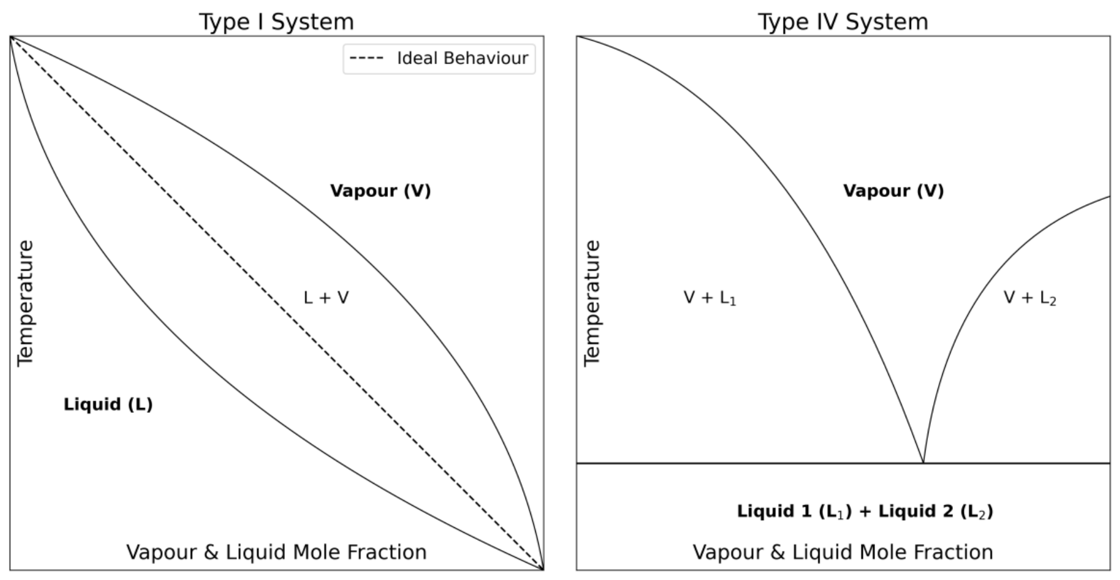
Modelling Miscibility Using Machine Learning

Joshua Cheung[1], Daniel Cançado[2], Thasmia Fathima[3], Joanna Grundy[2], Samantha Kanza[1] and Jeremy Frey[1]
University of Southampton [1] Chemistry and Chemical Engineering, [2] Electronics and Computer Science, [3] Engineering

