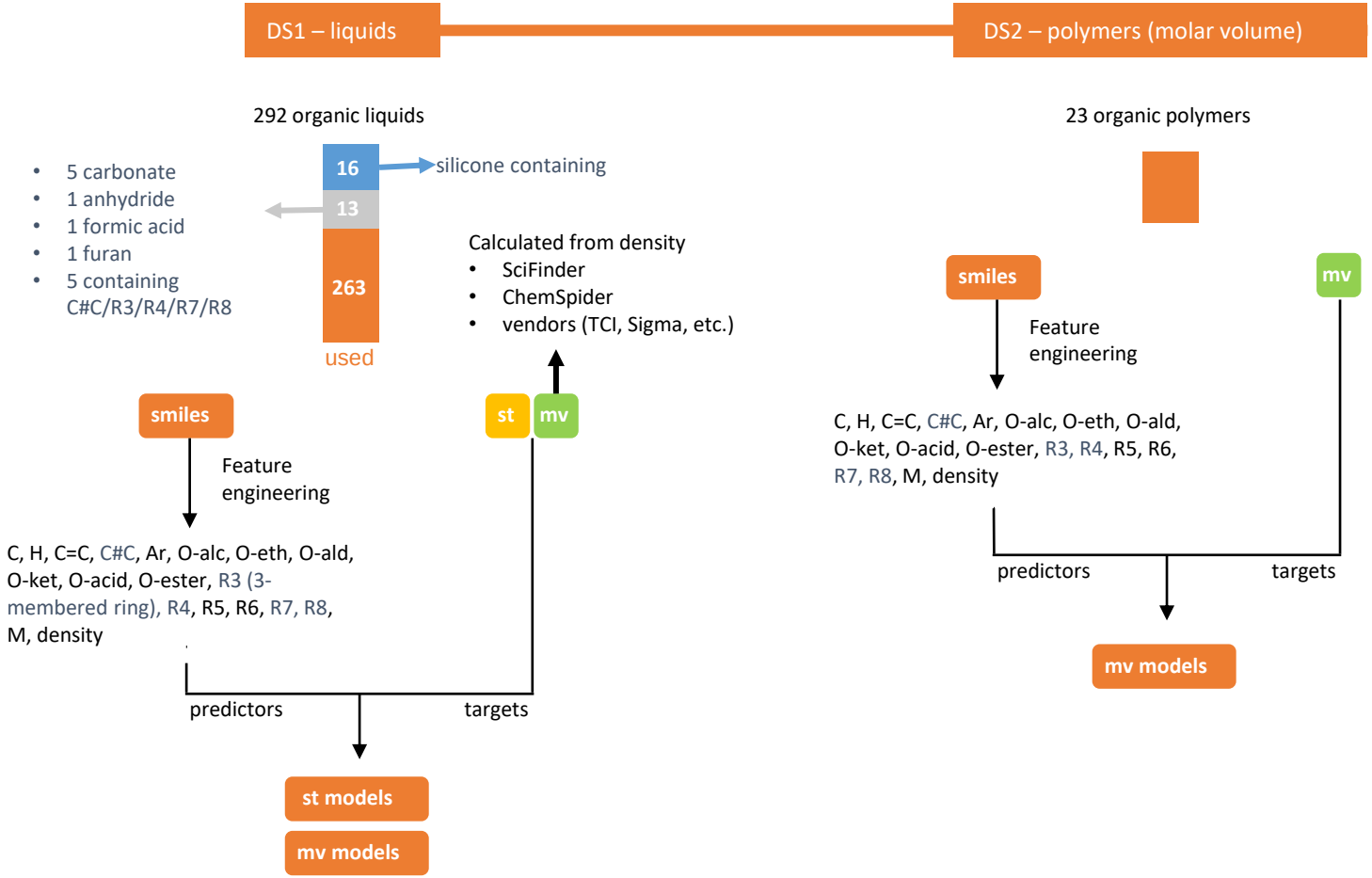
Surface tension modelling

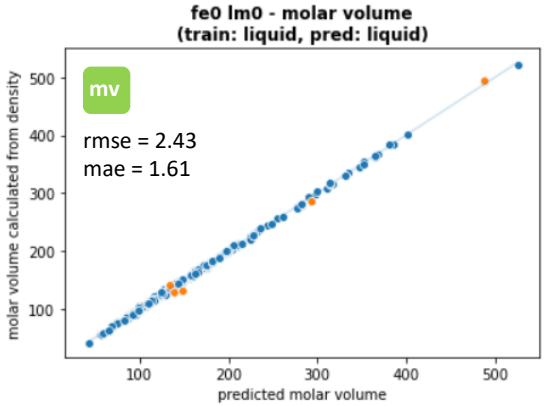


mv and density are the same information. One can be calculated from the other.

