MACHINE LEARNING INSIGHTS FOR PREDICTING ORAL DRUGS PROPERTIES

LUCIANA OLIVEIRA & MARÍA URIBURU GRAY

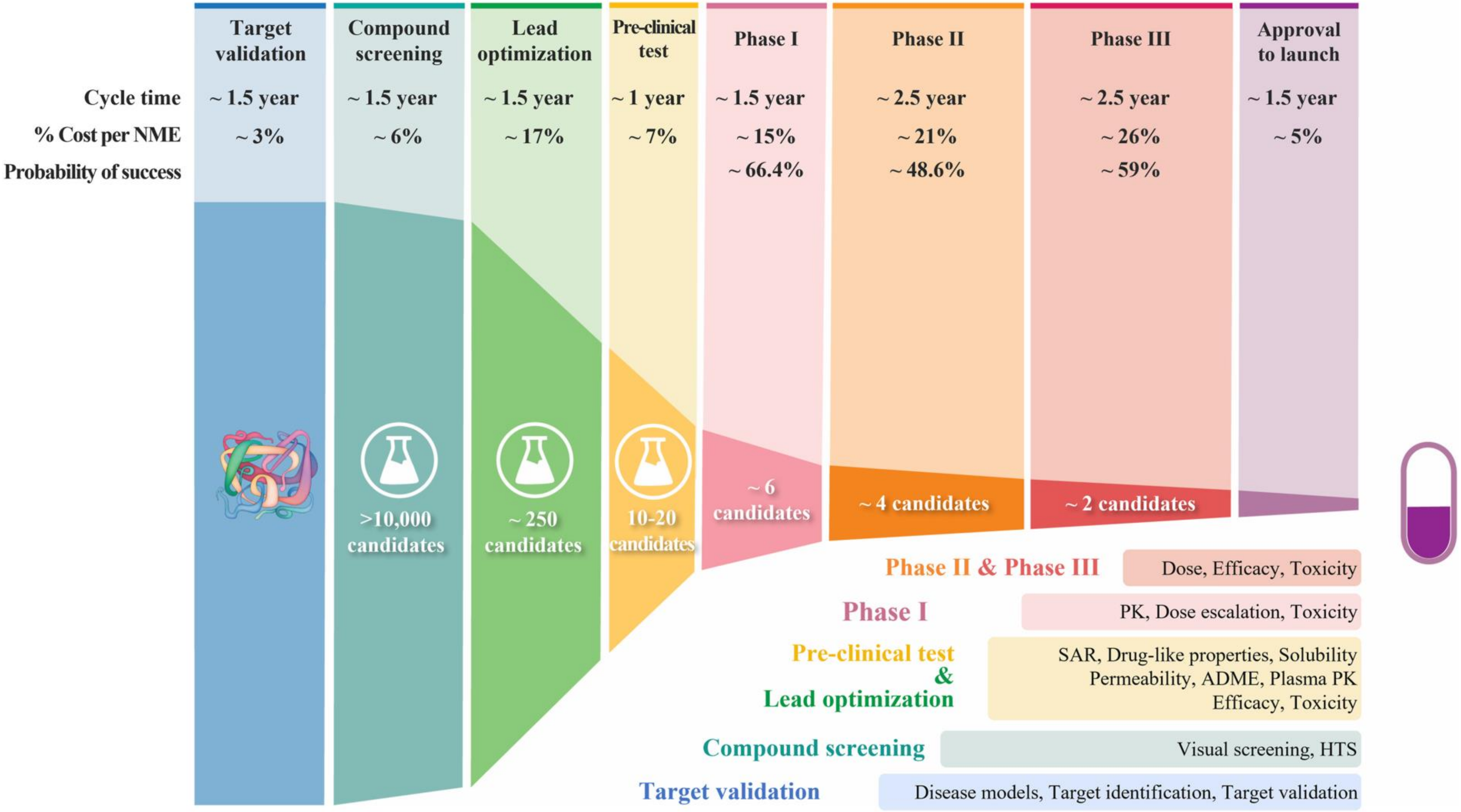
12/12/2024

AGENDA

- INTRODUCTION
- MAIN OBJECTIVE
- DATASET
- EXPLORATORY DATA ANALYSIS
- MACHINE LEARNING ALGORITHMS
- CONCLUSIONS

INTRODUCTION

Drug Discovery – Traditional Process



	Target validation	Compound screening	Lead optimization	Pre-clinical test	Phase I	Phase II	Phase III	Approval to launch
Cycle time	~ 1.5 year	~ 1.5 year	~ 1.5 year	~ 1 year	~ 1.5 year	~ 2.5 year	~ 2.5 year	~ 1.5 year
% Cost per NME	~ 3%	~ 6%	~ 17%	~ 7%	~ 15%	~ 21%	~ 26%	~ 5%
Probability of success					~ 66.4%	~ 48.6%	~ 59%	

Phase II & Phase III	Dose, Efficacy, Toxicity
Phase I	PK, Dose escalation, Toxicity
Pre-clinical test & Lead optimization	SAR, Drug-like properties, Solubility Permeability, ADME, Plasma PK Efficacy, Toxicity
Compound screening	Visual screening, HTS
Target validation	Disease models, Target identification, Target validation

7 phases

~ 10 years

~ $2.6 billion

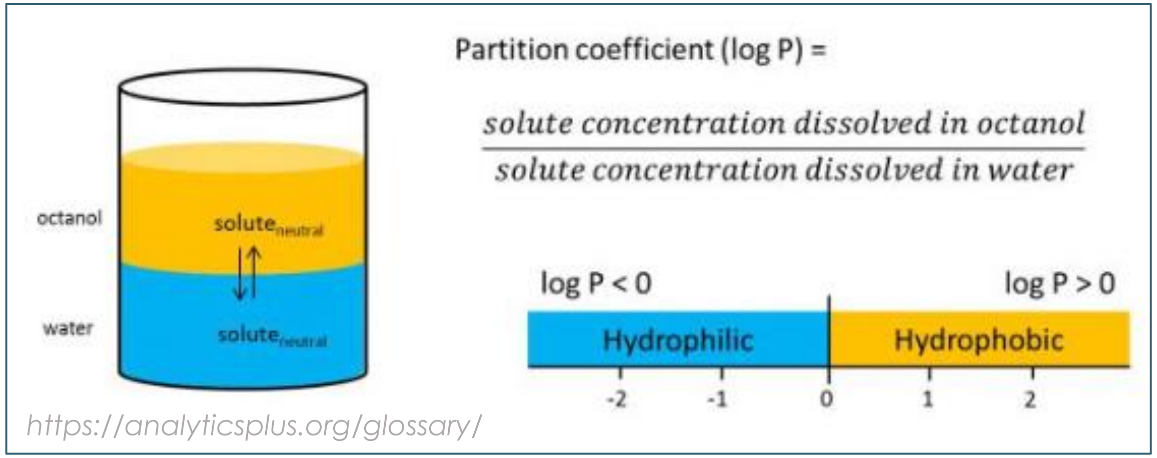
INTRODUCTION

Machine Learning in Drug Discovery

- **Supervised ML in pharmaceutics**
  - **Molecular Property Prediction**
    - **The Lipinski's Rule**
      - Molecular weight
      - **Hydrophobicity or Lipophilia**
        - **X Log P**
        - **Mannhold Log P**
      - H-Bond donors
      - H-Bond acceptors
  - Pharmacokinetics Modelling
  - Chemical Synthesis Prediction
  - Pharmacogenomics Modelling
  - Toxicity and Biodisponibility

Empirical rule

to assess whether a molecule has suitable chemical properties for oral bioavailability.



Partition coefficient (log P) =

$$\frac{\textit{solute concentration dissolved in octanol}}{\textit{solute concentration dissolved in water}}$$