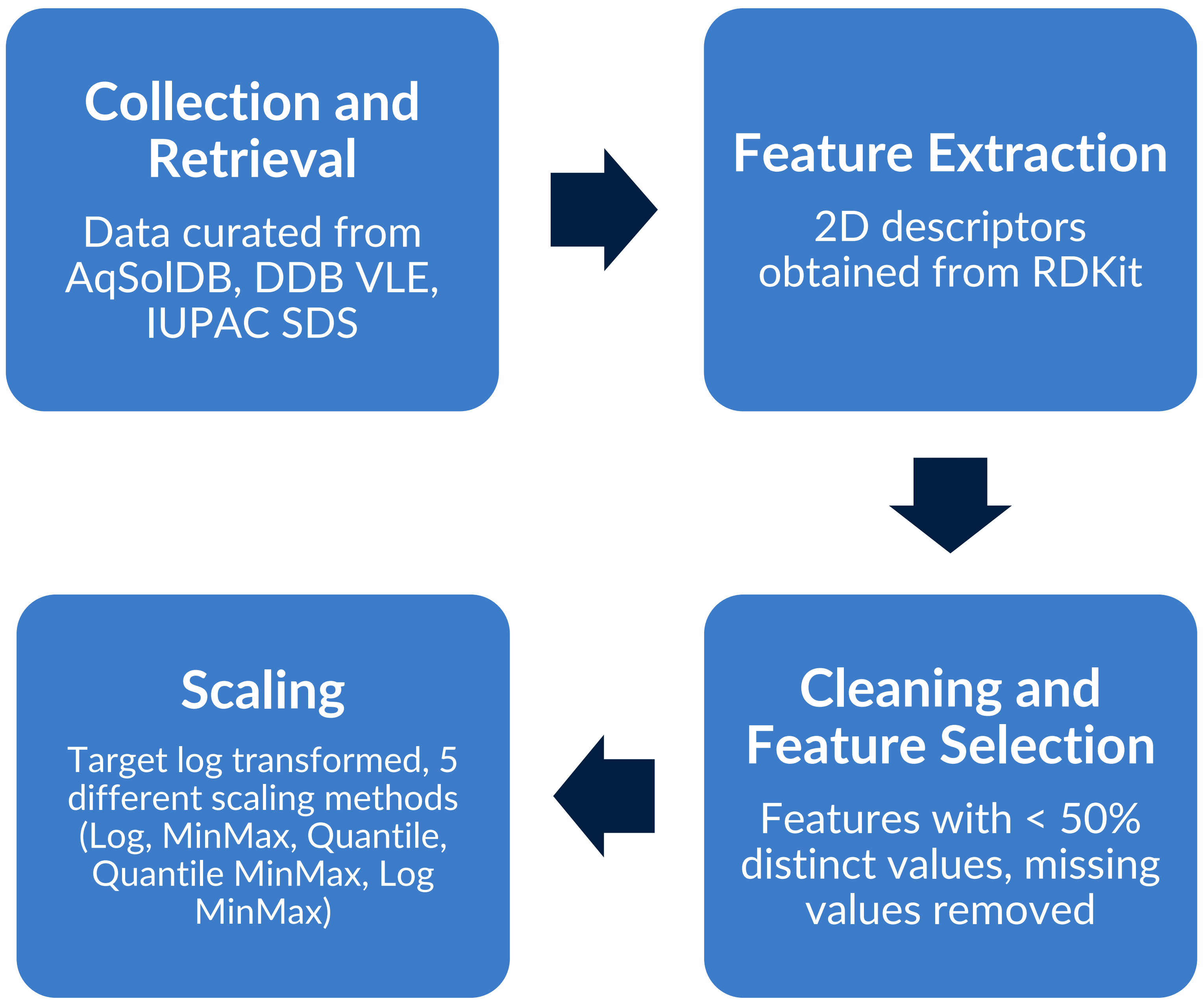
Motivation

- Miscibility is an important component of solubility, and has many applications, e.g. drug discovery, flow chemistry, polymer blends, and more
- Traditionally modelled with quantum mechanics simulations, QSPR
- Machine learning was used to model small molecule miscibility for a binary organic mixture
- Huge wealth of data available, but not necessarily in an accessible format



Methodology

Data Curation and Processing



Collection and Retrieval
Data curated from AqSolDB, DDB VLE, IUPAC SDS

Feature Extraction
2D descriptors obtained from RDKit

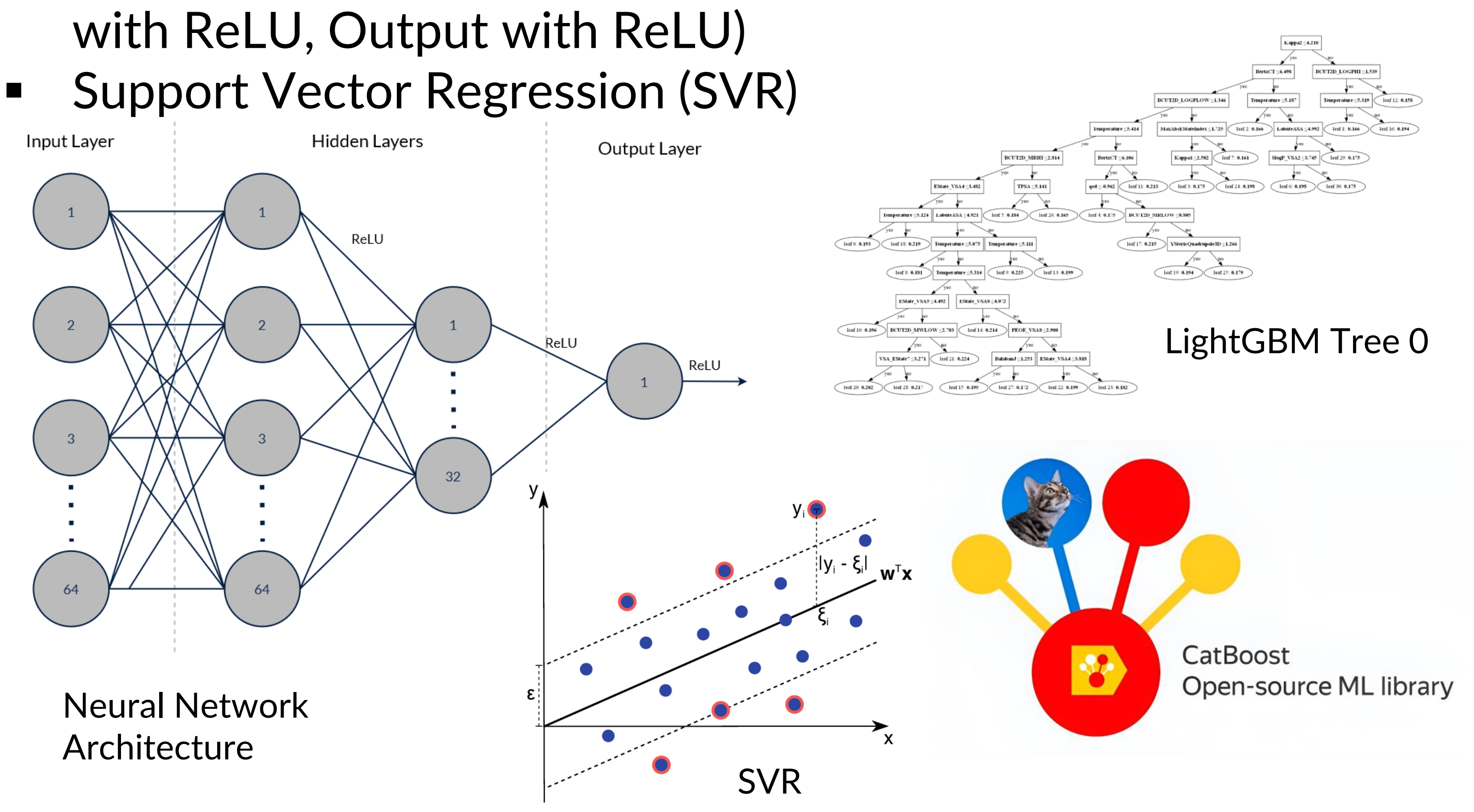
Cleaning and Feature Selection
Features with < 50% distinct values, missing values removed

