Summary - 2021/1/25

DHR

Feature importances were estimated using 3 tests:

- *PLS-Regression Analysis (PLSR)* — magnitudes of x-loadings relate to importance of each feature w.r.t. target variable

- *Gradient Boosting Regression (GBR)* — ensemble model where successive individual decision trees are fitted against the residuals, generally less prone to overfitting than random forest, provides direct estimate of feature importances (based on number of splits in each tree containing a given feature)

- *Permutation Test (PER)* — A trained regression model (SVR trained on combined dataset I talked about before) is used to make predictions with each individual feature shuffled. The feature importance is related to the degree to which the MSE increases when a feature is shuffled relative to the non-shuffled dataset

Feature selection was performed using sequential feature removal tests:

1. Start with complete feature set and corresponding feature importances

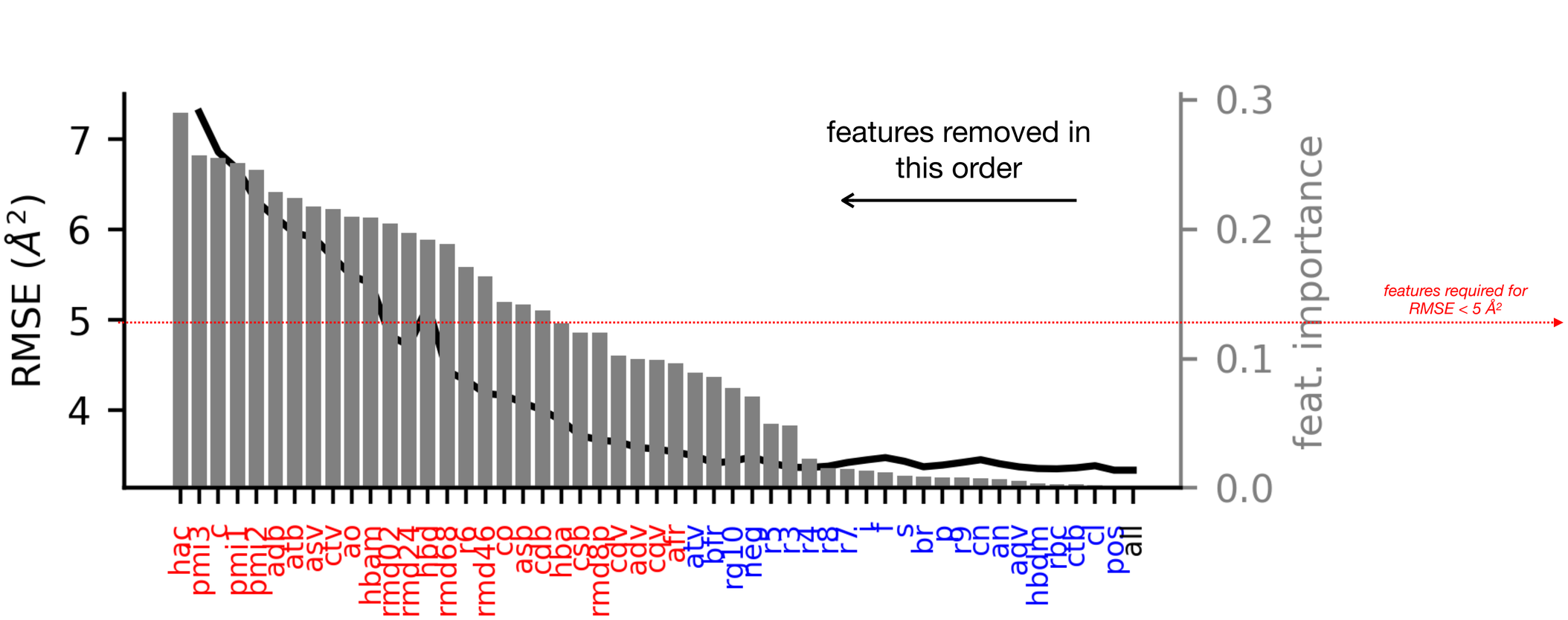
2. record the prediction performance (RMSE)

