Boiling Point Predictor

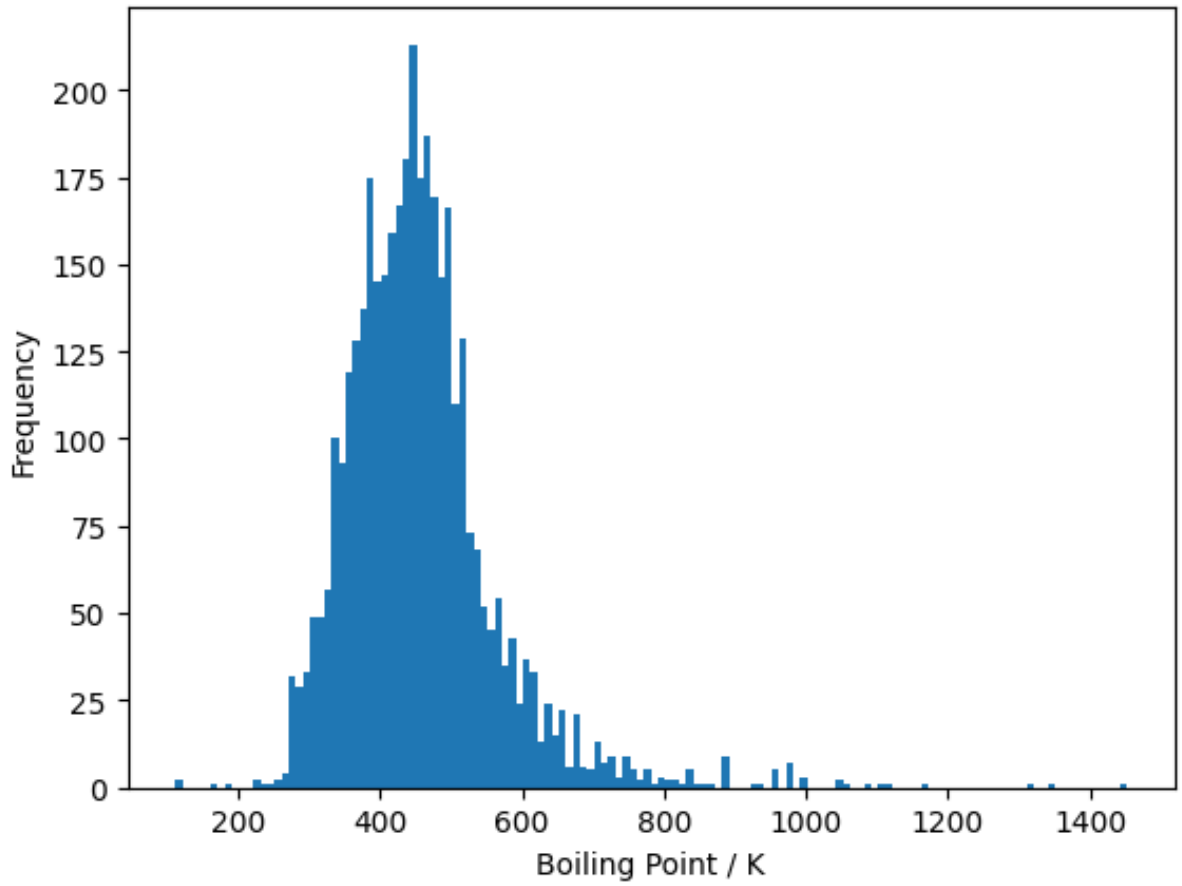
Predict the boiling points of organic molecules using molecular fingerprints and descriptors.