https://analyticsplus.org/glossary/

INTRODUCTION

Machine Learning in Drug Discovery

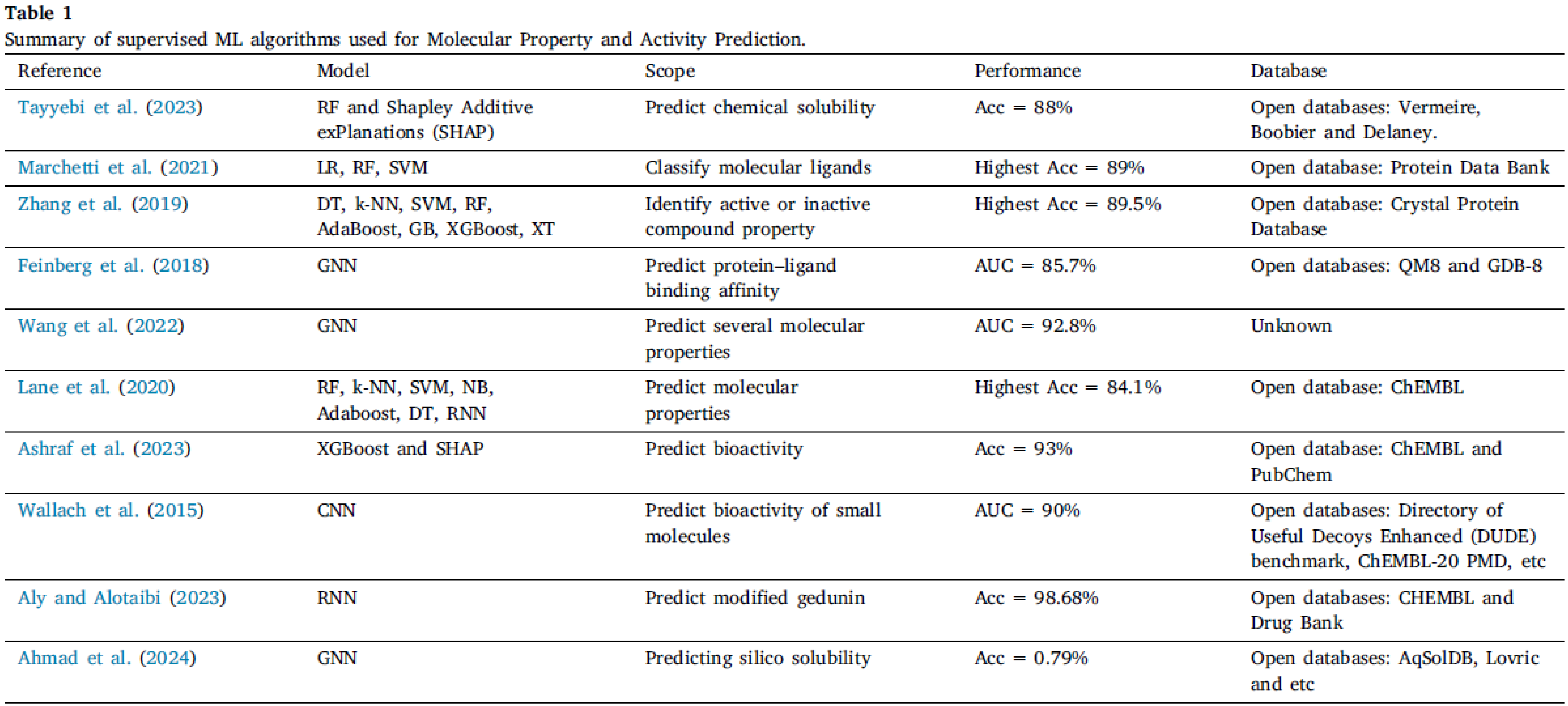
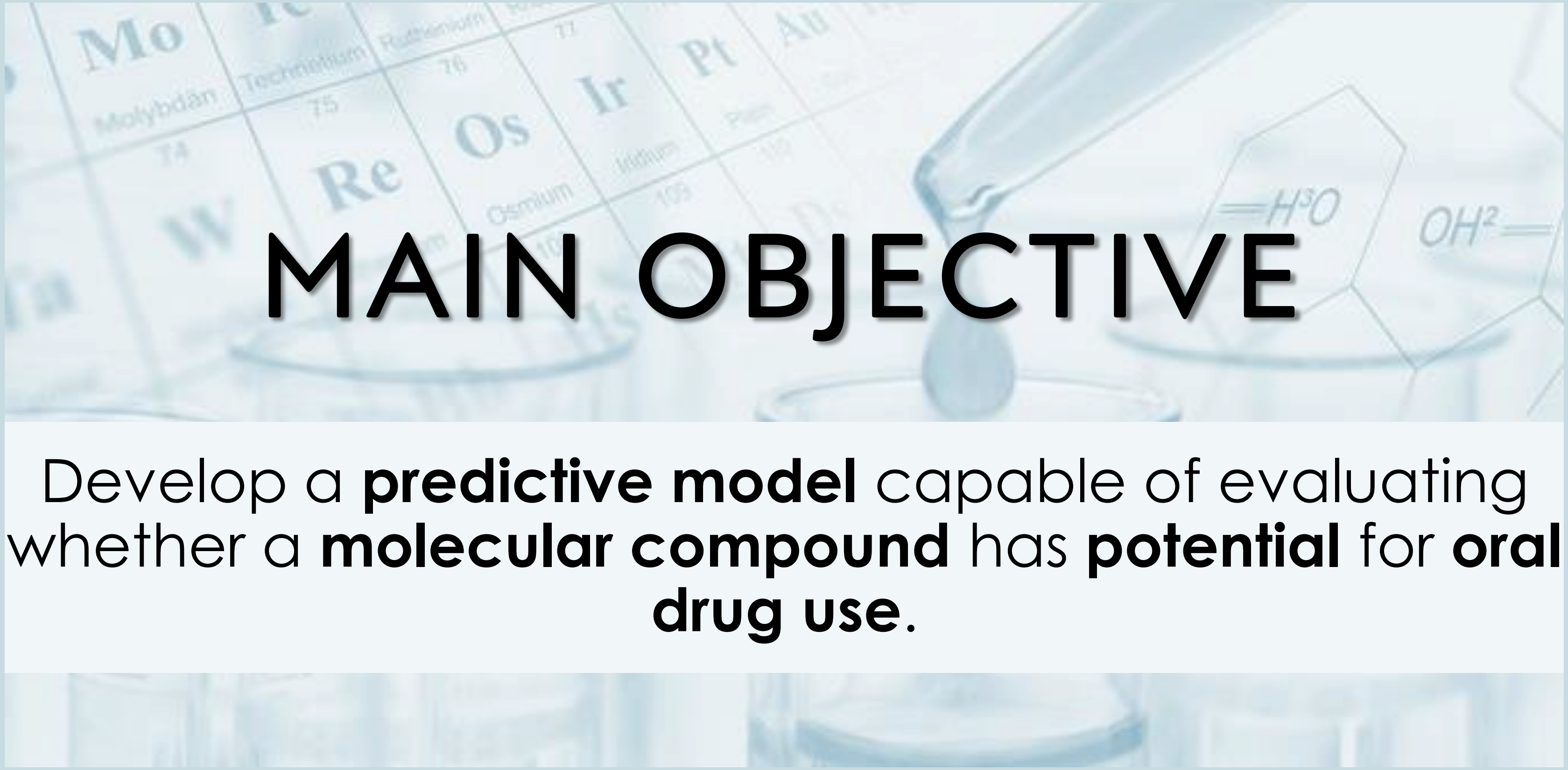


Table 1
Summary of supervised ML algorithms used for Molecular Property and Activity Prediction.

Reference	Model	Scope	Performance	Database
Tayyebi et al. (2023)	RF and Shapley Additive exPlanations (SHAP)	Predict chemical solubility	Acc = 88%	Open databases: Vermeire, Boobier and Delaney.
Marchetti et al. (2021)	LR, RF, SVM	Classify molecular ligands	Highest Acc = 89%	Open database: Protein Data Bank
Zhang et al. (2019)	DT, k-NN, SVM, RF, AdaBoost, GB, XGBoost, XT	Identify active or inactive compound property	Highest Acc = 89.5%	Open database: Crystal Protein Database
Feinberg et al. (2018)	GNN	Predict protein–ligand binding affinity	AUC = 85.7%	Open databases: QM8 and GDB-8
Wang et al. (2022)	GNN	Predict several molecular properties	AUC = 92.8%	Unknown
Lane et al. (2020)	RF, k-NN, SVM, NB, Adaboost, DT, RNN	Predict molecular properties	Highest Acc = 84.1%	Open database: ChEMBL
Ashraf et al. (2023)	XGBoost and SHAP	Predict bioactivity	Acc = 93%	Open database: ChEMBL and PubChem
Wallach et al. (2015)	CNN	Predict bioactivity of small molecules	AUC = 90%	Open databases: Directory of Useful Decoys Enhanced (DUDE) benchmark, ChEMBL-20 PMD, etc
Aly and Alotaibi (2023)	RNN	Predict modified gedunin	Acc = 98.68%	Open databases: CHEMBL and Drug Bank
Ahmad et al. (2024)	GNN	Predicting silico solubility	Acc = 0.79%	Open databases: AqSolDB, Lovric and etc



MAIN OBJECTIVE

Develop a **predictive model** capable of evaluating whether a **molecular compound** has **potential** for **oral drug use**.

DATASET

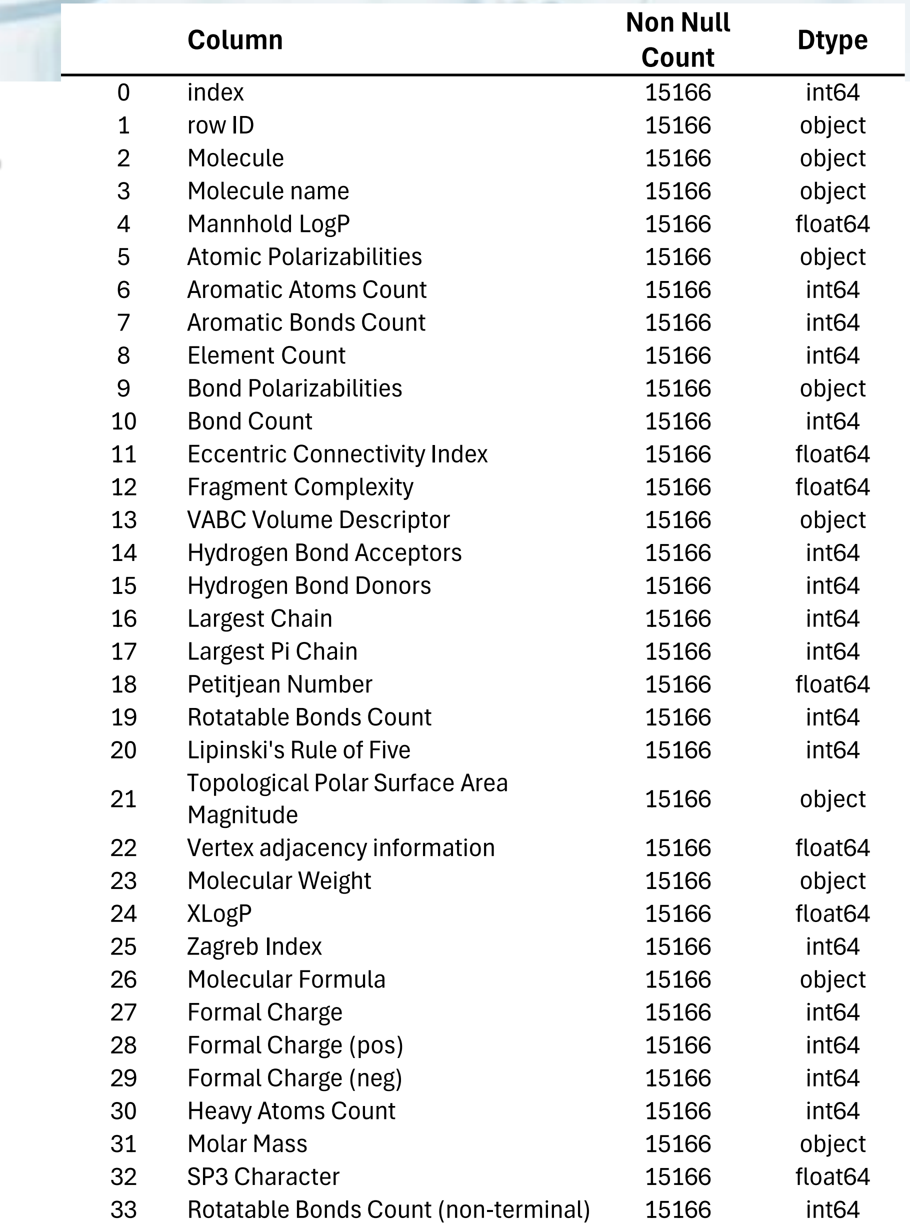
Wikipedia Molecules Properties Dataset

Molecular Properties Dataset from Wikipedia

[34 columns x 15166 rows]

This dataset is a collection of **molecular properties from various chemical substances.** Each entry represents a unique molecule and contains detailed information about its **chemical structure and characteristics**, including:

- **Molecular Structure**: Textual representation.
- **Physicochemical Properties**: Molecular weight, hydrophobicity (LogP or Mannhold LogP), polar surface area, etc.
- **Structural Properties**: Number of aromatic bonds, pi-chain length, etc.
- **Atomic Properties**: Polarizability of atoms and bonds.
- **Additional Information**: Molecular formula, formal charge, etc.