3. remove the least important feature

4. Repeat steps 2 and 3 until only a single feature remains

5. Repeat entire process using feature importances from all 3 tests (PLSR, GBR, PER)

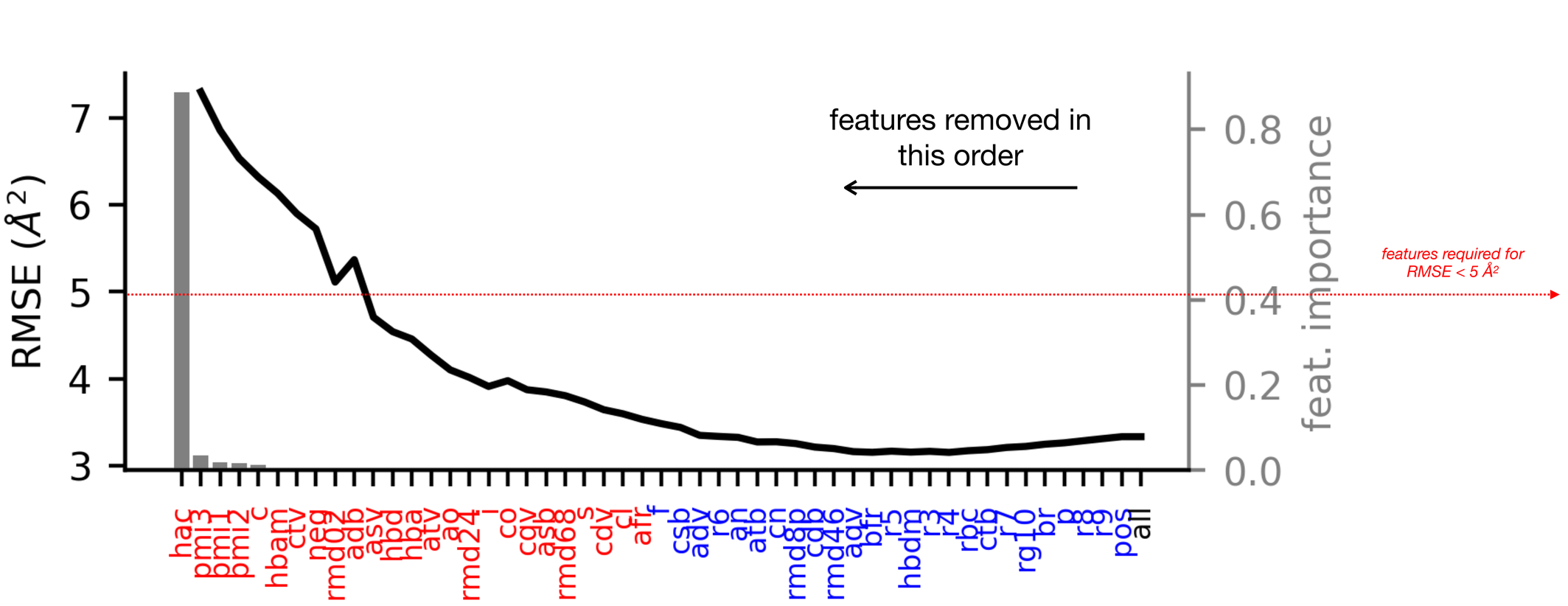
PLSR - sequential feature removal



blue if feature removal increases RMSE by < 5% of the complete feature set, **red** otherwise