Learning Objectives

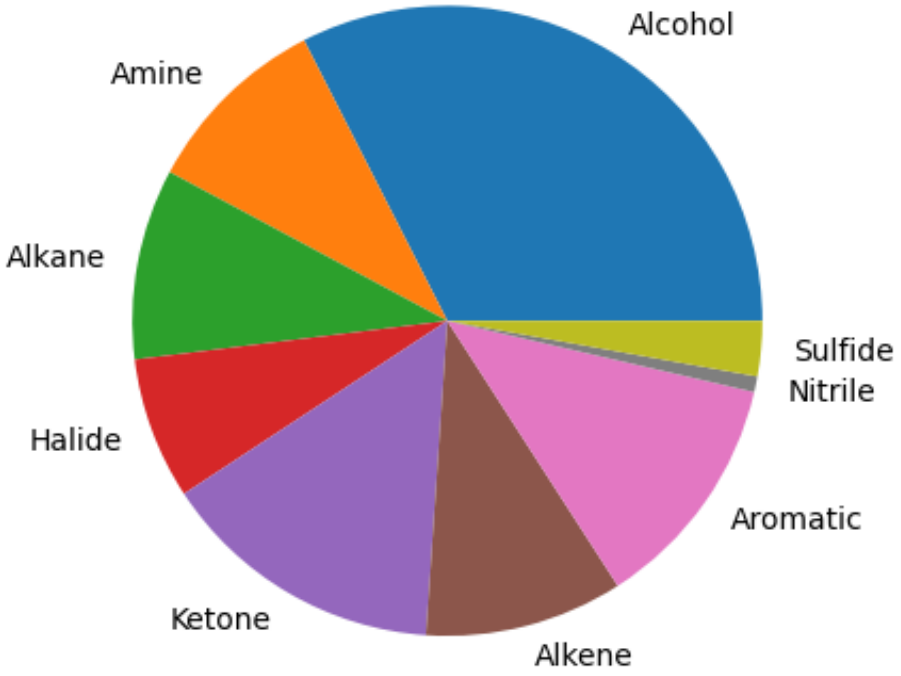
- Molecular fingerprinting
- Molecular descriptors
- SHAP model interpretation

Visualising the Data

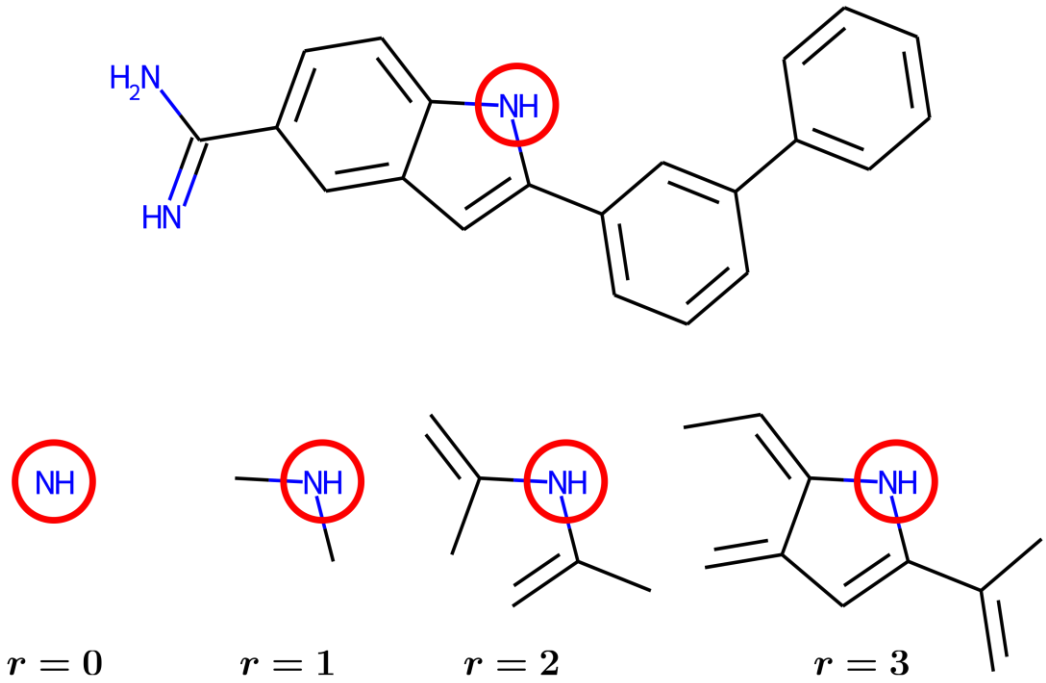
The boiling points (in K) and SMILES representations of 3790 diverse molecules were collected. The composition of the dataset is visualised below.



