Evaluation of Compound Selectivity in PPAR Family using Machine Learning Modelling

Oliver Scott , Dewei Ni, Run Chen Xu, AW Edith Chan

Wolfson Institute for Biomedical Research, University College London, Cruciform Building, Gower St, London WC1E 6BT, UK

- Peroxisome proliferator-activated receptors (PPAR) are members of the nuclear hormone receptors superfamily (NHR)
- PPAR is responsible for regulating many different lipid-related genes, e.g. the metabolism and transport of cholesterol and lipids
- There are three different types of PPAR receptor; PPARα, PPARδ, and PPARγ, with different localizations and specializations
- Some existing agonists are selective. For example, GW501516 is 1000 fold more selective for PPARδ than α or γ

- Cheminformatics is an established fields in drug discovery
- Machine learning has been used successfully in many areas of drug discovery
- Machine learning can be used for pattern recognition and high-level statistical modelling to learn relationships among a large set of chemical compounds
- Large databases of chemical and biological data are readily available, such as ChEMBL
- Open source machine learning algorithms are publicly available

Aims

Predict bioactivities and classify PPAR selectivity using multiple machine learning models

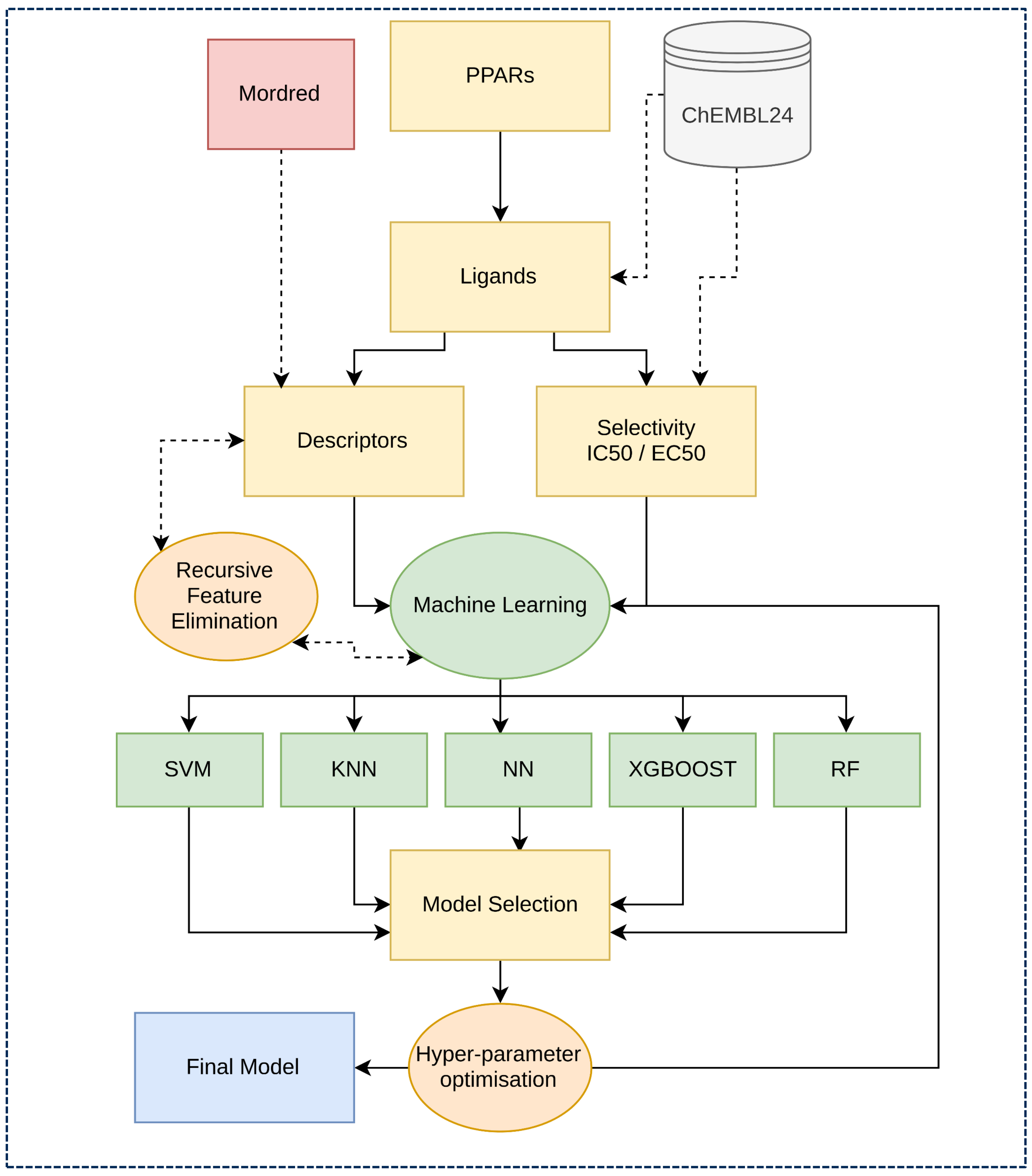
ChEMBL Dataset

# Compounds	PPAR-α	PPAR-γ	PPAR-δ
IC50	911	1573	545
EC50	2459	3081	1333

# Data points	PPAR-α	PPAR-γ	PPAR-δ
IC50	1225	2031	813
EC50	3378	4188	1857

- Chemical structure (SMILES), IC50, and EC50 data were extracted using amino acid sequences of the 3 subtypes (BLASTp)
- PPARα – Q07869, PPARδ – Q03181, PPARγ – P37231

Methods & Workflow



Mordred

PPARs

ChEMBL24

Ligands

Descriptors

Selectivity IC50 / EC50

Recursive Feature Elimination

Machine Learning

SVM

KNN

NN

XGBOOST

RF

Model Selection

Final Model

Hyper-parameter optimisation

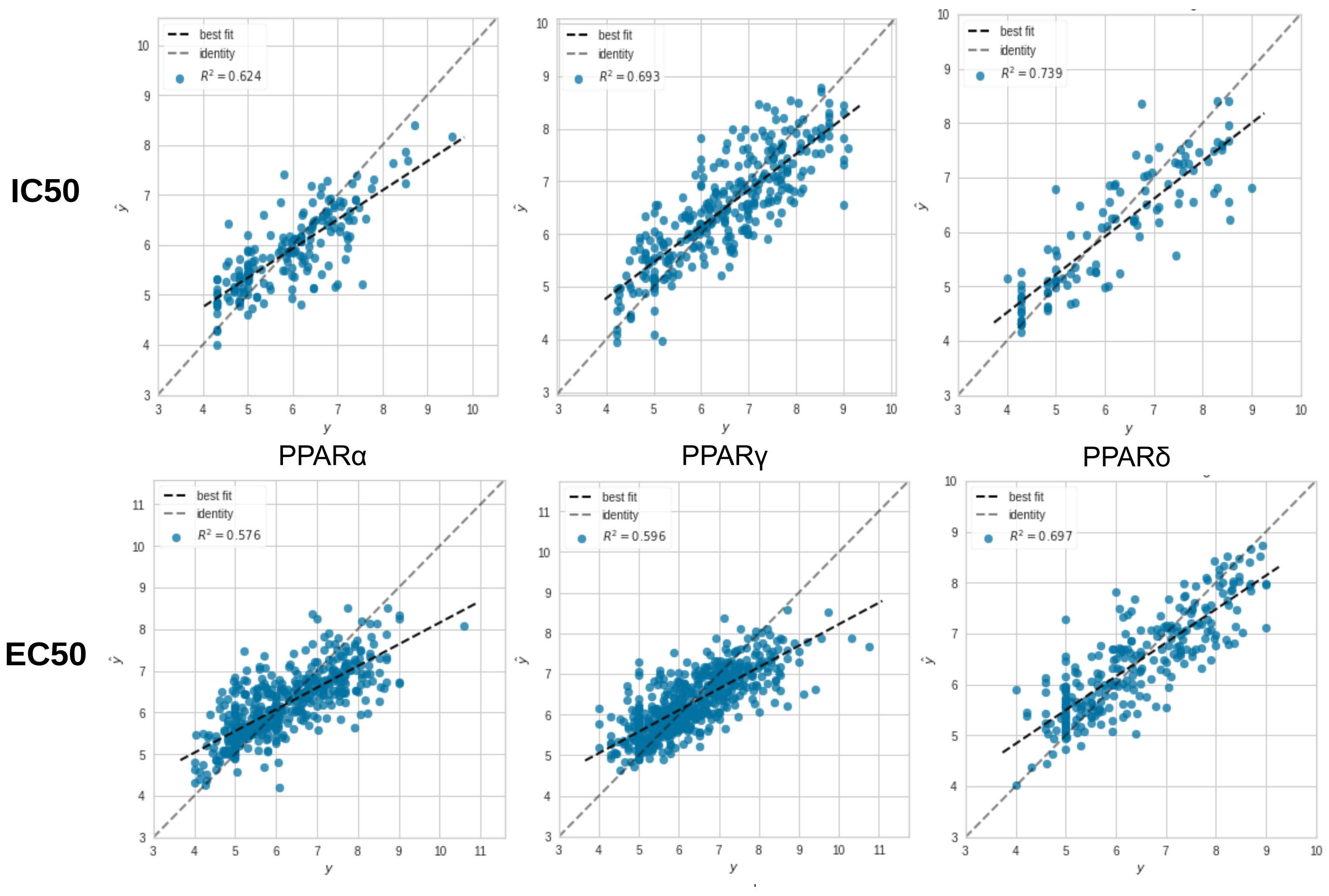
- Models implemented using Python 3.6, sckit-learn, keras and XGBoost
- 1826 initial molecular descriptors calculated with Mordred (1613 2D, 213 3D)
- ML Algorithms:

Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), Extreme Gradient Boosting (XGBOOST)

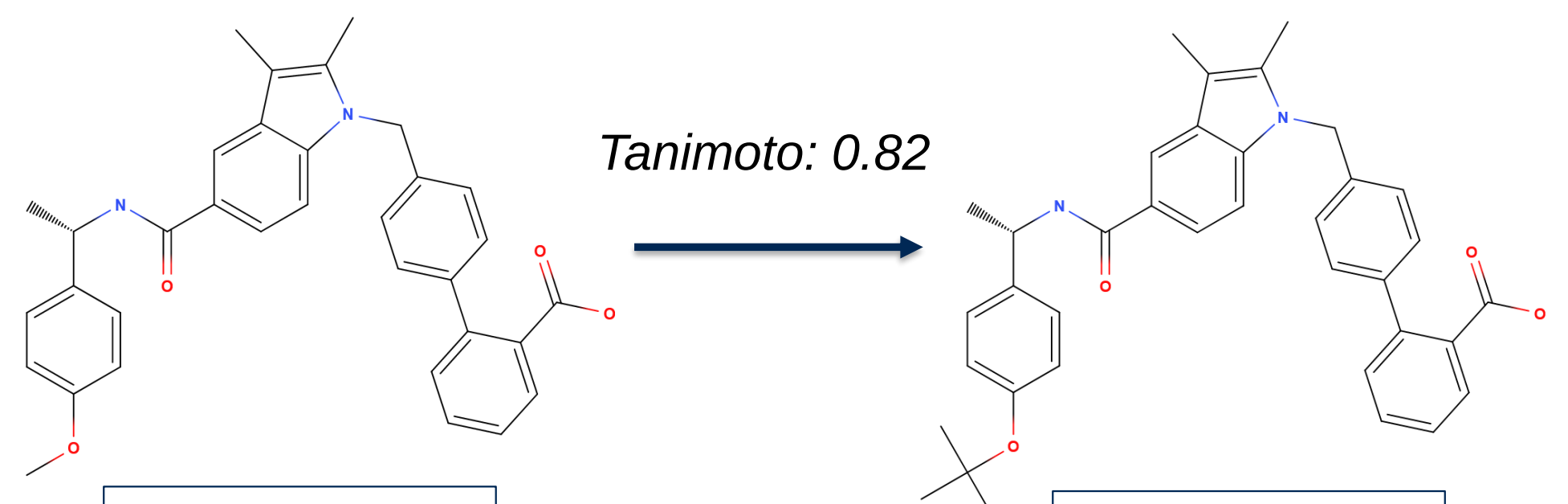
- Feature selection using recursive feature elimination
- Hyper-parameter optimisation with Bayesian optimisation