feature
hac
pmi3
c
pmi1
pmi2
adb
atb
asv
ctv
ao
hbam
rmd02
rmd24
hbd

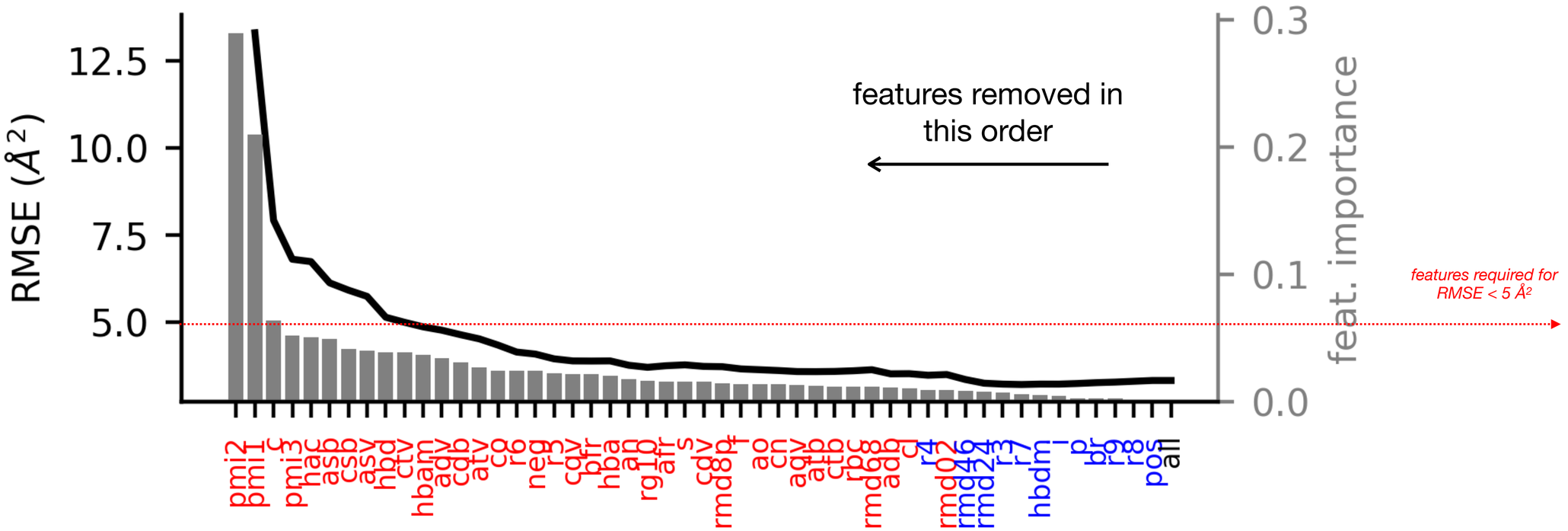
GBR - sequential feature removal



feature
hac
pmi3
pmi1
pmi2
c
hbam
ctv
neg
rmd02
adb
asv

blue if feature removal increases RMSE by < 5% of the complete feature set, **red** otherwise