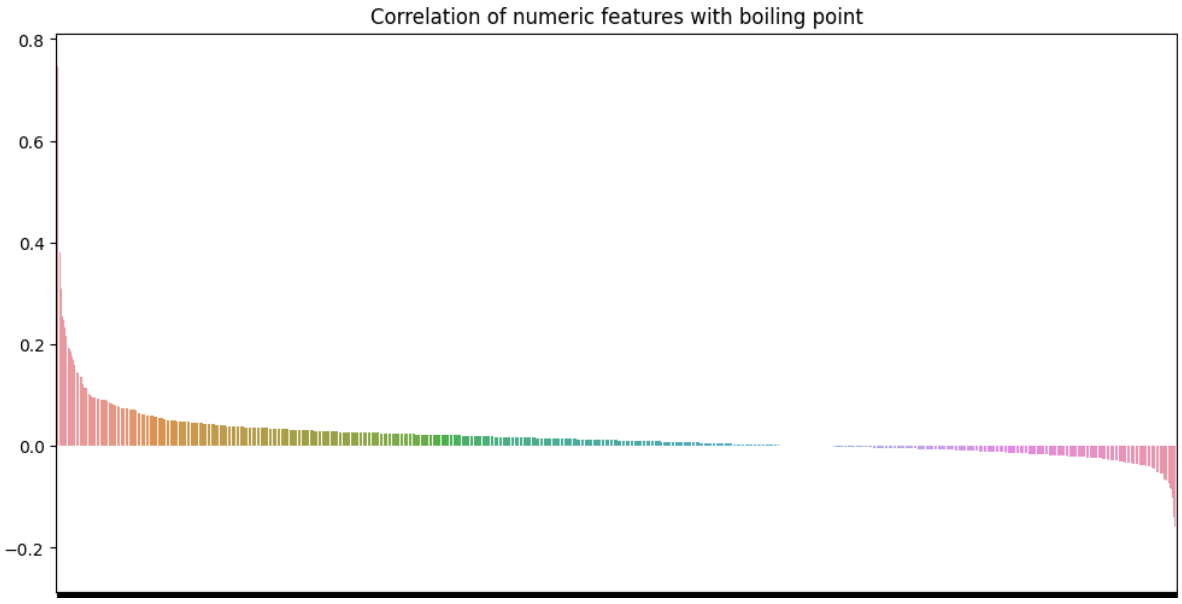
The distribution of boiling points is skewed with a tail towards higher boiling points.


Alcohol
Amine
Alkane
Halide
Ketone
Alkene
Aromatic
Nitrile
Sulfide

Morgan Fingerprints



Morgan extended connectivity fingerprints were constructed for each molecule using RDKit, using a radius of 3 and a length of 2048. Features that had a variance of less than 0.1% were dropped, resulting in a total of 1205 features for training. The correlation of these features with the boiling point is shown below.



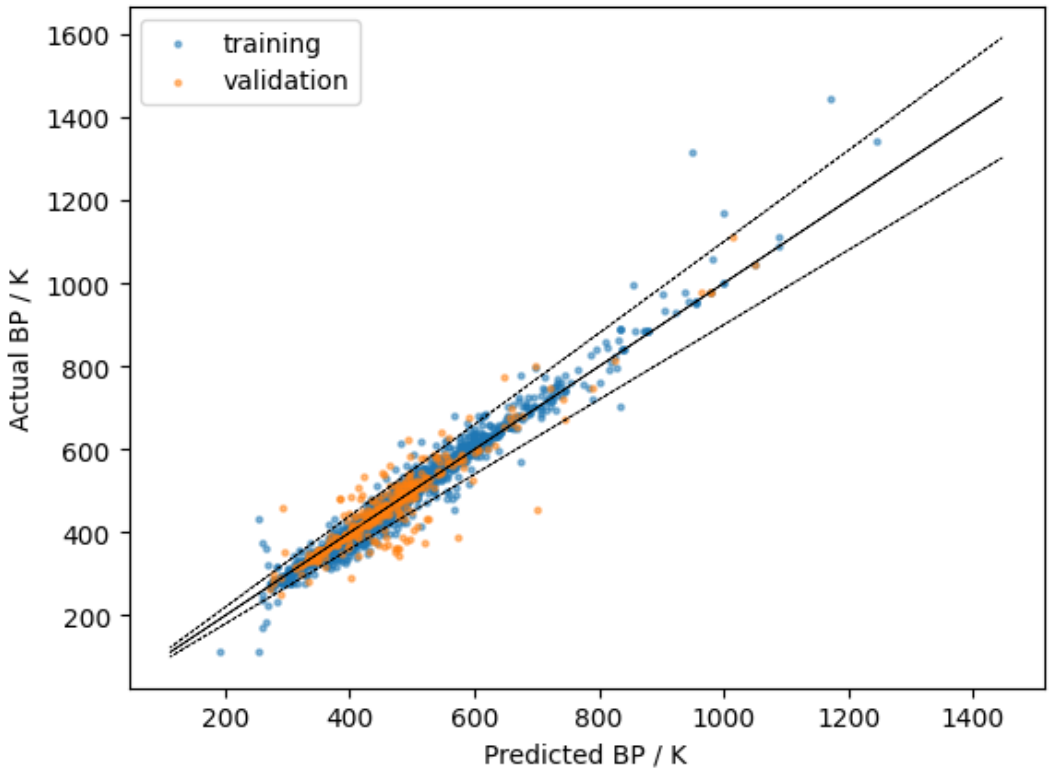
Evidently, most features in the fingerprints are only weakly correlated with the boiling point.