Surface tension modelling – linear model (DS1 – liquids)

Regressors: C, H, C=C, (C#C), Ar, O-alc, O-eth, O-ald, O-ket, O-acid, O-ester, (R3), (R4), R5, R6, (R7), (R8)

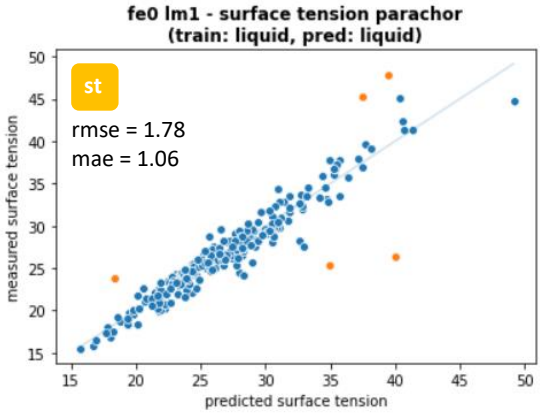
Features in bracket are excluded



$$mv = \theta_1 X_1 + \theta_2 X_2 + \ldots$$

measured_st	molecule	density	mv
26.4	Hexyl formate	0.990	131.502020
25.3	5-Methyl-2-hexanone	0.888	128.590090
24.2	2,2,4,4,6,8,8-Heptamethylnonane	0.793	285.558638
28.9	Triethylhexanoin	0.950	495.464211
23.8	Propyl acetate	0.836	122.168660

- Good fit
- Mistakes in SciFinder's densities:
  - Hexyl formate → 0.879
  - 5-Methyl-2-hexanone → 0.814
  - Propyl acetate → 0.888
- # TODO: correct this

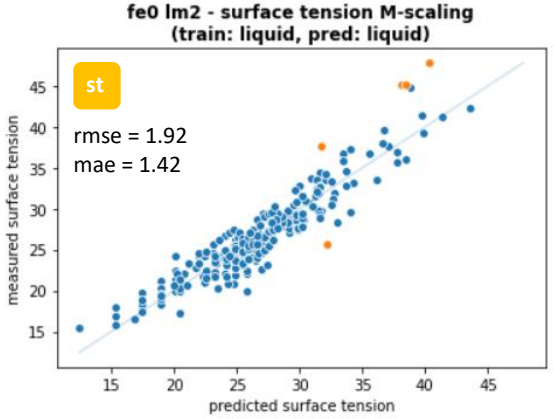


Parachor model $\gamma^{1/4}.mv = \theta_1 X_1 + \theta_2 X_2 + \ldots$

X_i = fragment counts

measured_st	molecule	density	mv
26.4	Hexyl formate	0.9900	131.502020
25.3	5-Methyl-2-hexanone	0.8880	128.590090
47.8	1,2-Ethanediol	1.1135	55.741356
45.2	Trimethylene glycol	1.0529	72.271821
23.8	Propyl acetate	0.8360	122.168660

- Good fit, but prone to error (due to power term)
- 3/5 outliers match mv model's outliers, the other 2 are **diols**



$$\gamma = \theta_1 X_1 + \theta_2 X_2 + \ldots$$

X_i = fragment counts / M

measured_st	molecule	density	mv
47.8	1,2-Ethanediol	1.1135	55.741356
45.1	Triethylene glycol	1.1250	133.488000
45.2	Trimethylene glycol	1.0529	72.271821
25.6	Paraldehyde	0.9923	133.184521
37.7	Dimethyl maleate	1.1570	124.568712