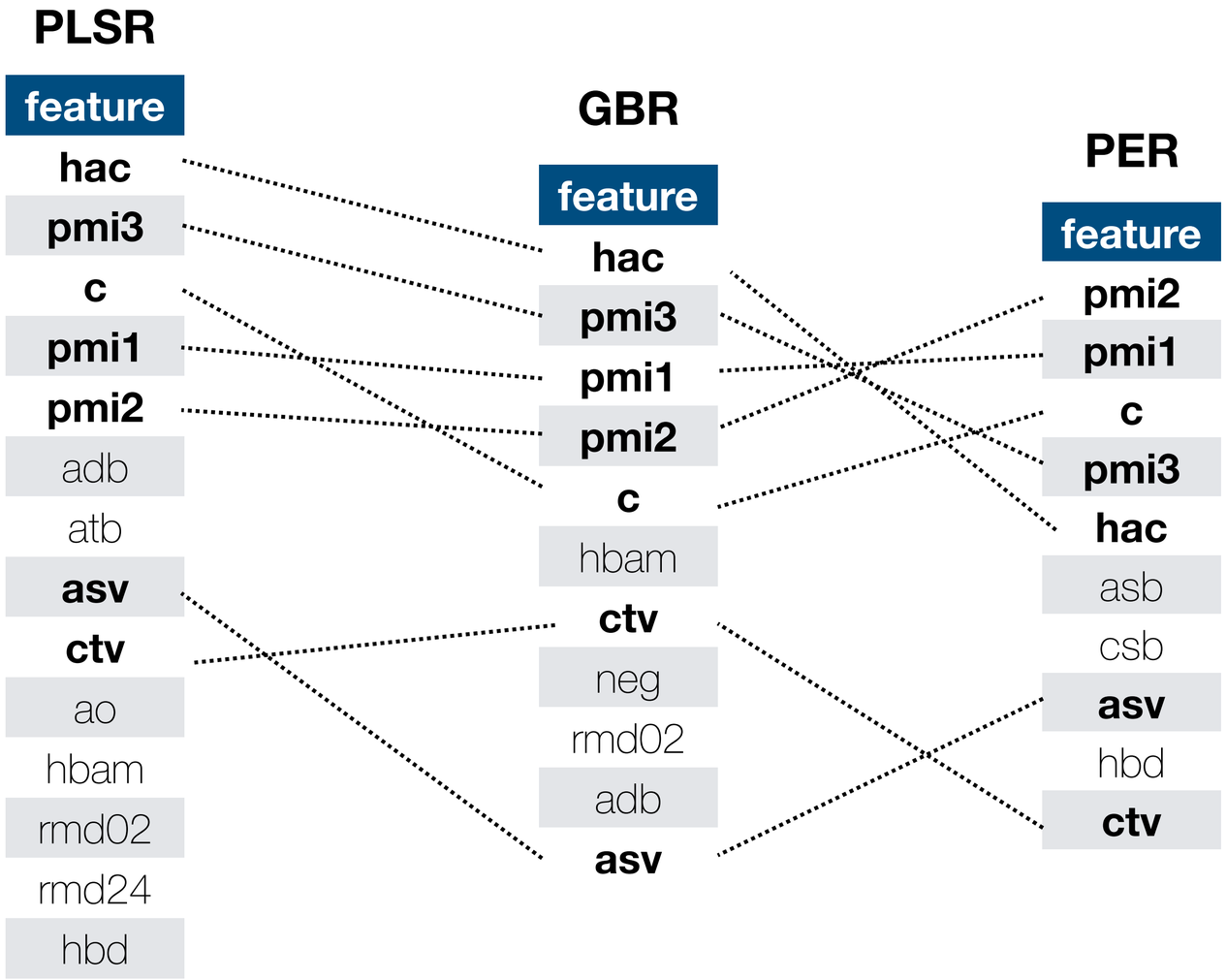
PER - sequential feature removal



feature
pmi2
pmi1
c
pmi3
hac
asb
csb
asv
hbd
ctv

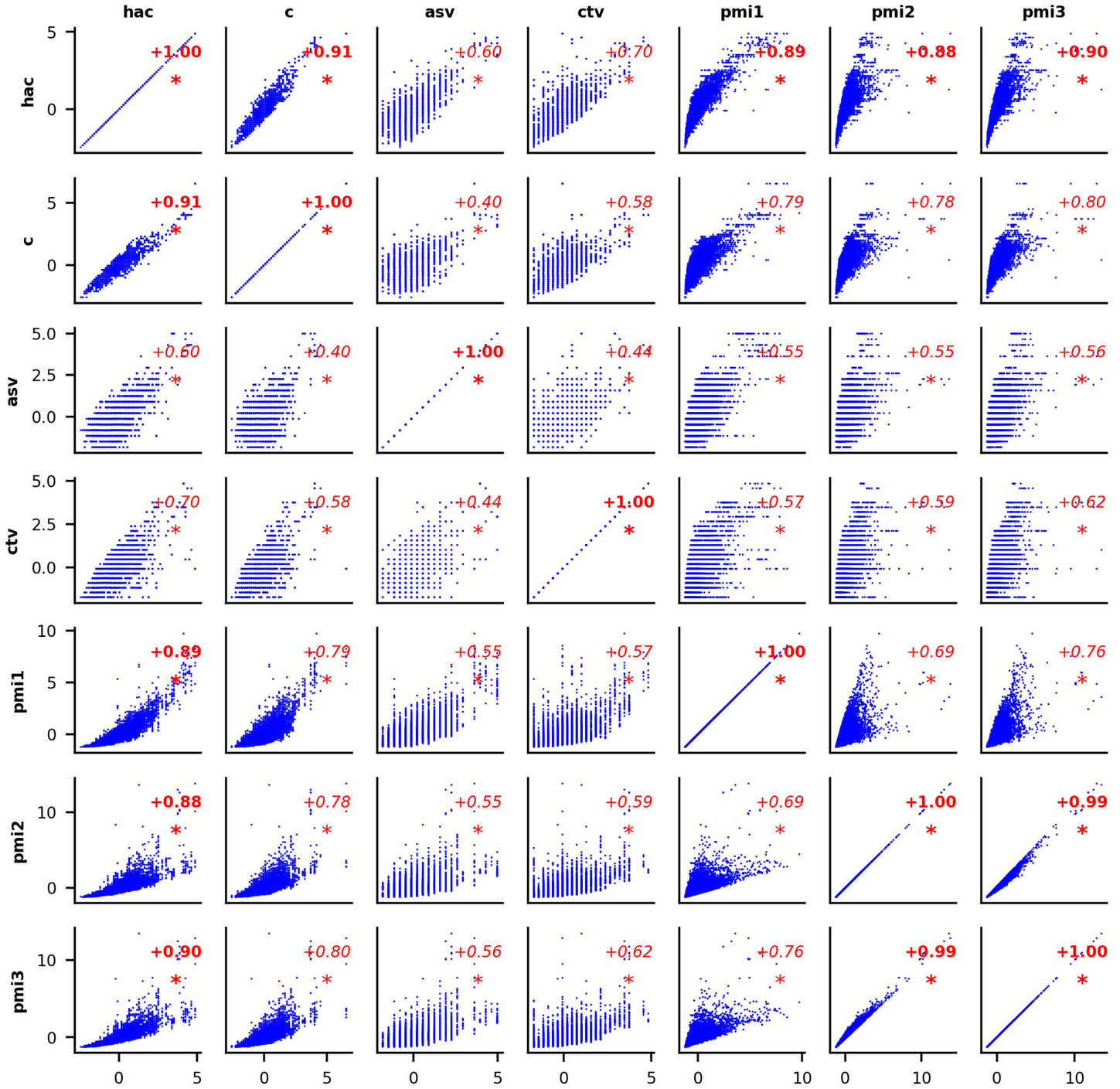
blue if feature removal increases RMSE by < 5% of the complete feature set, **red** otherwise

