acs.org/doi/abs/10.1021/ci00057a005
- https://en.wikipedia.org/wiki/Druglikeness
- https://pubmed.ncbi.nlm.nih.gov/22823020/
- https://pubs.acs.org/doi/10.1021/ci100050t
- https://en.wikipedia.org/wiki/Chemical_similarity#:~:text=The%20most%20popular%20similarity%20measure,0.85%20(for%20Daylight%20fingerprints).
- https://www.azom.com/article.aspx?ArticleID=19609
- https://en.wikipedia.org/wiki/Polar_surface_area
- https://www.youtube.com/watch?v=U86Qn9V33Y8
- https://thedatascientist.com/performance-measures-rmse-mae/
- https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/#:~:text=Root%20Mean%20Square%20Error%20(RMSE)%20is%20the%20standard%20deviation%20of,the%20line%20of%20best%20fit.
- https://www.researchgate.net/figure/nterpretation-of-typical-MAPE-values_tbl1_257812432
- https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/#:~:text=The%20RMSE%20is%20the%20square,an%20absolute%20measure%20of%20fit.&text=Lower%20values%20of%20RMSE%20indicate%20better%20fit.
- https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
- https://cran.r-project.org/web/packages/MultivariateRandomForest/MultivariateRandomForest.pdf
- https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
- www.pubchem.com

- www.webbook.nist.gov/chemistry