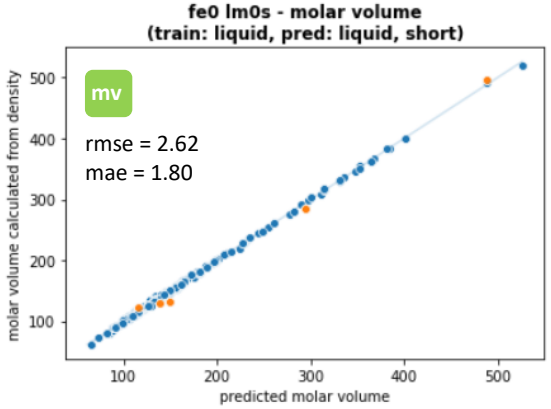
- Direct relationship between fragments and st has a degree of linearity
- Simple, but possibility of missing features. E.g. at $15 < x < 20$.

Surface tension modelling – linear model (DS1 – liquids) - short

Regressors: C, H, C=C, (C#C), Ar, (O-alc), O-eth, (O-ald), O-ket, (O-acid), O-ester, (R3), (R4), R5, R6, (R7), (R8)

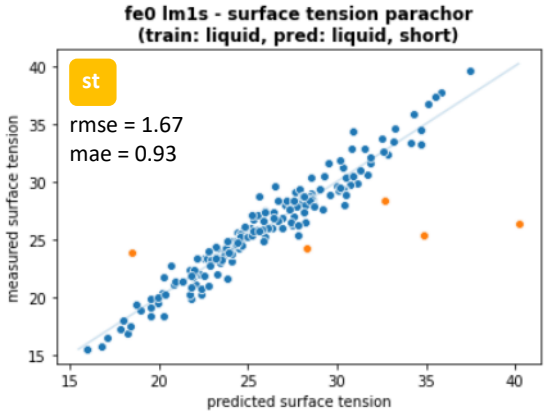
Features in bracket are excluded

Short = excluding alcohols, acids, aldehydes



$$mv = \theta_1 X_1 + \theta_2 X_2 + \ldots$$

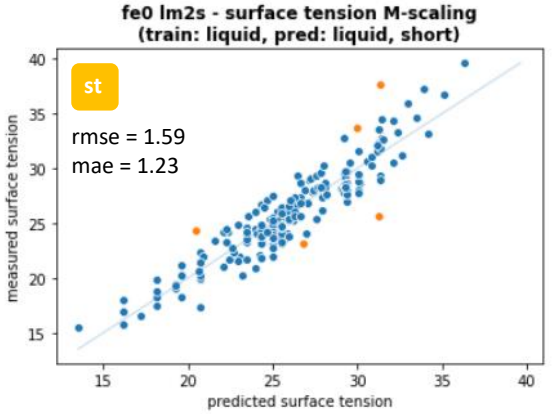
measured_st	molecule	density	mv
26.4	Hexyl formate	0.990	131.502020
25.3	5-Methyl-2-hexanone	0.888	128.590090
24.2	2,2,4,4,6,8,8-Heptamethylnonane	0.793	285.558638
28.9	Triethylhexanoin	0.950	495.464211
23.8	Propyl acetate	0.836	122.168660



$$\gamma^{1/4}.mv = \theta_1 X_1 + \theta_2 X_2 + \ldots$$

X_i = fragment counts

measured_st	molecule	density	mv
26.4	Hexyl formate	0.9900	131.502020
25.3	5-Methyl-2-hexanone	0.8880	128.590090
23.8	Propyl acetate	0.8360	122.168660
28.3	1,2,3-Trimethylbenzene	0.8944	134.386181
24.2	2,2,4,4,6,8,8-Heptamethylnonane	0.7930	285.558638