Minimal feature set selected as common components of all 3 feature removal tests



Reduced feature set enhances performance (relative to MQN or MD3D alone) with less overfitting

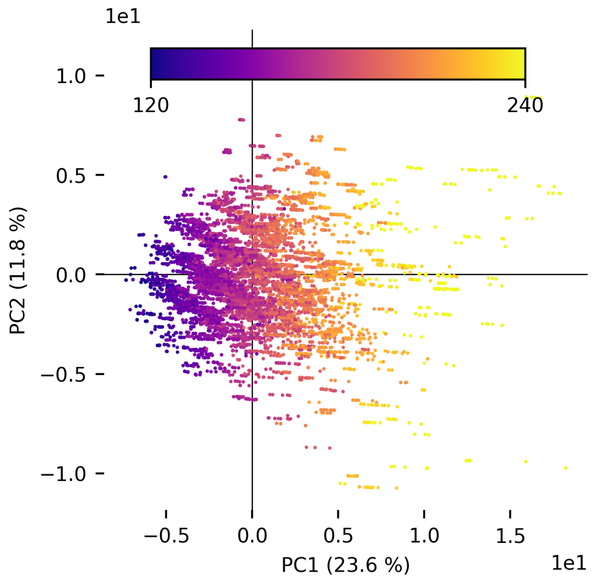
There is still correlation among the features in this reduced feature set