Scaling
Target log transformed, 5 different scaling methods (Log, MinMax, Quantile, Quantile MinMax, Log MinMax)

Machine Learning Models

Models tried:

- RandomForest
- Gradient Boosted Decision Trees (GOSS, DART, trad variants)
- CatBoost
- Neural Network (Input, 64 neurons with ReLU, 32 neurons with ReLU, Output with ReLU)
- Support Vector Regression (SVR)



Neural Network Architecture

SVR

LightGBM Tree 0

Results

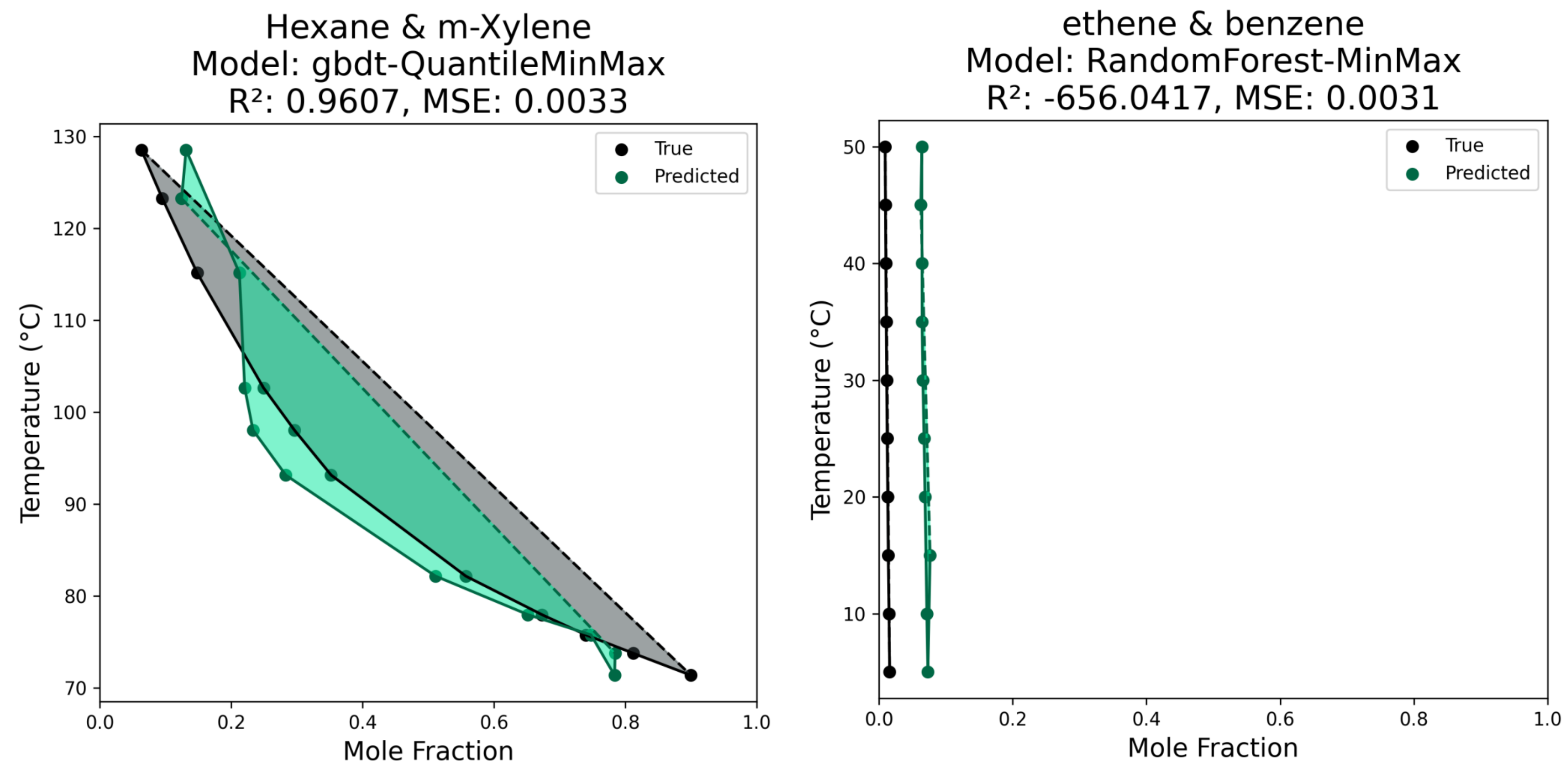
- Overall, models using log transformed mole fractions performed much better than those with unscaled mole fractions (Max R^2 0.93 vs 0.62, lowest MSE 2.80 vs 0.03)
- Scaled mole fractions improve the overall results, but have no significant effect on results by compound pair (solvent and solute)
- When comparing scores by compound pair (> 5 data points), some models appeared to vastly outperform others

