Computational Epidemiology – PS3

Anirudh Rao

Summary

- The descriptors for the compounds provided was obtained using the ChemDes web server.
- 1875 descriptors (PaDEL) were obtained after giving the SMILES data as input.
- A regression QSAR model was built using Python based on the descriptors obtained and the pIC50 data.
- After training and validation of the model, the pIC50 for the 'BLINDED' compounds was found and analysed.

BLINDED pIC50 Scores

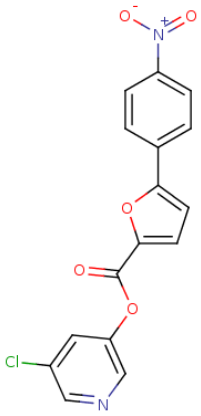
Compound No.	SMILES	Predicted pIC50
6	ClC1=CC(OC(=O)C2=C3C=CC=CC3=CC=C2)=CN=C1	-0.60950964
12	CC1=CC(O)=NC(SCC(=O)NC2=CC=C(OC3=CC=C(Cl)C=C3)C=C2)=N1	-2.183443378
24	CC1=C(C=C(O1)C(C)(C)C)C1=NNC(NS(=O)(=O)C2=CC=CS2)=C1	-1.568560844
25	ClC(Cl)=C(Cl)C(=O)OC1=CC=C(C=C1)S(=O)(=O)C1=CC=C(OC(=O)C(Cl)=C(Cl)Cl)C=C1	-1.283476567
31	O=C1N(CC2=CC=C3C=CC=CC3=C2)C2=CC=C(C=C2C1=O)S(=O)(=O)N1CCCCC1	-0.839339837
37	IC1=CC=C2N(C\C=C\C3=CC4=CC=CC=C4S3)C(=O)C(=O)C2=C1	-0.436155768
43	CC1CCCCN1S(=O)(=O)C1=CC2=C(NC(=O)C2=O)C=C1	-0.359989254
59	[O-][N+](=O)C1=CC=C(C=C1)C1=CC=C(O1)C(=O)OC1=CN=CC(Cl)=C1	0.478992592
70	CC(SC1=NC(C2=CC=CC=C2)=C(C#N)C(=O)N1)C(=O)NC1=CC=C(Cl)C=C1	-2.05000698
80	CC(C)C1=CC=C(NC(=O)CSC2=NC(=CC=N2)C2=CC=CS2)C=C1	-1.601485981

IC50 is the half maximal inhibitory concentration. A lower IC50 is desirable for an inhibitor as it achieves half of the maximum possible inhibition at a lower concentration.

pIC50 is the negative logarithm of IC50. Thus, a higher value of pIC50 implies a greater potency of the inhibitor.

From the predicted pIC50 values, **Compound No. 59** seems to be the most potent inhibitor among the 'BLINDED' compounds.

Using the Smi2DDepict web server, the structure of Compound No. 59 is obtained.



Model Validation

After performing a train-test split on the 'unblinded' compounds,

- the R^2 score for the model was found to be 0.774
- the LOO-q^2 score for the model was found to be 0.485

According to the Golbraikh–Tropsha model acceptability criteria[1],

- The model should have a high LOO-q^2 score (> 0.5).

  Our model has a score of 0.485, which doesn't meet this criterion, but comes very close to it.

- Correlation coefficient R between the predicted and actual activities for an external test set should be close to 1.

  Our model's coefficient is 0.88, which is quite close to 1.

- The slope of regression line should be close to 1.

  For our model, the slope is 0.8, which is again close to 1.

Thus, our model doesn't completely satisfy the Golbraikh–Tropsha criteria, but comes very close to it.

[1] Golbraikh, A., & Tropsha, A. (2002). Beware of q^2! Journal of Molecular Graphics and Modelling, 20(4), 269–276. doi:10.1016/s1093-3263(01)00123-1