	Column	Non Null Count	Dtype
0	index	15166	int64
1	row ID	15166	object
2	Molecule	15166	object
3	Molecule name	15166	object
4	Mannhold LogP	15166	float64
5	Atomic Polarizabilities	15166	object
6	Aromatic Atoms Count	15166	int64
7	Aromatic Bonds Count	15166	int64
8	Element Count	15166	int64
9	Bond Polarizabilities	15166	object
10	Bond Count	15166	int64
11	Eccentric Connectivity Index	15166	float64
12	Fragment Complexity	15166	float64
13	VABC Volume Descriptor	15166	object
14	Hydrogen Bond Acceptors	15166	int64
15	Hydrogen Bond Donors	15166	int64
16	Largest Chain	15166	int64
17	Largest Pi Chain	15166	int64
18	Petitjean Number	15166	float64
19	Rotatable Bonds Count	15166	int64
20	Lipinski's Rule of Five	15166	int64
21	Topological Polar Surface Area Magnitude	15166	object
22	Vertex adjacency information	15166	float64
23	Molecular Weight	15166	object
24	XLogP	15166	float64
25	Zagreb Index	15166	int64
26	Molecular Formula	15166	object
27	Formal Charge	15166	int64
28	Formal Charge (pos)	15166	int64
29	Formal Charge (neg)	15166	int64
30	Heavy Atoms Count	15166	int64
31	Molar Mass	15166	object
32	SP3 Character	15166	float64
33	Rotatable Bonds Count (non-terminal)	15166	int64

DATASET

Cleaning data process

- Normalisation of column names:

  Lower(), replace(), rename()

- Transformation of numeric columns:

  To_numeric()

- Data evaluation:

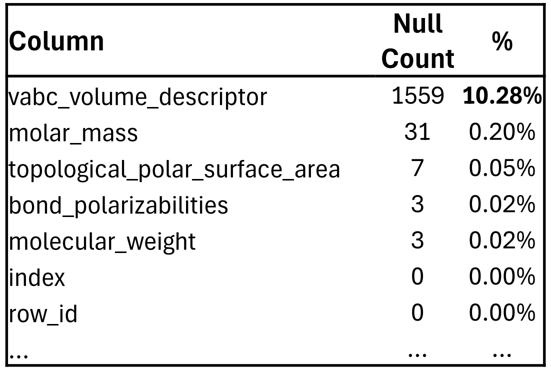
  Isnull().sum(), duplicated()

- Drop rows containing null values:

  [‘molar_mass’, ‘tpsa’, ‘bond_polarizabilities,
  ‘molecular_weight’]

- Drop columns:

  [‘vabc_volume_descriptor’, 'row_id', 'molecule_name’]

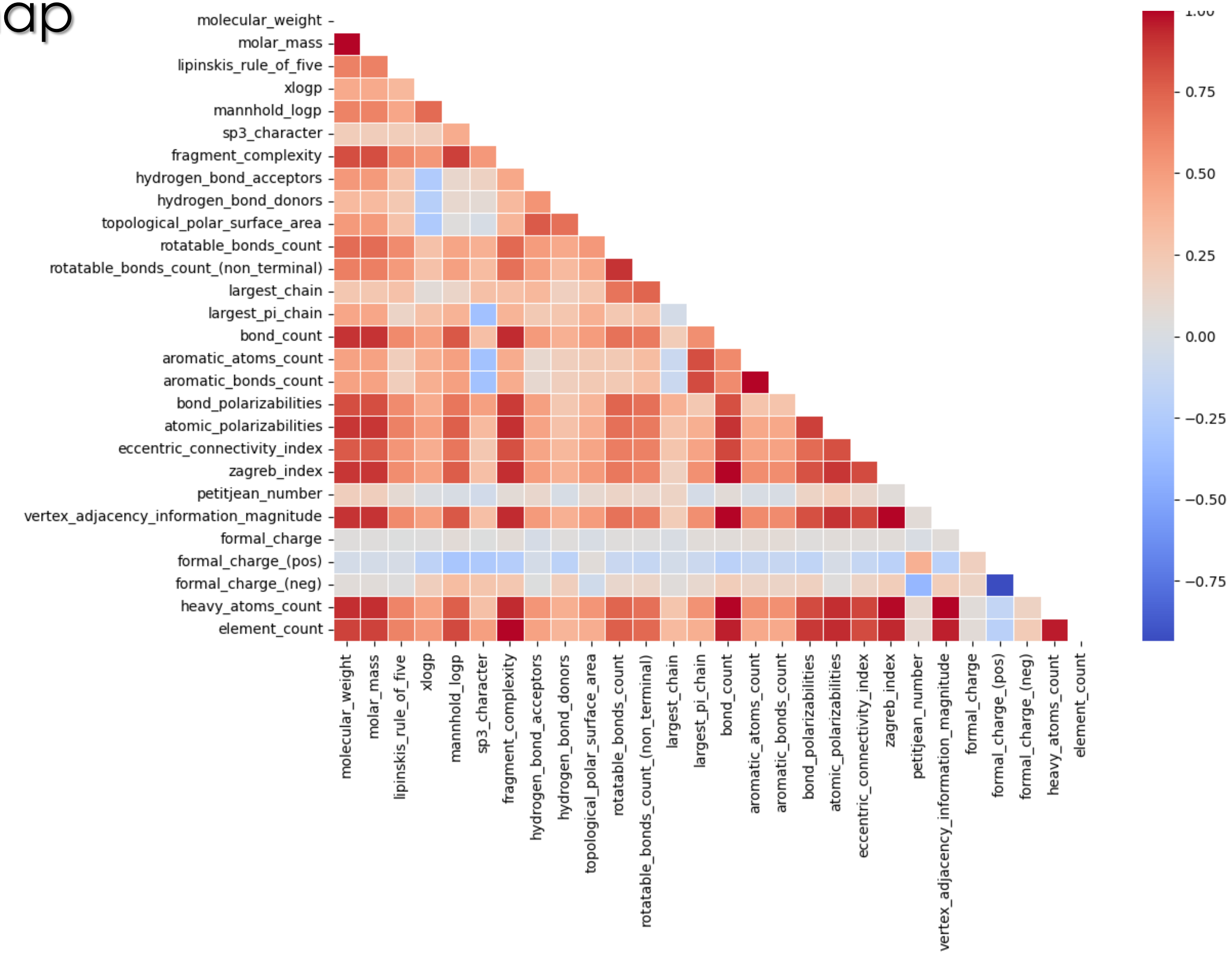


Column	**Null Count**	**%**
vabc_volume_descriptor	1559	**10.28%**
molar_mass	31	0.20%
topological_polar_surface_area	7	0.05%
bond_polarizabilities	3	0.02%
molecular_weight	3	0.02%
index	0	0.00%
row_id	0	0.00%
...	...	...