The total data was split into a training and testing set in a 9:1 split. An XGBoost model was fitted to the training data. Its hyperparameters were tuned using a Randomised CV Search with 5-fold cross validation.

The final model achieved a training score of 0.9470 and a testing score of 0.8117.

A Random Forest Regressor was also fitted to the training data using 50 estimators. This achieved higher training and testing scores of 0.9612 and 0.8325, respectively.

A plot of the prediction vs actual boiling points for the Random Forest model is given below. Lines have been drawn showing the +/- 10% confidence interval.



Testing Data

SMILES	Predicted BP / K	Actual BP / K	Difference
SCCS	428.111	419.250	2.1%
COC(CCl)=O	396.771	403.250	1.6%
CCCCCCCC#CCCC	484.928	483.160	0.4%
CCC[N+](=O)[O-]	395.123	404.450	2.3%
CCC(C)C(SC1C)N=C1C	477.263	344.250	38.6%

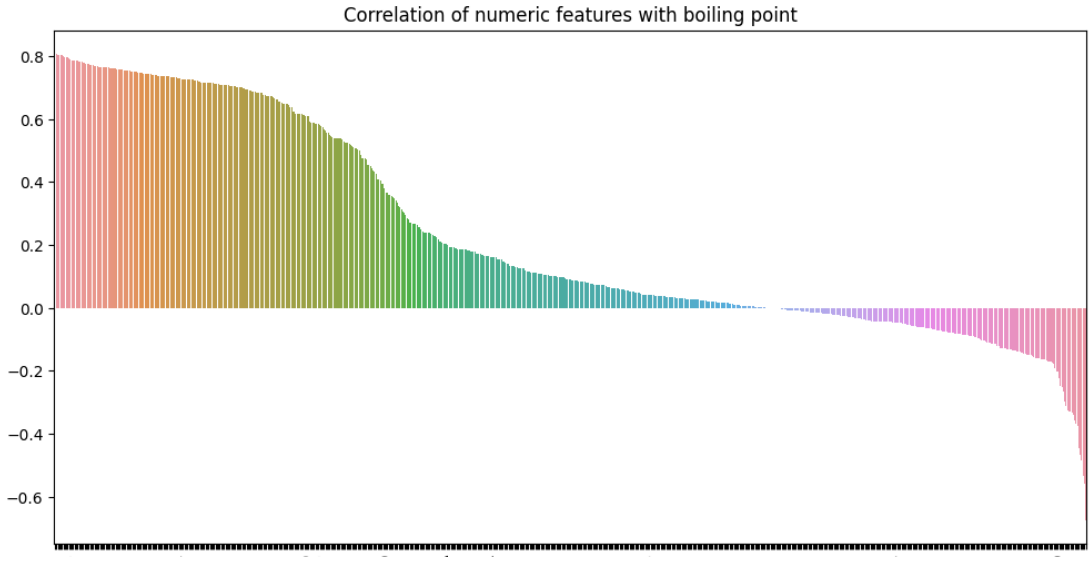
Least Accurate Predictions

SMILES	Predicted BP / K	Actual BP / K	Difference
C	254.000	111.650	127.5%
[2H]C([2H])([2H])[2H]	190.848	112.150	70.2%
C=C	259.615	169.350	53.3%
SCCCCCCCCCS	574.548	389.250	47.6%
CC	266.233	184.550	44.3%

The model appears to perform very poorly on the simplest molecules, giving wildly inaccurate results for methane, ethane, and ethene.