Algorithms	Count		Total
	Log Scaled Results	Unscaled Results	
LightGBM (GOSS)	16	20	36
LightGBM (Bagging)	16	14	30
RandomForest	9	15	24
Neural Network	10	8	18
LightGBM (DART)	6	5	11

- The model appeared to predict best for alkanes and carbonyls (including carboxylic acids in the carbonyl category)
- Much higher proportion of those functional groups in the best predicted compound pairs (> 5 data points) than in the dataset

Functional Group	% in top 20 compound pairs	% Train	% Test
Alkanes (unbranched, n > 3)	42.1	10	13.1
Carbonyls (inc acids)	36.8	23.6	25.6
Benzene Rings	29	23.6	26.2
Carbonyls (excl acids)	7.89	17.2	17.2
Furan Rings	2.63	0.63	0.84

The percentage of compounds in each subset (top 20 highest scoring compound pairs, training set, test set) that contained a given functional group



2 of the predicted phase diagrams: Hexane and m-Xylene displaying high precision and accuracy, Ethene and Benzene illustrating high precision but low accuracy, both for type I systems

Conclusions and Future Work

- It is possible to predict and plot liquid phase temperature-composition diagrams for a binary mixture of organic compounds using machine learning
- Data scarcity and quality proved challenging, with issues such as unit conversion and data not being in an accessible format (16 968 data points in total, only 510 compound pairs with more than 4 data points); ideally, there would be more compound pairs with > 4 data points
- Improving the size and quality of the dataset could lead to better model performance
- Immiscible compounds (IV systems) were largely underrepresented in the dataset. Oversampling them could improve predictions.
- Further expansion: building a secondary database of predicted phase diagrams, which can then be used to train a classifier.

Walas, S. M. (1985) ‘5 - Phase Diagrams’, in Walas, S. M. (ed.) *Phase Equilibria in Chemical Engineering*. Butterworth-Heinemann, pp. 245–297. doi: 10.1016/B978-0-409-95162-2.50013-0.
Ke, G. et al., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, pp.3146–3154.
Breiman, L. (2001) ‘Random Forests’, *Machine Learning*, 45, pp. 5–32. doi: 10.1023/A:1010933404324.
Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), pp.2825–2830.
Rosenbaum, Lars & Dörr, Alexander & Bauer, Matthias & Boeckler, Frank & Zell, Andreas. (2013). Inferring multi-target QSAR models with taxonomy-based multi-task learning. Journal of cheminformatics. 5. 33. 10.1186/1758-2946-5-33.
Sorkun, M. C., Khetan, A. and Er, S. (2019) ‘AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds’, *Scientific Data*. doi: 10.1038/s41597-019-0151-1.
Iupac (2023) *Solubility Data Series*. Available at: https://iupac.org/what-we-do/databases/solubility-data-series/.
‘RDKit: Open-Source Cheminformatics Software’ (2023). Available at: https://www.rdkit.org/.

This work was funded by EPSRC through grants EP/X032701/1, EP/X032663/1 and EP/W032252/1 – PSDI (Physical Sciences Data Infrastructure (PSDI) https://www.psdi.ac.uk/.