EXPLORATORY DATA ANALYSIS

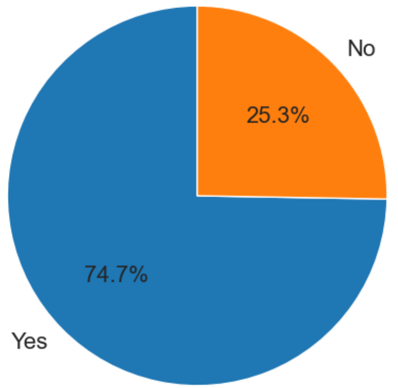
Correlation map

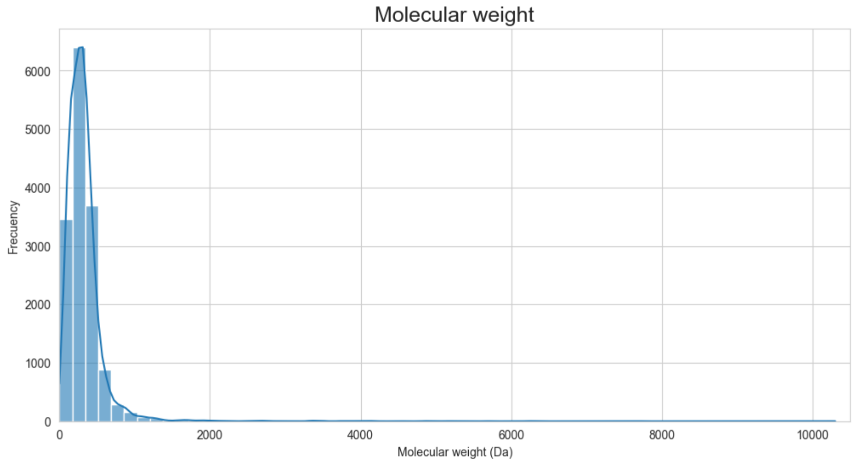
Method: Spearman

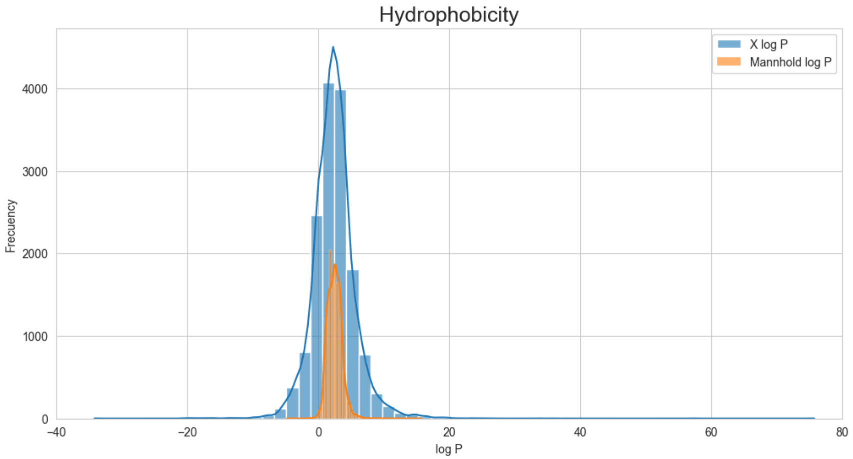


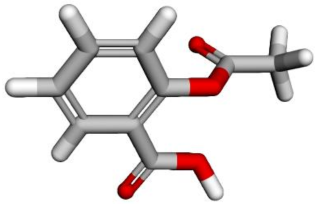
EXPLORATORY DATA ANALYSIS

Molecules that meet Lipinski's criteria

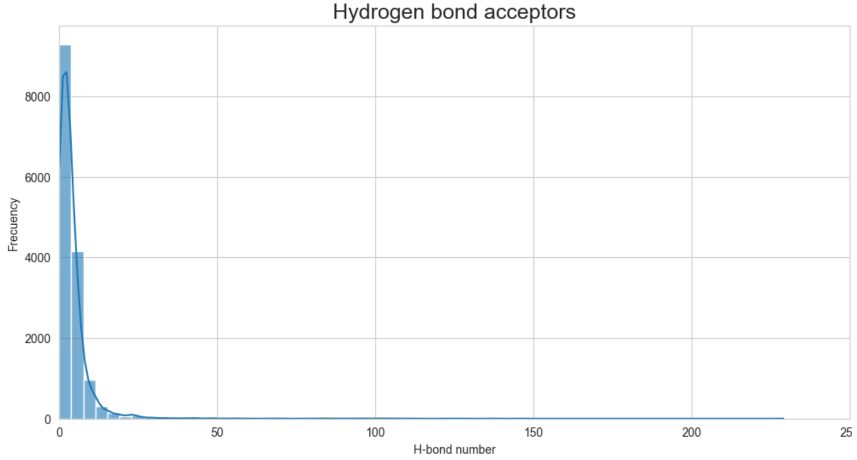


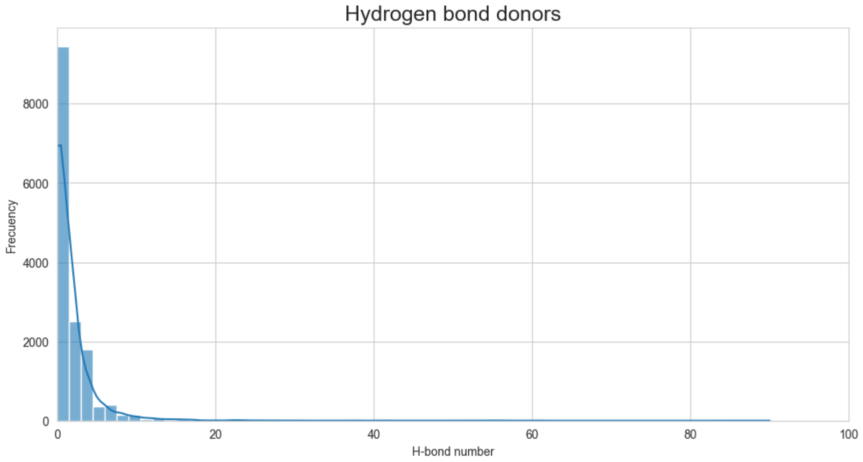

Molecular weight

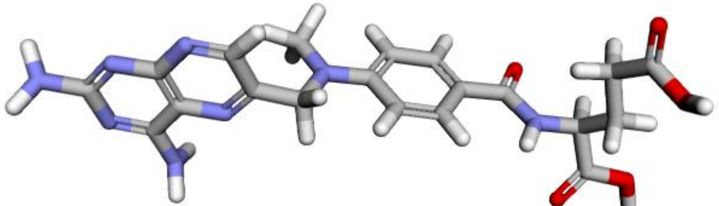

Hydrophobicity



Acetylsalicylic acid – Aspirin ($C_9H_8O_4$)
Pain killer and fever reduction


Hydrogen bond acceptors

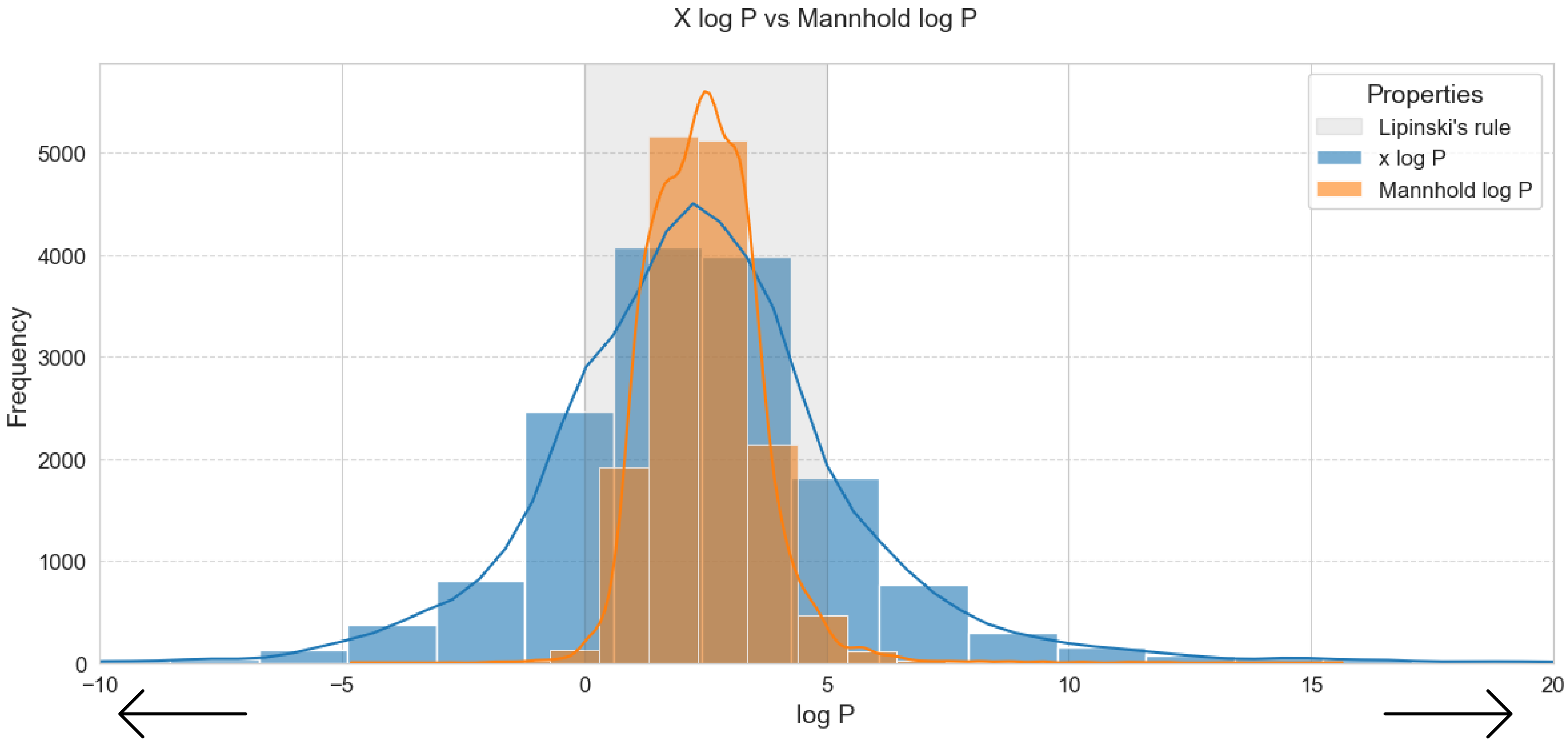

Hydrogen bond donors



Methotrexate ($C_{20}H_{22}N_8O_5$)
Treatment of cancer and autoimmune diseases

EXPLORATORY DATA ANALYSIS