Molecular Descriptors

RDKit and Mordred were used to create 2048 2D and 3D descriptors for each molecule. Features that had a variance of less than 1% were dropped, resulting in a total of 1424 descriptors for training. The correlation of these features with the boiling point is shown below.

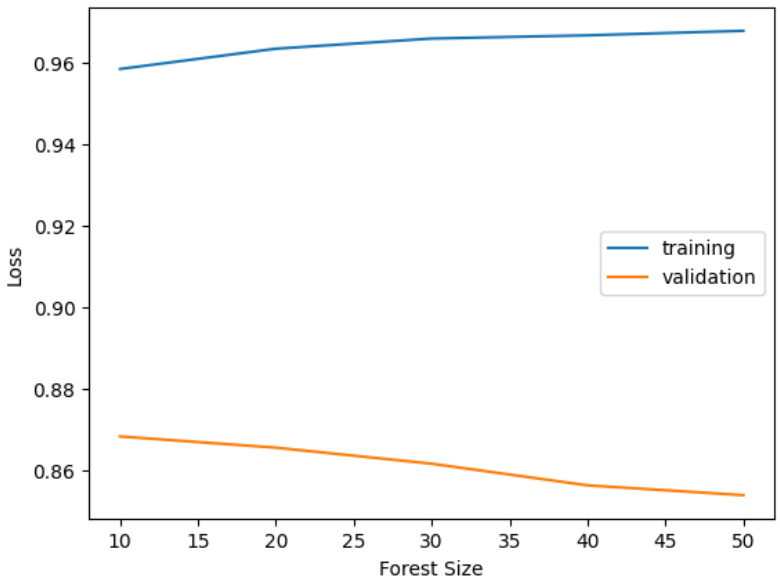


These descriptors are significantly more correlated with the boiling point than the features from the Morgan fingerprints.

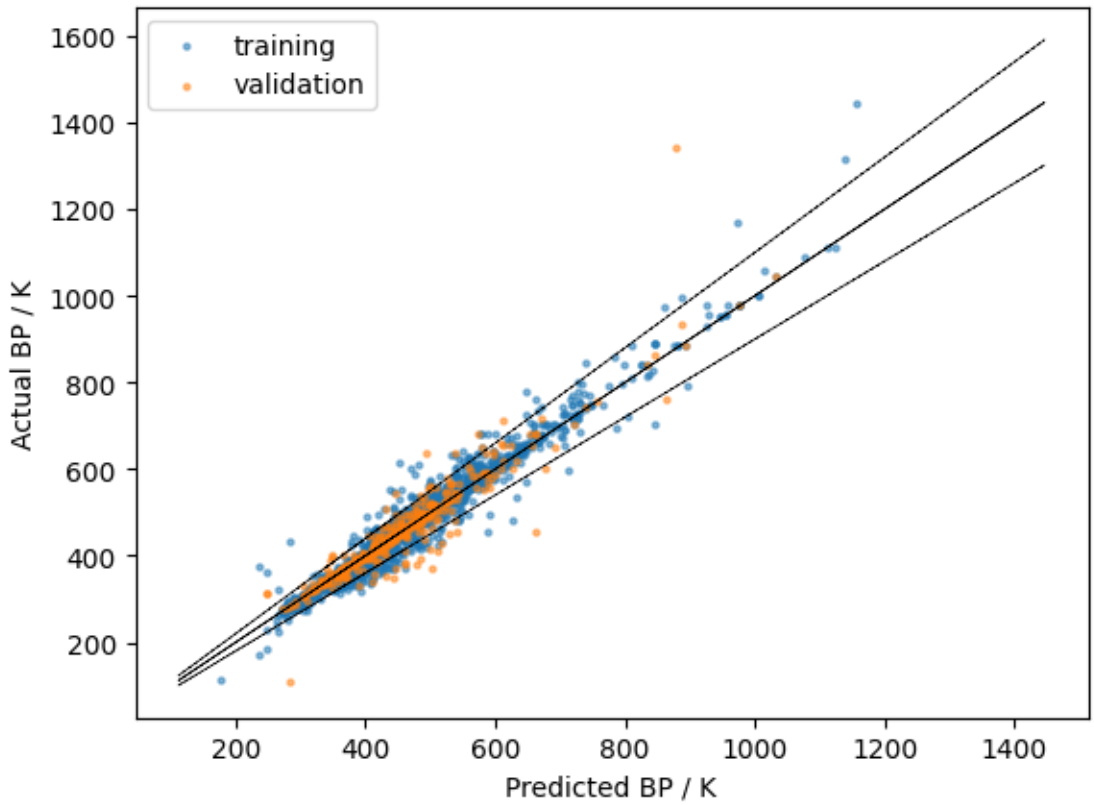
The total data was again split into a training and testing set in a 9:1 split and an XGBoost model was fitted to the training data. Its hyperparameters were tuned using a Randomised CV Search with 5-fold cross validation.