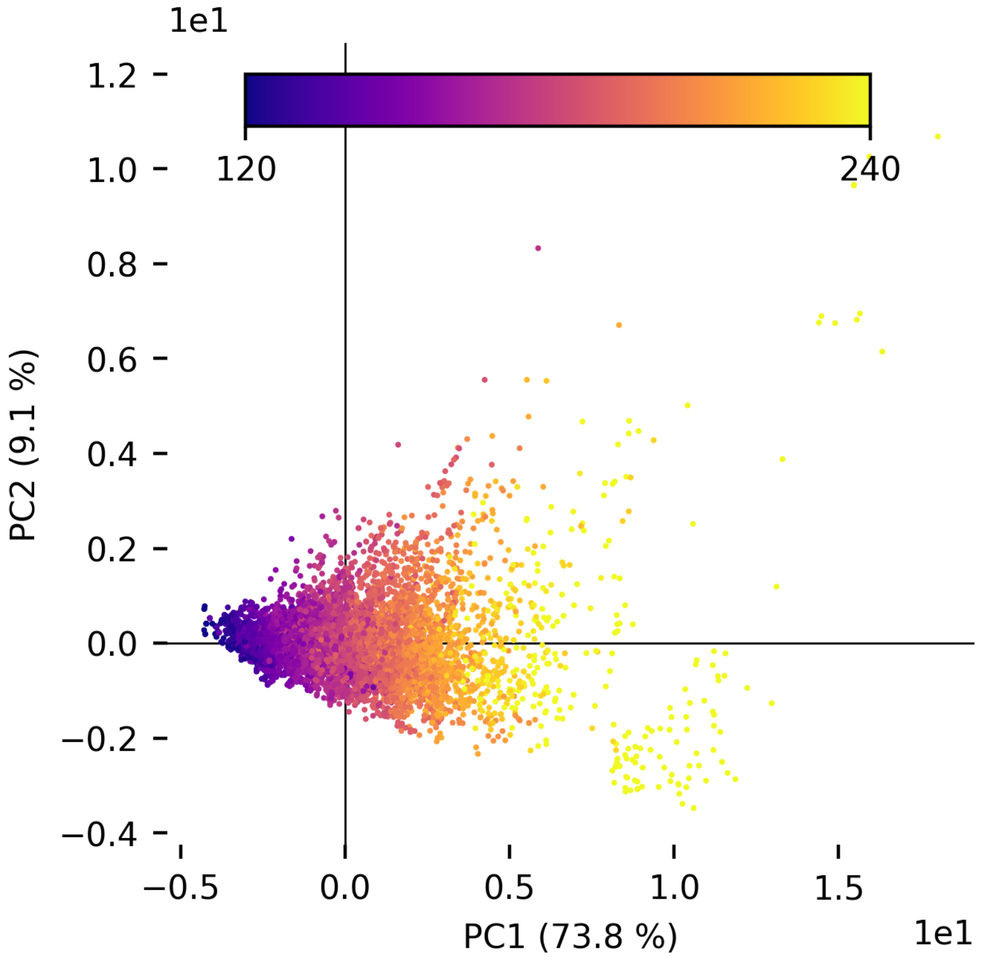


*Spearman rank correlation coefficient (**bold** if > 0.8)*
** if p < 0.01*

PCA - reduced feature set

(complete feature set)





- 4 components required to explain 95% of the variance (compared to 27 components for complete feature set)
- a large proportion of the variance in this reduced feature set is related to CCS

Reduced feature set still has a lot of support vectors in trained predictive model