Hydrophobicity determination

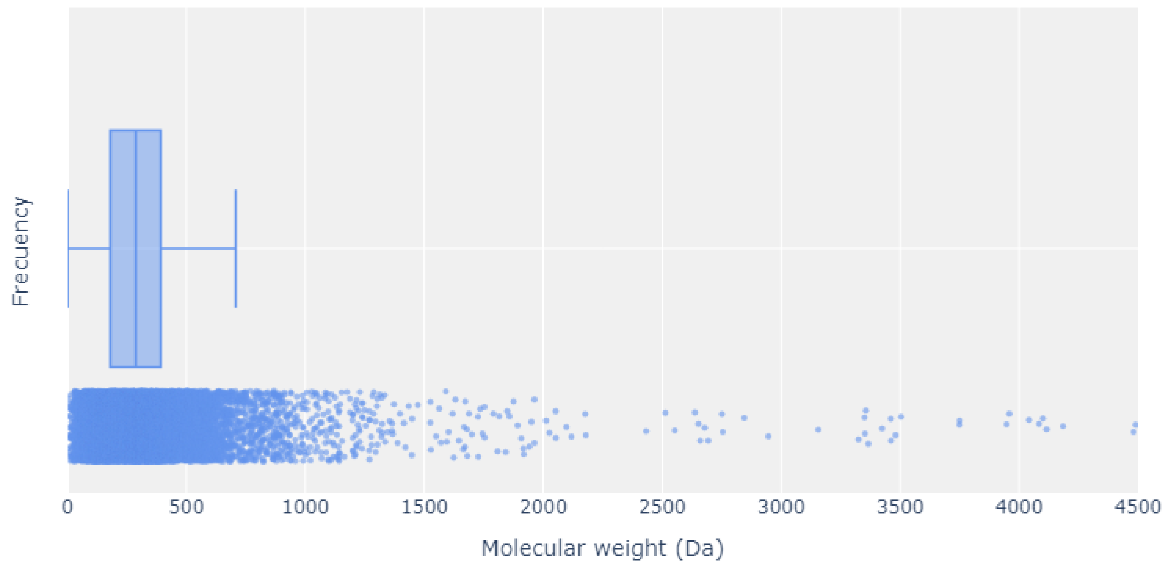


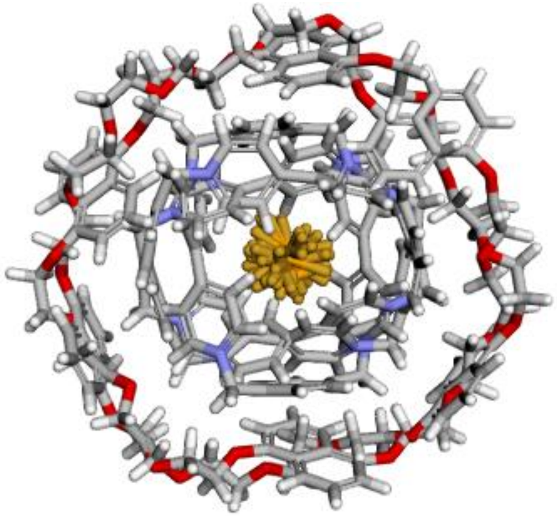
Hydrophilic

Hydrophobic

EXPLORATORY DATA ANALYSIS

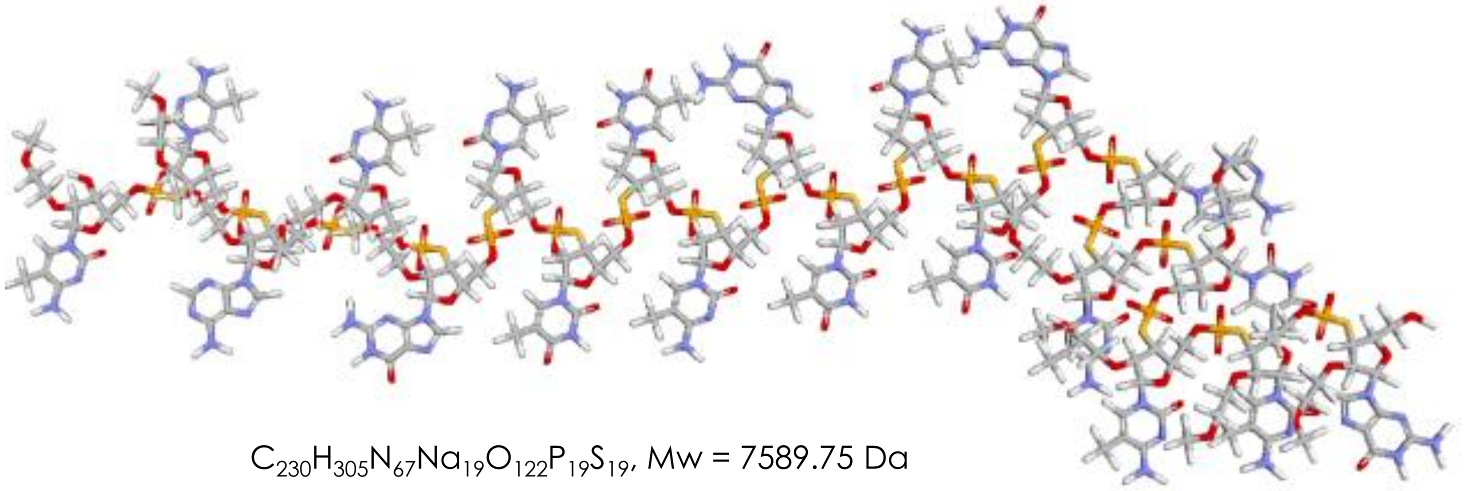
Molecular weight distribution





$C_{216}H_{228}F_{72}N_{12}O_{30}P_{12}$, Mw = 5209.23 Da

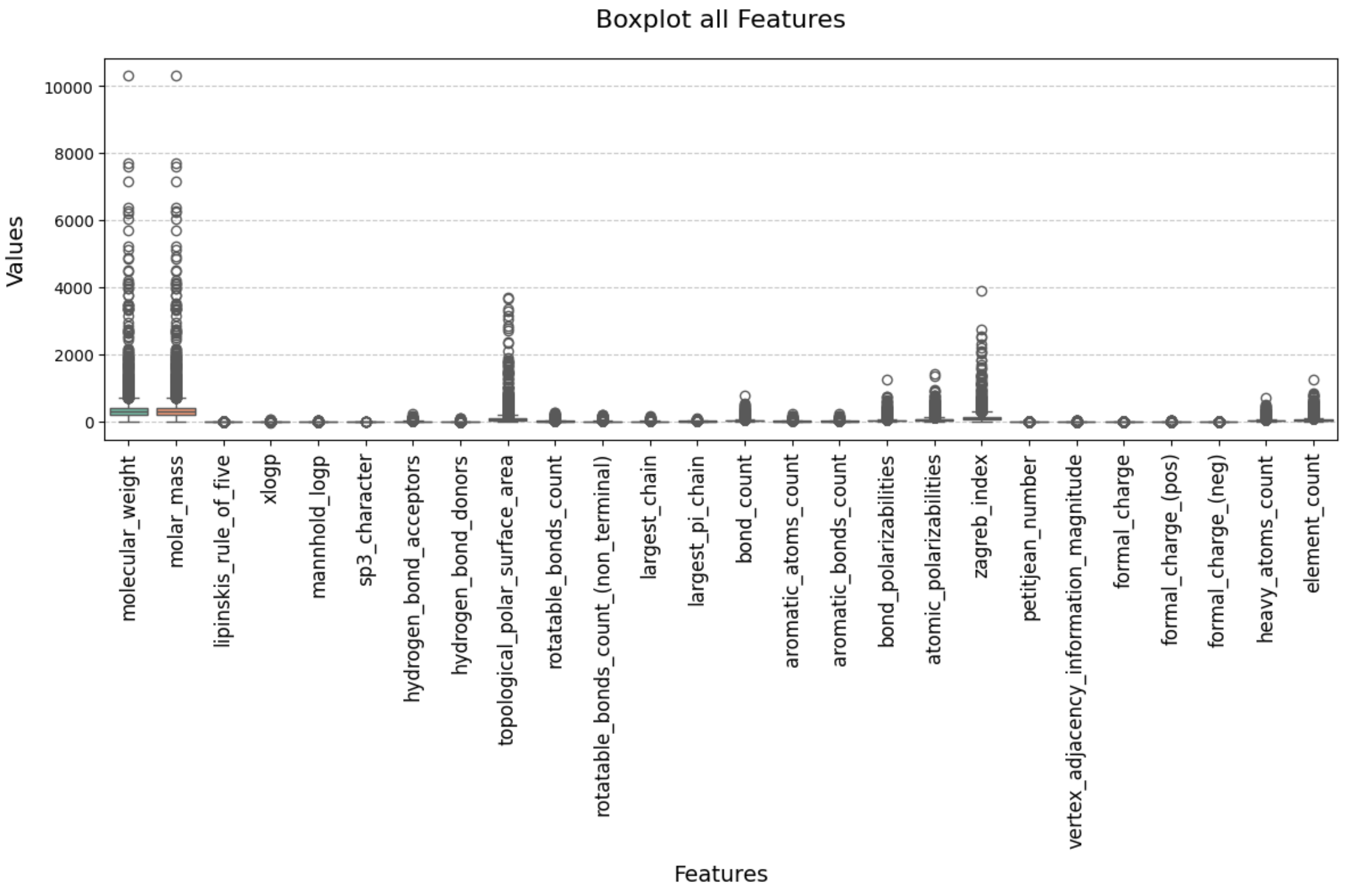
- Many outliers observed in the boxplot.
- Those outliers correspond to large-sized molecules.
- RDKit Python Library for molecule representation.

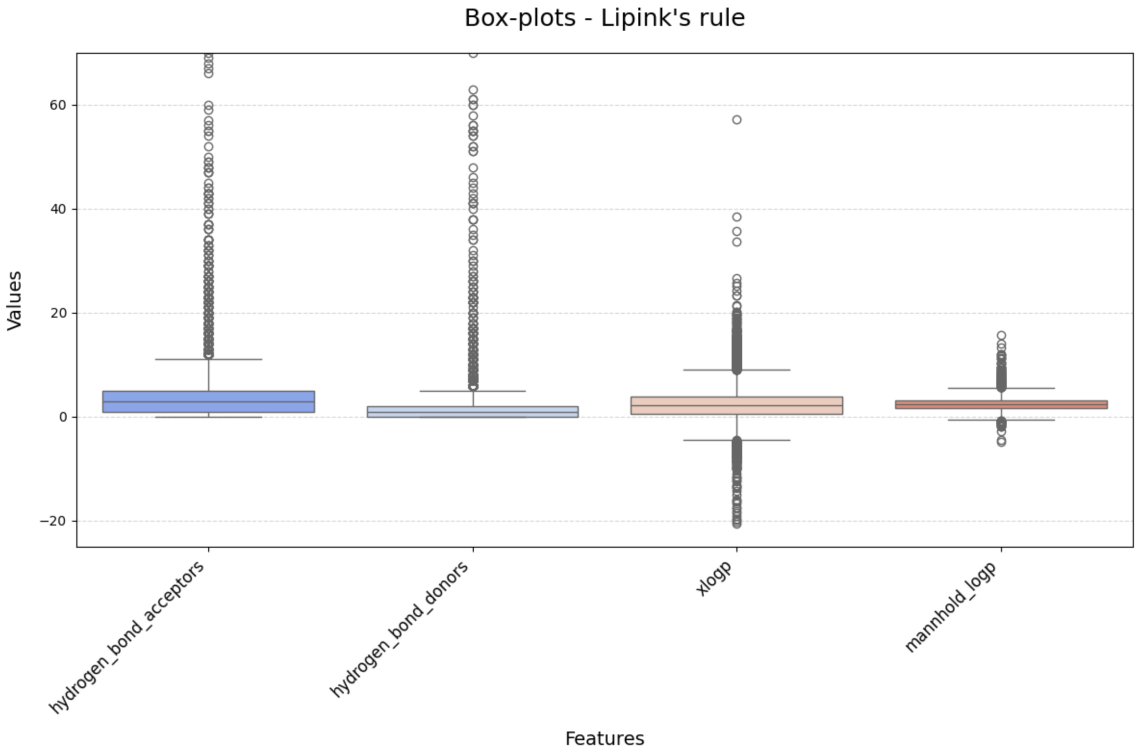


$C_{230}H_{305}N_{67}Na_{19}O_{122}P_{19}S_{19}$, Mw = 7589.75 Da

EXPLORATORY DATA ANALYSIS

Boxplots



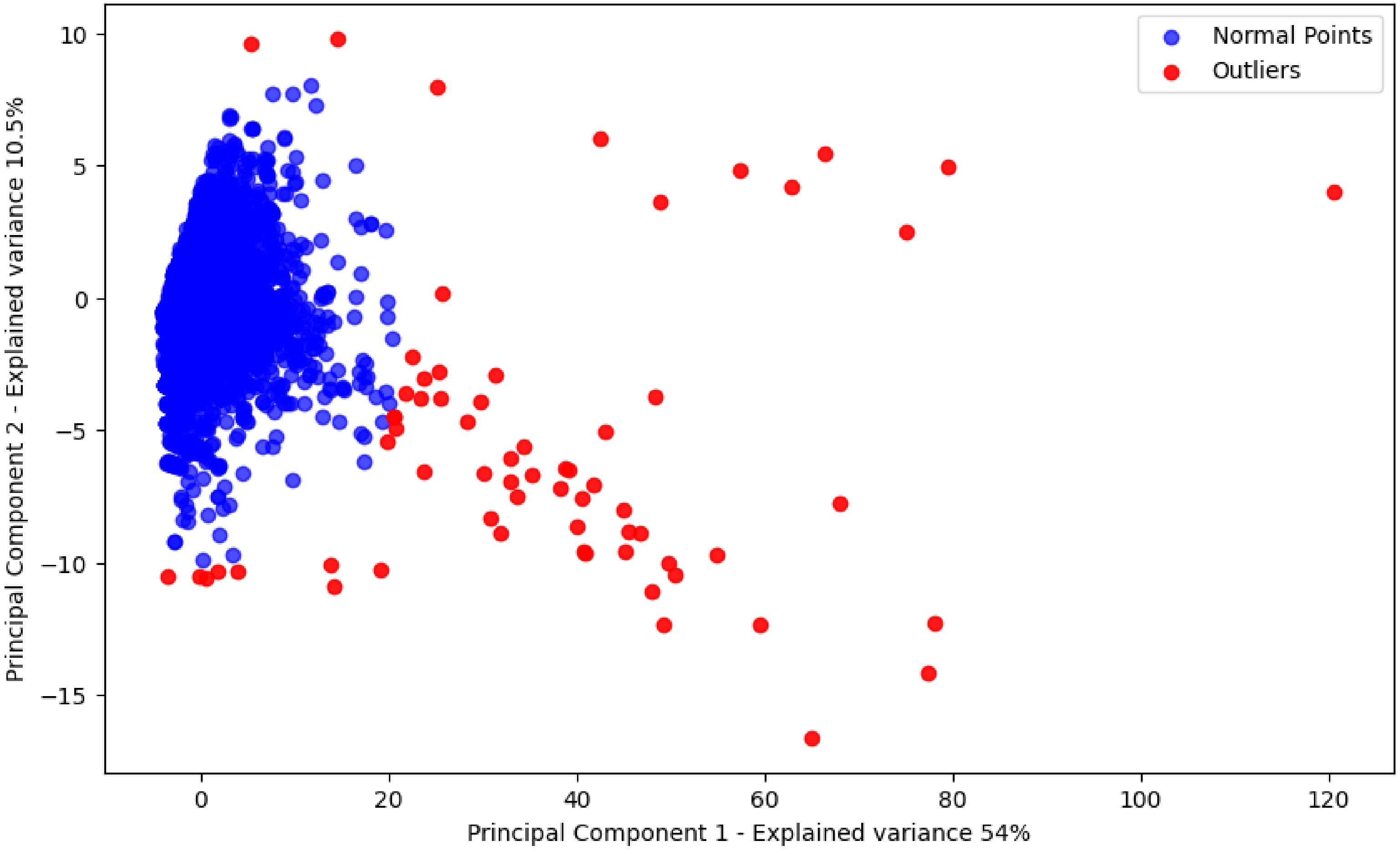


💊 Many outliers observed in the boxplot corresponding to large-sized molecules.