Results and Discussion

Subtype	Dataset	r^2	MSE	MLSE	MAE
PPAR-α	IC50	0.624	0.432	0.0091	0.505
PPAR-γ	IC50	0.693	0.462	0.0083	0.542
PPAR-δ	IC50	0.739	0.519	0.0092	0.534
PPAR-α	EC50	0.576	0.588	0.0109	0.586
PPAR-γ	EC50	0.596	0.502	0.0091	0.529
PPAR-δ	EC50	0.697	0.456	0.0086	0.524



- R^2 for the training set is close to 1, while it is around 0.6 to 0.7 for the test set, for both IC50, EC50 and each subtype
- High performance on the training set indicates a degree of overfitting
- Models trained with IC50 data show smaller error than EC50, this could due to intrinsic variability within the assay type
- Our results are similar to previous QSAR results. However, the applicability domain of previous QSAR models is small due to the small amount of training data

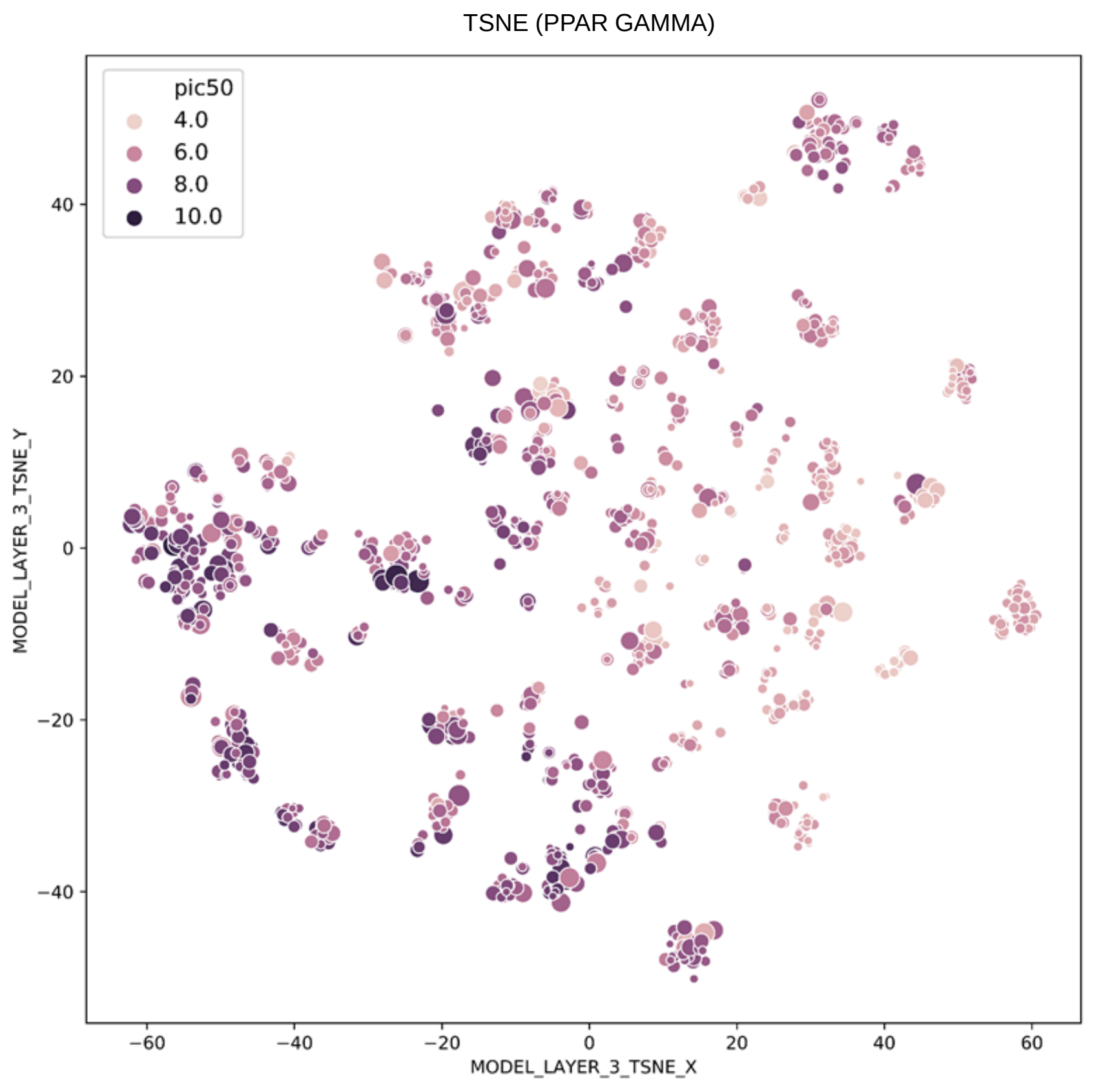


Tanimoto: 0.82

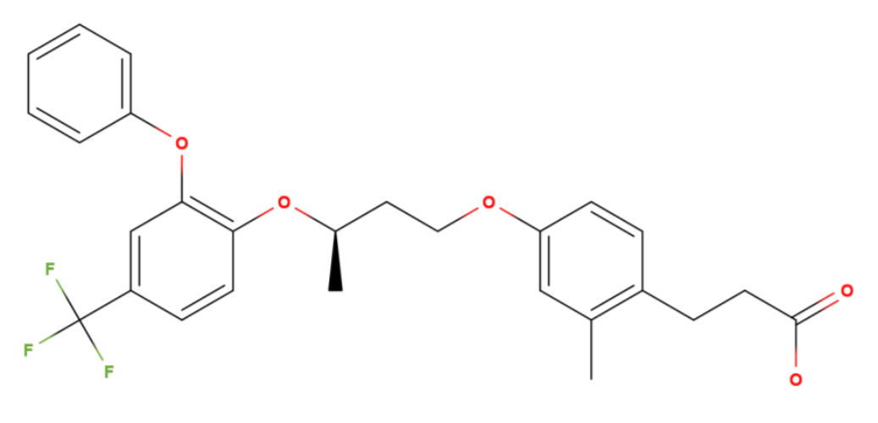
Actual: 6.4
Predicted: 7.2

Actual: 9.1
Predicted: 7.5

- Our models struggle to model activity cliffs where a small change in structure leads to a large change in activity, especially when the structural feature is not in the training set