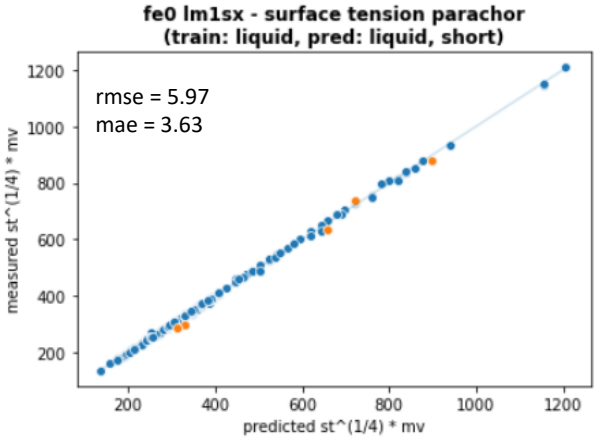


$$\gamma = \theta_1 X_1 + \theta_2 X_2 + \ldots$$

X_i = fragment counts / M

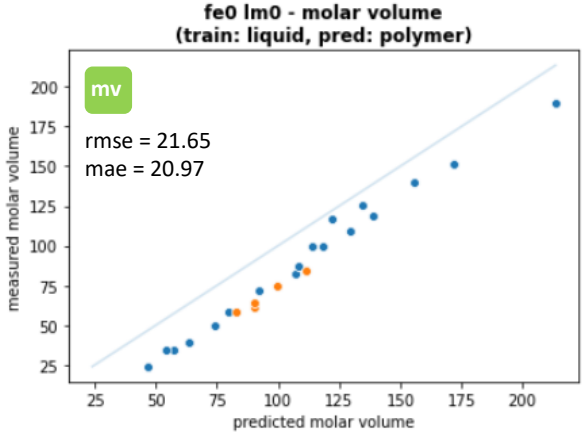
measured_st	molecule	density	mv
37.7	Dimethyl maleate	1.1570	124.568712
25.6	Paraldehyde	0.9923	133.184521
24.3	Methyl formate	0.9742	61.642373
33.7	Tetraethylene glycol dimethyl ether	1.0100	220.080198
23.1	Ethyl orthoformate	0.8909	166.350881

Surface tension modelling – linear model (DS1 – liquids, DS2 - Polymers)



measured_st	molecule	density	mv
26.4	Hexyl formate	0.990	131.502020
24.2	2,2,4,4,6,8,8-Heptamethylnonane	0.793	285.558638
25.3	5-Methyl-2-hexanone	0.888	128.590090
29.6	Methyl palmitate	0.852	317.437793
28.0	Neopentyl glycol di(2-ethylhexanoate)	0.931	382.972073

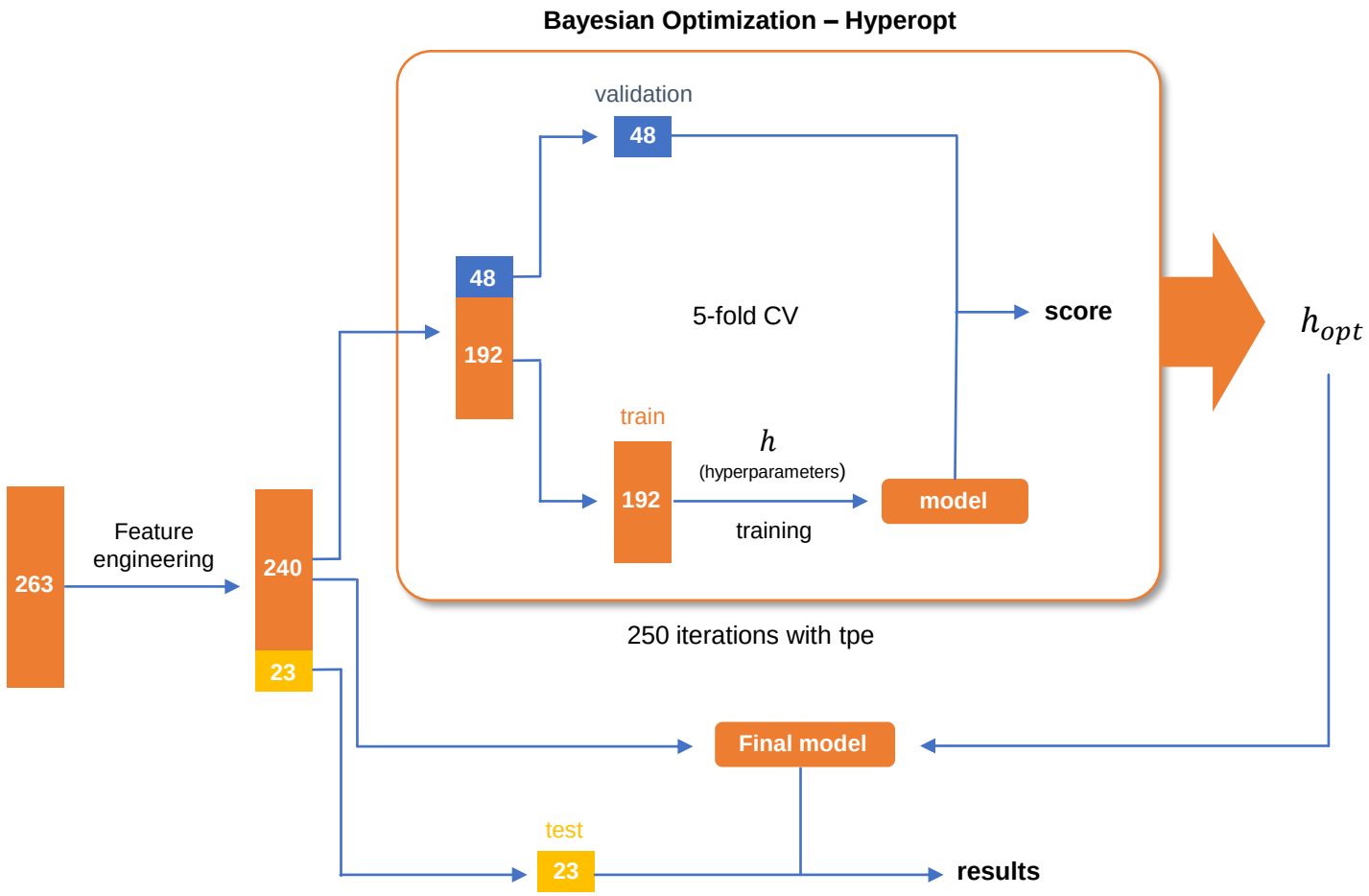
Before transformation, linearity holds. This is hardly special since the surface tension reduced to the power of 1/4th is highly swamped by mv which is proven to be a linear combination of fragments