EXPLORATORY DATA ANALYSIS

PCA analysis

PCA Scatter Plot with Possible Outliers



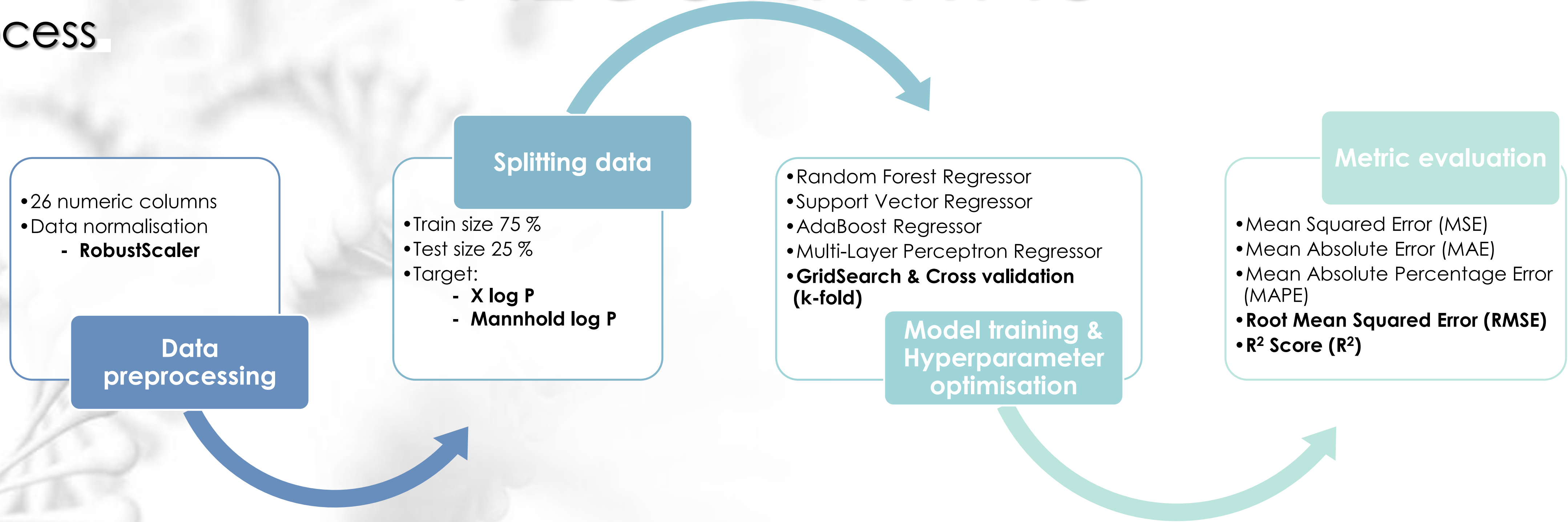
StandardScale was applied before PCA analysis

Mahalanobis distance: Compute how far each point is from the mean in PCA space,

Set a threshold using **Chi-square** distribution (confidence level = 95%)

MACHINE LEARNING ALGORITHMS

Process



Data preprocessing

- 26 numeric columns
- Data normalisation
  - **RobustScaler**

Splitting data

- Train size 75 %
- Test size 25 %
- Target:
  - **X log P**
  - **Mannhold log P**

Model training & Hyperparameter optimisation

- Random Forest Regressor
- Support Vector Regressor
- AdaBoost Regressor
- Multi-Layer Perceptron Regressor
- **GridSearch & Cross validation (k-fold)**

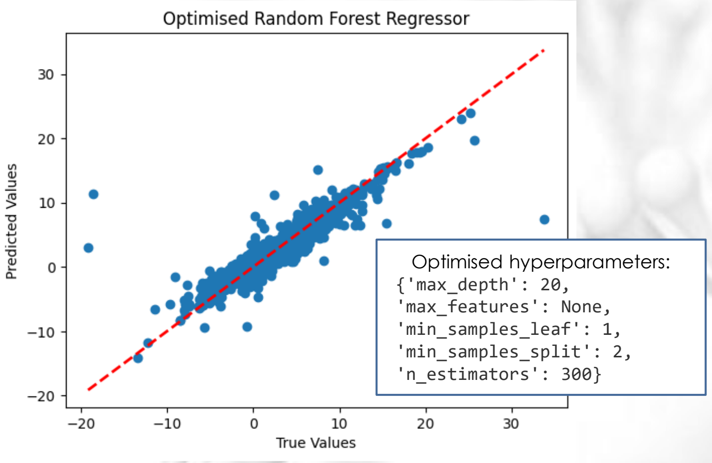
Metric evaluation

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- **Root Mean Squared Error (RMSE)**
- **R^2 Score (R^2)**

Random Forest Regressor	Support Vector Regressor	AdaBoost Regressor	Multi-Layer Perceptron
• Combines **decision trees** to reduce variance. • Handles **nonlinear data** well. • Less effective for high-dimensional data.	• Uses a **hyperplane** to minimize errors. • Ideal for **small**, well-distributed datasets. • Sensitive to scaling; computationally expensive for large datasets.	• Combines weak predictors, **focusing on errors.** • Effective with moderate noise. • Prone to overfitting with outliers.	• Artificial **neural network** for nonlinear data. • Requires proper setup and large datasets. • Susceptible to overfitting if not well-tuned.

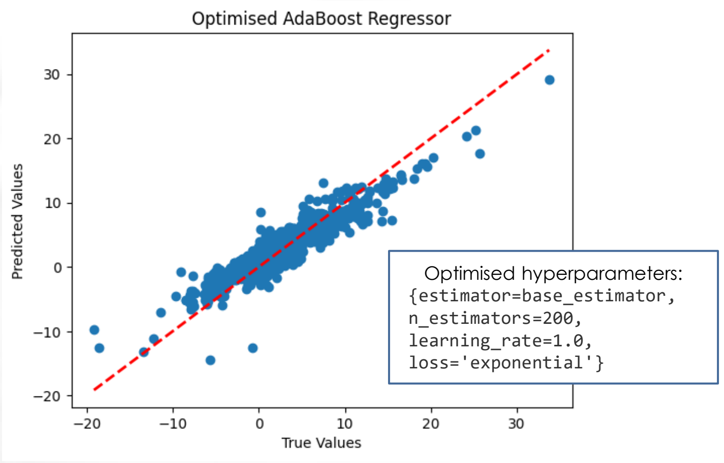
MACHINE LEARNING ALGORITHMS

Results for Target: **X log P** – True vs Predicted values



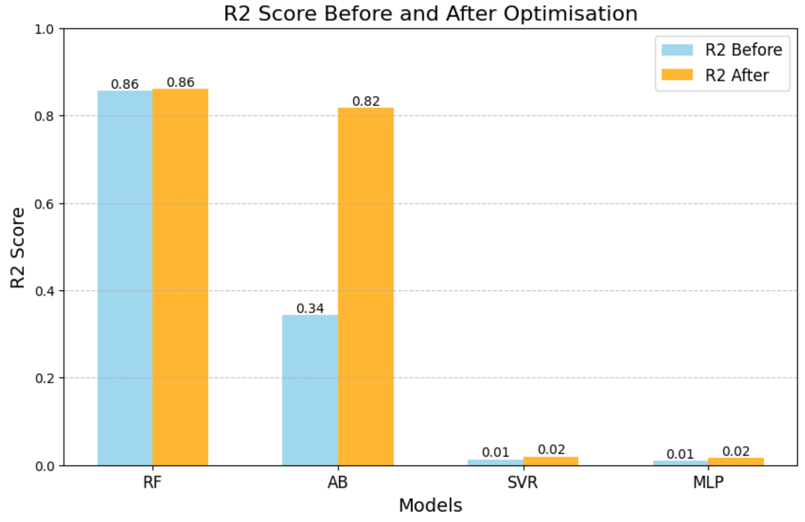
Optimised hyperparameters:

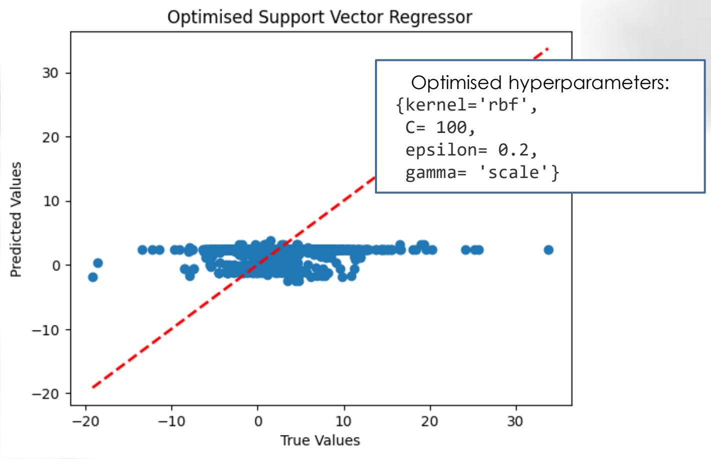
```
{'max_depth': 20,
'max_features': None,
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 300}
```



Optimised hyperparameters:

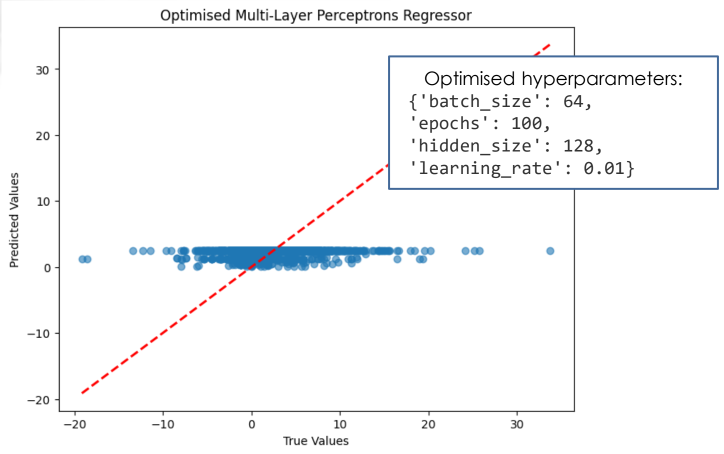
```
{estimator=base_estimator,
n_estimators=200,
learning_rate=1.0,
loss='exponential'}
```





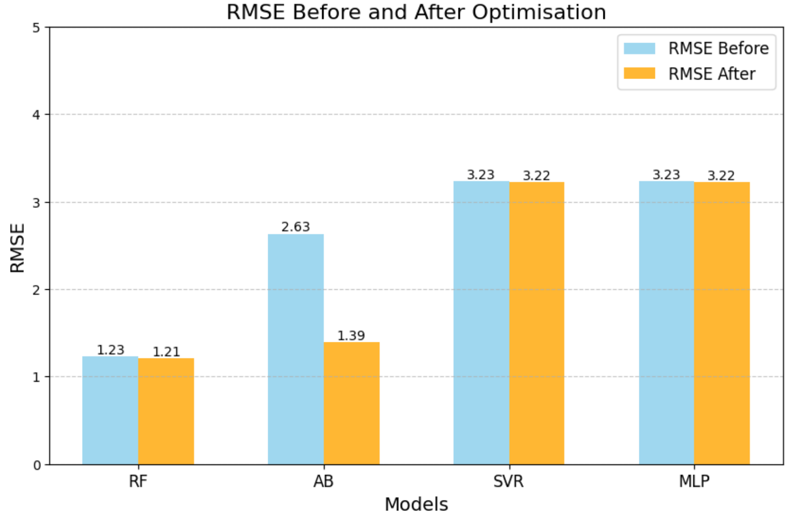
Optimised hyperparameters:

```
{kernel='rbf',
 C= 100,
 epsilon= 0.2,
 gamma= 'scale'}
```