- TSNE representation shows the model learns that similar structures share similar bioactivity

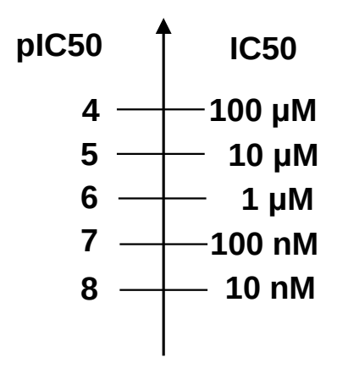


Web output example



CHEMBL213315

pIC50	PPAR-α	PPAR-γ	PPAR-δ
Actual value	5.98	8.22	8.30
Predictive value	5.89	7.61	8.10
Selectivity	0.7	0.9	1



Conclusion

- Our models predict the pIC50 for the 3 subtypes within half a log unit.
- The models provide predicted pIC50 or pEC50 values and compare PPAR selectivity
- The results could help in filtering or screening potential compounds in future studies
- The models will be available from a web interface

References:

oliver.scott.17@ucl.ac.uk

1. Dunning KR, Anastasi MR, Zhang VJ, Russell DL, Robker RL. Regulation of Fatty Acid Oxidation in Mouse Cumulus-Oocyte Complexes during Maturation and Modulation by PPAR Agonists. PLOS ONE 2014;9:e87327.
2. Moller DE. The Mechanisms of Action of PPARs. Annual Review of Medicine 2002;53:409-35.
3. Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics 2018;10:4.
4. RDKit: Open-source cheminformatics; http://www.rdkit.org
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.