polymer	measured_mv	ru
polyisobutylene	61.4	['isobutylene', nan]
poly(methyl methacrylate)	84.1	['methyl methacrylate', nan]
poly(1-butene)	64.4	['1-butene', nan]
polybutadiene(trans)	58.1	['butadiene', nan]
polyisoprene(iso)	74.7	['isoprene', nan]

- Predicted mv of polymers are consistently below measured value. This is easily explained by the increase in density as degree of polymerization increases
- Dataset is very limited due to absence of molecular weight. The overestimation of molar volume interestingly seems to be constant.
- Without correction of molar volume, it is impossible to use liquids dataset to predict st of polymers. Even then, parachor model is very sensitive to error in molar volume.
- When density can be accurately measured, there is a good chance polymer st can be predicted with linear model, at least within a class of similar polymers.
- #TODO: build linear model for mv from DS2

Surface tension modelling – XGB (DS1 – liquids)



Workflow

263 datapoints are split into test set (23) and train set (240).

Test data points are picked as suggested by the literature from which DS1 is taken from.

Training loop

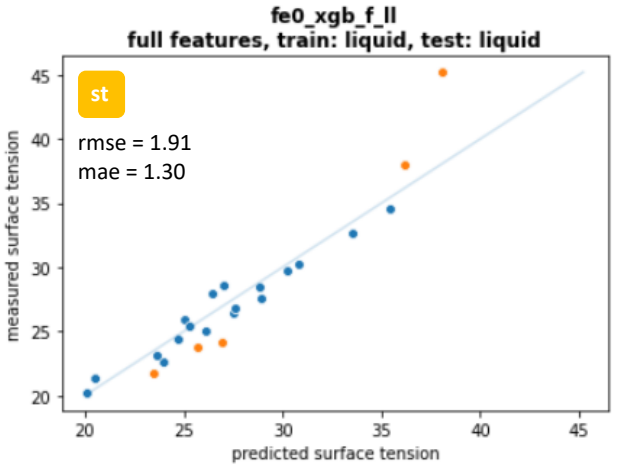
For each iteration (out of 250):

- Train set is split into 5 (fixed seed)
- For each split (out of 5):
  - 1 split → validation set
  - 4 splits → cv train set
  - XGB is trained on cv train set and tested on validation set
- Average score for 5 splits is computed
- BO with tpe suggests new hyperparameters for next iteration

Final model is trained on the full train set (240) with best hyperparameters.