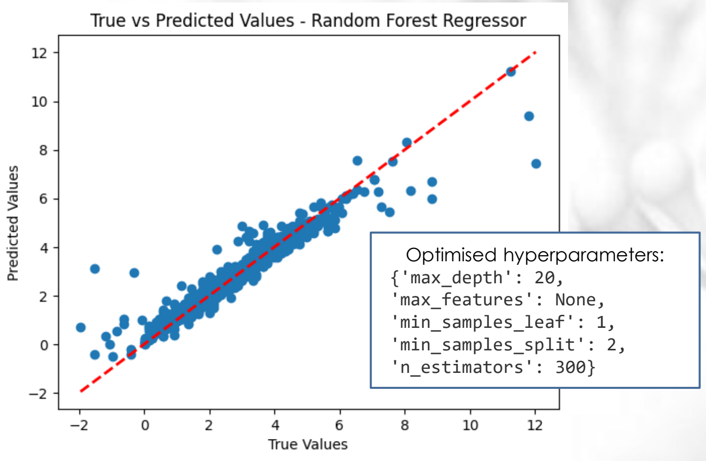
Optimised hyperparameters:

```
{'batch_size': 64,
'epochs': 100,
'hidden_size': 128,
'learning_rate': 0.01}
```



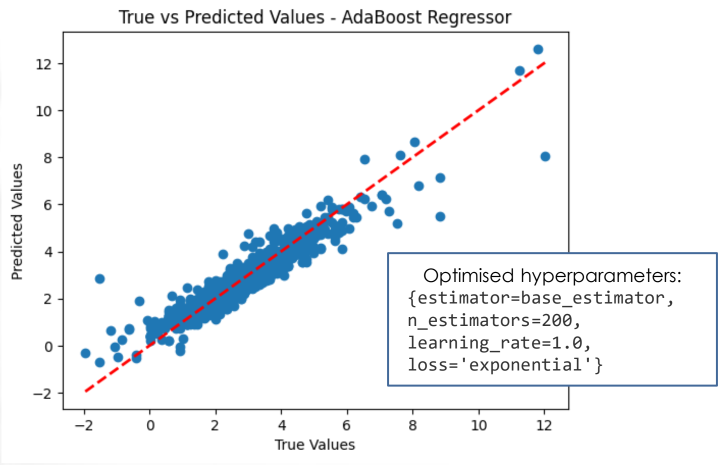
MACHINE LEARNING ALGORITHMS

Results for Target: **Mannhold X log P** – True vs Predicted values


True vs Predicted Values - Random Forest Regressor

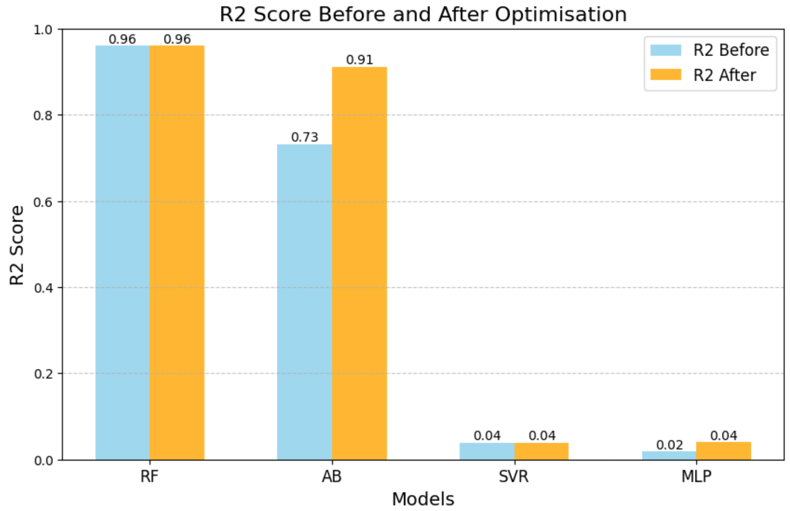
Optimised hyperparameters:

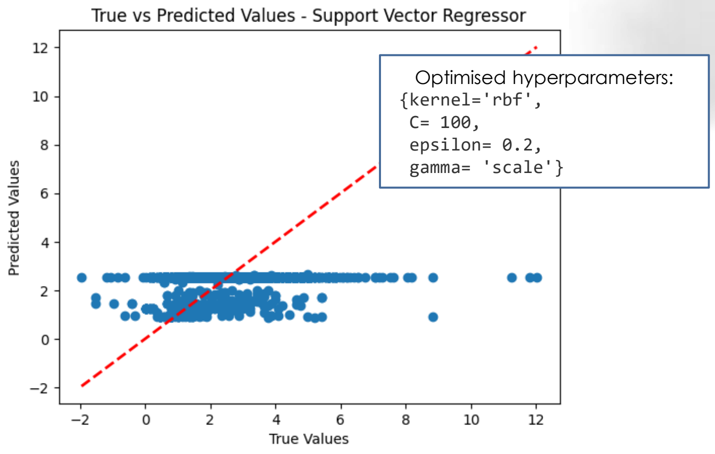
```
{'max_depth': 20,
'max_features': None,
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 300}
```


True vs Predicted Values - AdaBoost Regressor

Optimised hyperparameters:

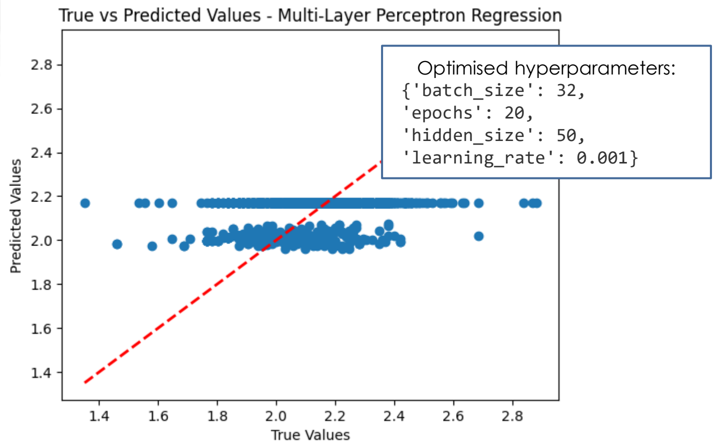
```
{estimator=base_estimator,
n_estimators=200,
learning_rate=1.0,
loss='exponential'}
```


R2 Score Before and After Optimisation


True vs Predicted Values - Support Vector Regressor

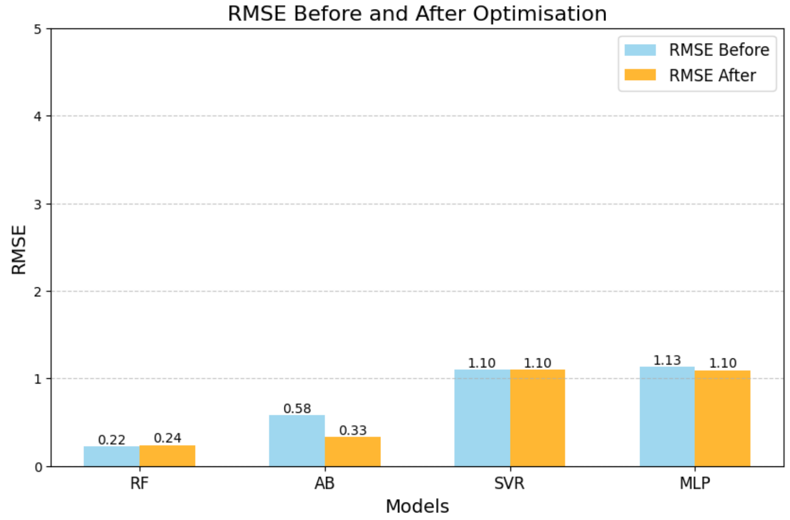
Optimised hyperparameters:

```
{kernel='rbf',
 C= 100,
 epsilon= 0.2,
 gamma= 'scale'}
```


True vs Predicted Values - Multi-Layer Perceptron Regression

Optimised hyperparameters:

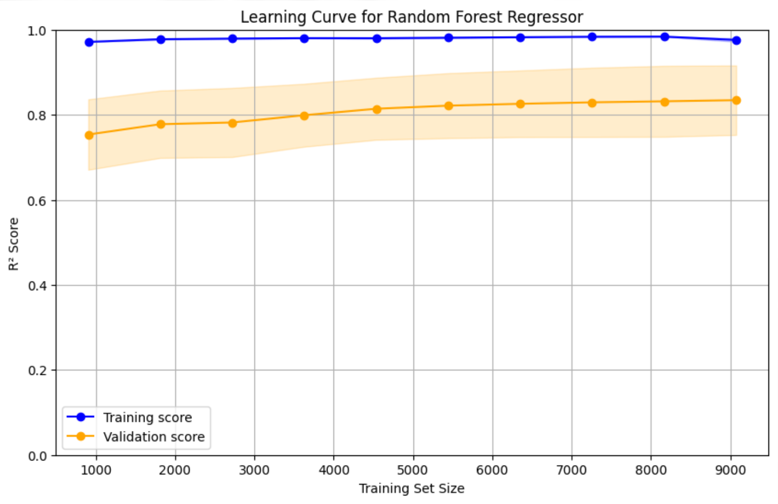
```
{'batch_size': 32,
'epochs': 20,
'hidden_size': 50,
'learning_rate': 0.001}
```