The final model achieved a training score of 0.9891 and a testing score of 0.8156. This very high training score compared to the lower testing score could imply some overfitting for the model.

Several Random Forest Regressors were also fitted to the training data using between 10 to 50 estimators. These all achieved higher testing scores than the XGBoost model, with the top performance using only 10 estimators (training score: 0.9585; testing score: 0.8683).



A plot of the prediction vs actual boiling points for the best performing Random Forest model is given below. Lines have been drawn showing the +/- 10% confidence interval.



Testing Data

SMILES	Predicted BP / K	Actual BP / K	Difference
SCCS	410.559	419.250	2.1%
COC(CCl)=O	398.176	403.250	1.3%
CCCCCCCC#CCCC	484.390	483.160	0.3%
CCC[N+](=O)[O-]	409.139	404.450	1.2%
CCC(C)C(SC1C)N=C1C	379.243	344.250	10.2%

Least Accurate Predictions

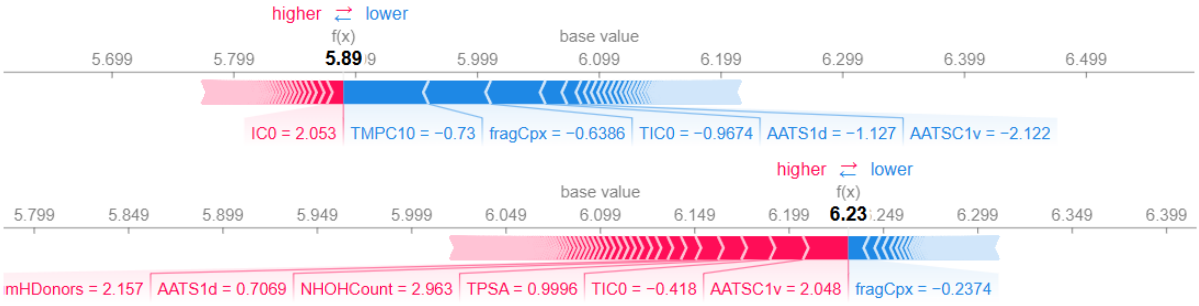
SMILES	Predicted BP / K	Actual BP / K	Difference
C	282.950	111.650	153.4%
[2H]C([2H])([2H])[2H]	177.640	112.150	58.4%
C=C	237.192	169.350	40.1%
O=CCCc1ccccc1	503.112	371.250	35.5%
CC	247.074	184.550	33.9%

94.4% of all molecules in the training and testing datasets scored a percentage difference smaller than 10%. Meanwhile, 84.8% of all molecules scored a percentage difference smaller than 5%, and 51.1% of all molecules scored a percentage difference smaller than 1%.

Model Insights

Permutation importance and SHAP values were used to uncover insights into the trained Random Forest Regressor.

The SHAP force plots demonstrate how each descriptor impacted the final model prediction, examples of which are given below.



Permutation importance can be used to rank the importance of each individual feature in the model as a whole. The top 10 features and their relative weights are given below.

Weight	Feature
0.1901 ± 0.0651	TMPC10
0.0894 ± 0.0205	fragCpx
0.0455 ± 0.0038	TIC0
0.0295 ± 0.0059	ZMIC1
0.0270 ± 0.0065	ATS0p
0.0264 ± 0.0060	AATSC1v
0.0202 ± 0.0104	NHOHCount
0.0114 ± 0.0048	VSA_EState3
0.0102 ± 0.0021	TPSA
0.0092 ± 0.0014	SMR

The most important descriptors identified by both the SHAP values and permutation importance are difficult to interpret as they correspond to abstract structural features of the molecules.

Conclusion

SMILES representations of molecules can be used to generate molecular fingerprints and descriptors, both of which can be used to predict molecular properties with a reasonable accuracy.

Molecular descriptors include more detailed information about the molecule and are more strongly correlated with boiling point. This translates to a more accurate performance on the testing dataset.