Final model is then tested on test set.

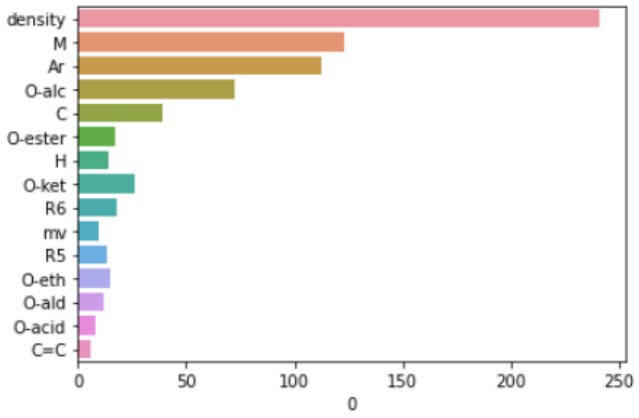
Surface tension modelling – XGB (DS1 – liquids)



Top 5 error

measured_st	molecule	density	mv
45.2	Trimethylene glycol	1.0529	72.271821
24.1	2,6-Dimethyl-4-heptanone	0.8062	176.435128
23.8	3-Methylbutanoic acid propyl ester	0.8620	167.301624
38.0	Benzaldehyde	1.0470	101.360076
21.8	Isopropyl acetate	0.8718	117.151870

Feature importances



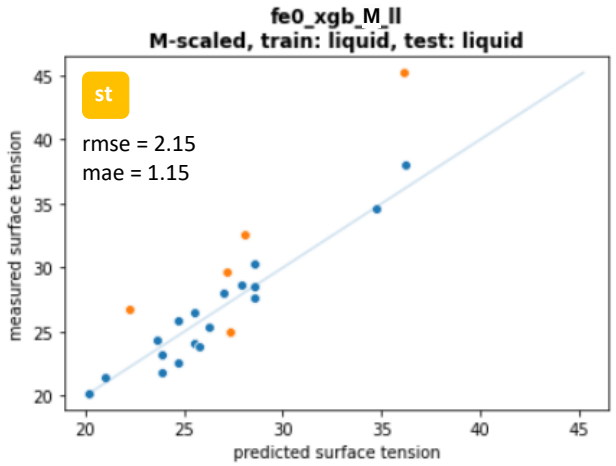
Features

C, H, C=C, (C#C), Ar, O-alc, O-eth, O-ald, O-ket, O-acid, O-ester, (R3), (R4), R5, R6, (R7), (R8), M, density, mv

Features in bracket are excluded

- Decent fit except for trimethylene glycol
- Most important feature is density. Conceptually, density is correlated to surface tension since both seems to arise from intermolecular force.
- # TODO: do all diols have large error in training? Are they underrepresented? The answer seems to be yes, as observed also in the linear model.
- # TODO: train “short” models without irrelevant functional groups in polyester (OH, acid, aldehyde)

Surface tension modelling – XGB (DS1 – liquids)



Top 5 error

measured_st	molecule	density	mv
45.2	Trimethylene glycol	1.0529	72.271821
26.8	Diethylene glycol diethyl ether	0.9070	178.863286
32.6	Methyl acetoacetate	1.0762	107.894443
29.7	Diethylene glycol monobutyl ether	0.9536	170.122693
25.0	Isovaleric acid	0.9310	109.702470

- Most simple, does not consider density or mv, but worst fit.