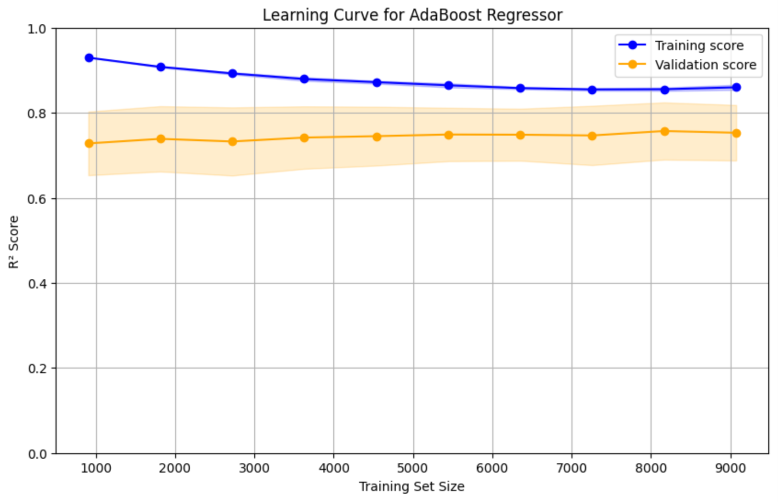

RMSE Before and After Optimisation

MACHINE LEARNING ALGORITHMS

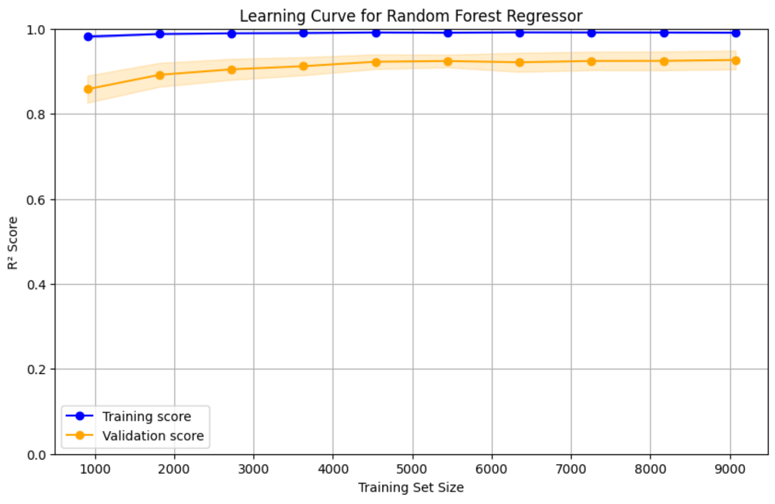
Model fitting evaluation

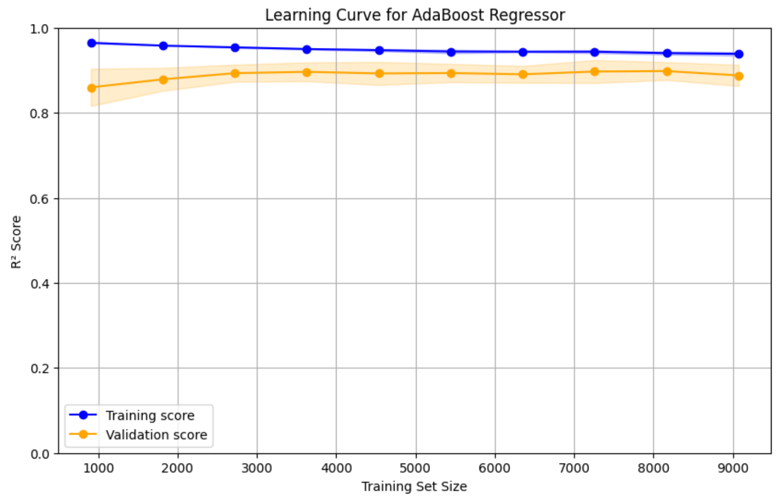
Target: X log P


Learning Curve for Random Forest Regressor


Learning Curve for AdaBoost Regressor

Target: Mannhold log P


Learning Curve for Random Forest Regressor

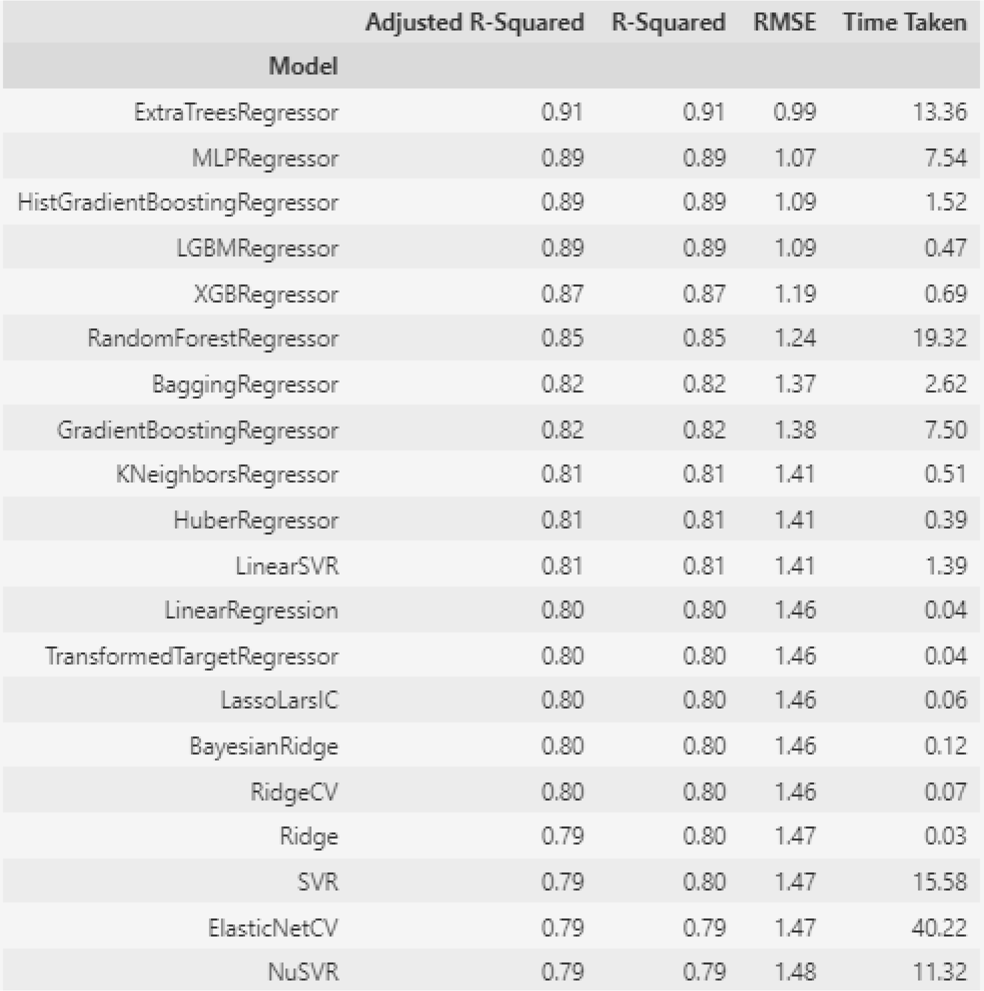

Learning Curve for AdaBoost Regressor

FUTURE WORK

Improvements

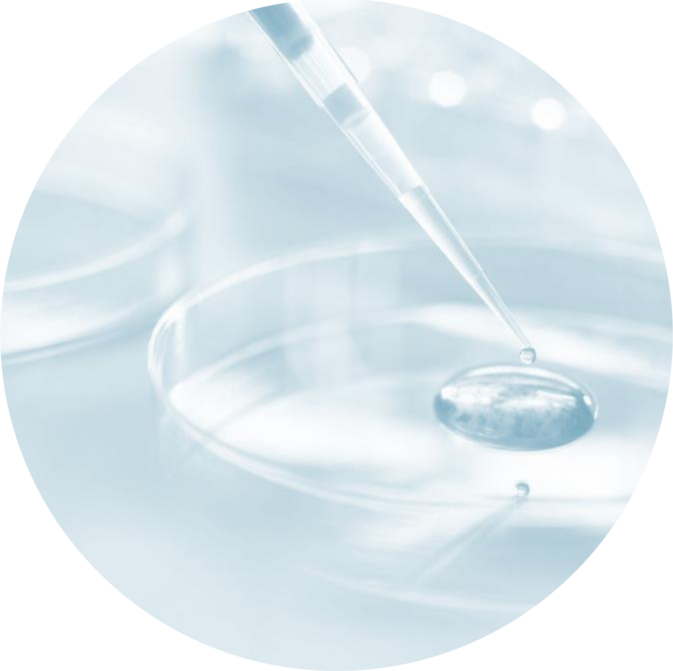
<u>LazyRegressor</u> is a Python library that quickly **compares** the performance of multiple **regression models** on a dataset. It provides metrics like R², RMSE, and execution time for various algorithms.

SVR and MLP may be optimised. Other models can be applied.



	Adjusted R-Squared	R-Squared	RMSE	Time Taken
Model				
ExtraTreesRegressor	0.91	0.91	0.99	13.36
MLPRegressor	0.89	0.89	1.07	7.54
HistGradientBoostingRegressor	0.89	0.89	1.09	1.52
LGBMRegressor	0.89	0.89	1.09	0.47
XGBRegressor	0.87	0.87	1.19	0.69
RandomForestRegressor	0.85	0.85	1.24	19.32
BaggingRegressor	0.82	0.82	1.37	2.62
GradientBoostingRegressor	0.82	0.82	1.38	7.50
KNeighborsRegressor	0.81	0.81	1.41	0.51
HuberRegressor	0.81	0.81	1.41	0.39
LinearSVR	0.81	0.81	1.41	1.39
LinearRegression	0.80	0.80	1.46	0.04
TransformedTargetRegressor	0.80	0.80	1.46	0.04
LassoLarsIC	0.80	0.80	1.46	0.06
BayesianRidge	0.80	0.80	1.46	0.12
RidgeCV	0.80	0.80	1.46	0.07
Ridge	0.79	0.80	1.47	0.03
SVR	0.79	0.80	1.47	15.58
ElasticNetCV	0.79	0.79	1.47	40.22
NuSVR	0.79	0.79	1.48	11.32

MACHINE LEARNING INSIGHTS FOR PREDICTING ORAL DRUGS PROPERTIES

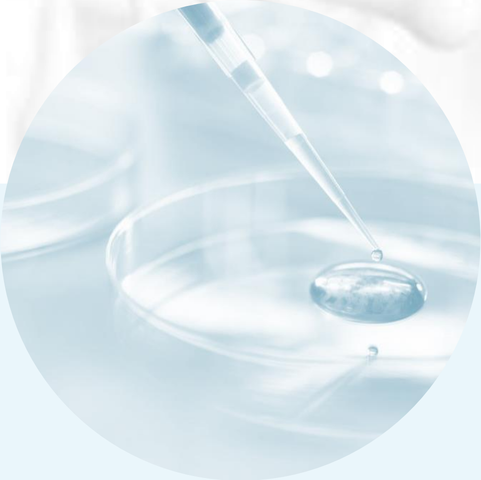


CONCLUSIONS

- **It was possible to use machine learning** algorithms to **predict two important parameters used** for drug discovery (X log P and Mannhold log P).
- **EDA** was essential and showed the needed of **RobustScale** standardization before applying the models.
- **SVR and MLP** were not suitable models for this dataset being also more complex to implement and optimize.
- **Random Forest and AdaBoost algorithm** showed the best model performances with low errors and good R^2 values.

REFERENCES

- *Analytics+*. (n.d.), https://analyticsplus.org/glossary/
- Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y. A. M. M., Gomaa, M. M., & Hassanien, A. E. (2023). Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, *56*(7), 5975–6037. https://doi.org/10.1007/s10462-022-10306-1
- Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, *25*(3), 1315–1360. https://doi.org/10.1007/s11030-021-10217-3
- Lipinski, C. A., Dominy, B. W., & Feeney, P. J. (1997). drug delivery reviews Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. In *Advanced Drug Delivery Reviews* (Vol. 23).
- Obaido, G., Mienye, I. D., Egbelowo, O. F., Emmanuel, I. D., Ogunleye, A., Ogbuokiri, B., Mienye, P., & Aruleba, K. (2024). Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Machine Learning with Applications*, *17*, 100576. https://doi.org/10.1016/j.mlwa.2024.100576
- Shin, K. K., Ong, L., Kayne, K. M., & Sow, W. (n.d.). *CM4044 Project 2 LogP Prediction*.
- Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? In *Acta Pharmaceutica Sinica B* (Vol. 12, Issue 7, pp. 3049–3062). Chinese Academy of Medical Sciences. https://doi.org/10.1016/j.apsb.2022.02.002



THANK YOU

LUCIANA OLIVEIRA & MARÍA URIBURU GRAY
DSPP02 – CODEOP

Extra slides

MACHINE LEARNING ALGORITHMS

X log P

Model	MSE	MAE	MAPE	RMSE	R^2
RF	1.47	0.64	25700739400780	1.21	0.8608
AB	1.93	1.05	91363261747935	1.39	0.8174
SVR	10.36	2.23	154865117739752	3.22	0.0193
MLP	10.39	2.26	164907121406810	1.07	0.0168

Mannhold log P

Model	MSE	MAE	MAPE	RMSE	R^2
RF	0.06	0.11	0.07	0.24	0.96
SVR	1.20	0.83	0.60	1.10	0.0390
AB	0.11	0.24	0.17	0.33	0.9109
MLP	1.19	0.83	0.65	1.095	0.0402