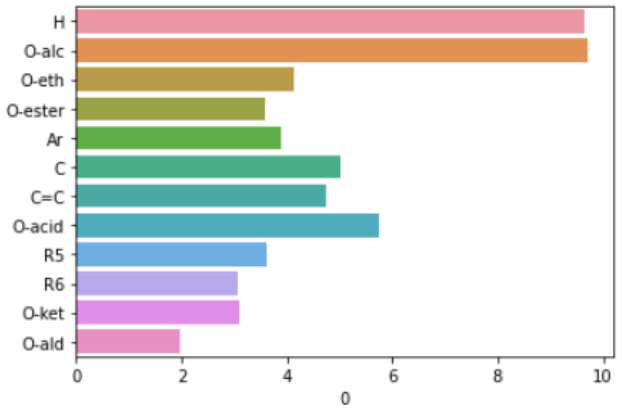
Features

C, H, C=C, (C#C), Ar, O-alc, O-eth, O-ald, O-ket, O-acid, O-ester, (R3), (R4), R5, R6, (R7), (R8), (M), (density), (mv)

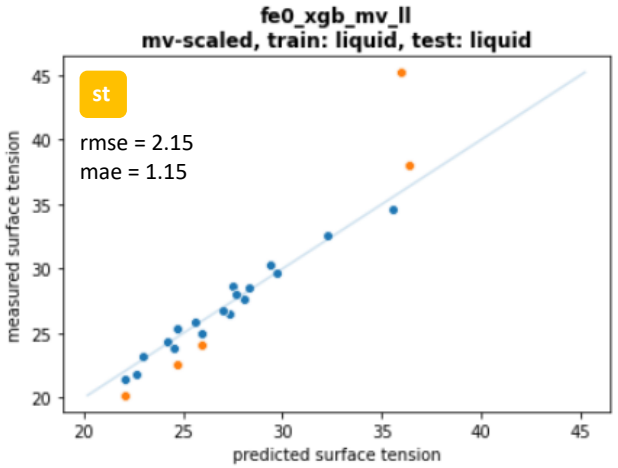
Features in bracket are excluded

Scaled by $\frac{1}{M}$

Feature importances



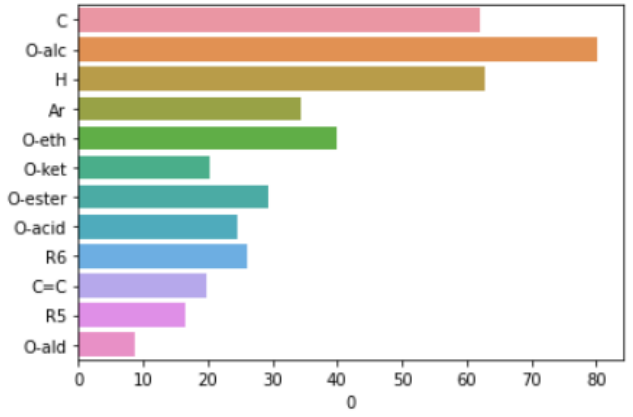
Surface tension modelling – XGB (DS1 – liquids)



Top 5 error

measured_st	molecule	density	mv
45.2	Trimethylene glycol	1.0529	72.271821
22.6	4-Methyl-2-pentanol	0.8075	126.534985
20.2	2,2,3-Trimethylpentane	0.7160	159.541899
24.1	2,6-Dimethyl-4-heptanone	0.8062	176.435128
38.0	Benzaldehyde	1.0470	101.360076

Feature importances



Features

C, H, C=C, (C#C), Ar, O-alc, O-eth, O-ald, O-ket, O-acid, O-ester, (R3), (R4), R5, R6, (R7), (R8), (M), (density), (mv)

Features in bracket are excluded

Scaled by $\frac{1}{mv}$

- Decent fit except for trimethylene glycol
- St seems to be a volume phenomenon, not mass (comparing to M-scaled model)
- Feature importances are more evenly spread, suggesting the importance of mv (and hence density)
- Presence of hydroxyl affects surface tension greatly
- C and H are possibly indicators of size, as larger molecules tend to have higher st because of higher density.
- Highly branched structure are potentially giving large error (entry 3&4 in top 5)
- # TODO: plot correlations. Density and st have some collinearity
- # TODO: try different feature engineering, maybe less atomic based more group based
- # TODO: determine the feature space explored in DS1. XGB does not extrapolate well

Important findings

- Density and mv are important data to predict st
- Chemical nature strongly influences st
- Assuming high accuracy of measurements, parachor model seems decent to predict st
- Scaling by 1/mv seems effective to open possibilities of predicting diverse polymers (e.g. multiple repeating units)
- Feature engineering can be refined, e.g. considering branching (CH3, CH2, CH, … instead of C atoms, H atoms)