From simulated to experimental :

PCR Melt Curve Prediction with Random Forest Regressor

Introduction

- Objective: Predict experimental melt curve from simulated melt curve in a given experimental setting

- Experimental melt curve data from

Moniri, Ahmad et al. "High-Level Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis." *Analytical chemistry* vol. 92,20 (2020): 14181-14188. doi:10.1021/acs.analchem.0c03298

- Simulated melt curve obtained with uMelt

Using target sequences and primers provided in the aforementionned paper

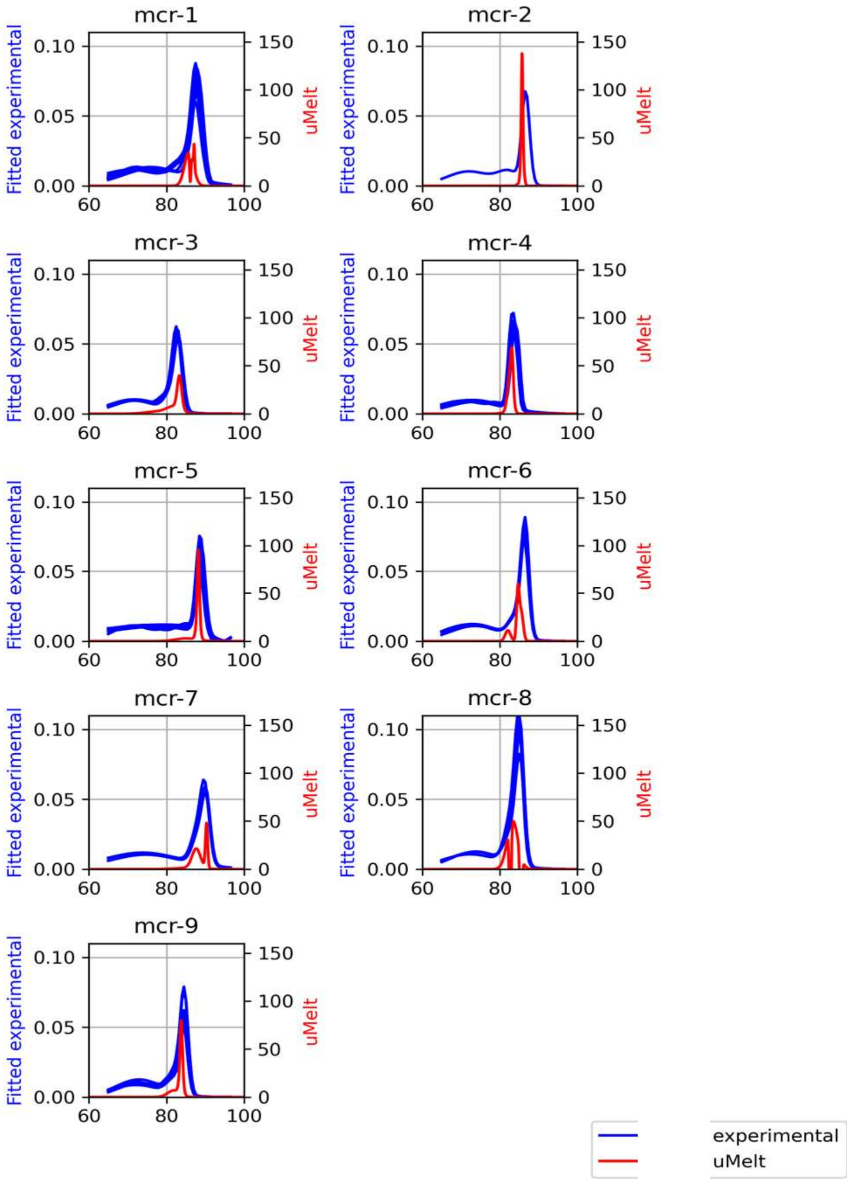
Simulated Melt Curves vs Experimental Melt Curves

- uMelt simulated curve peak close to experimental peak

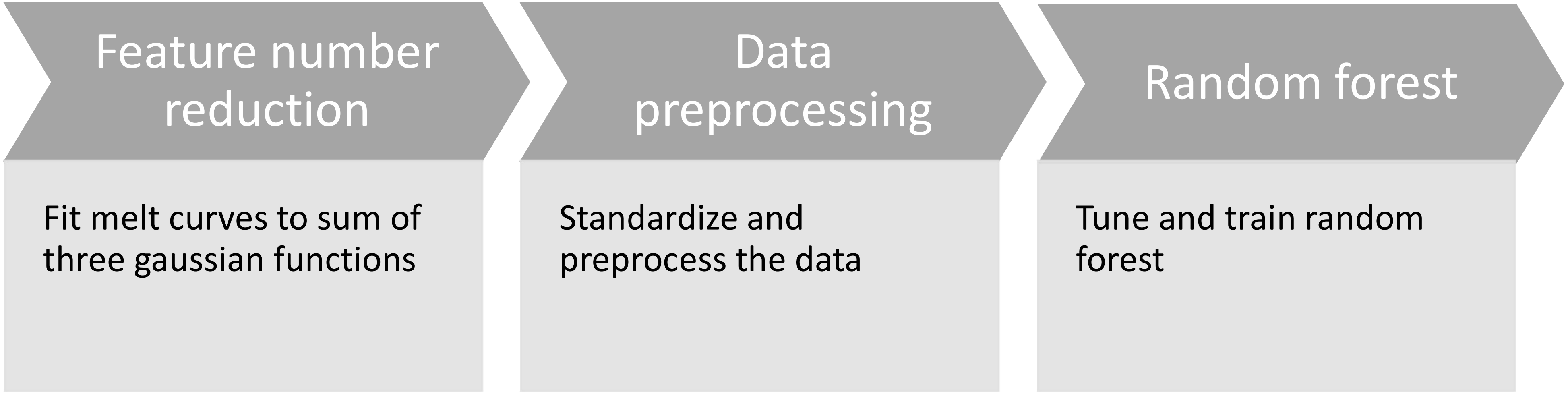
- Issue with scales for data standardization
  - uMelt melt curve is given by the inverse derivative of helicity
  - Experimental melt curves have been scaled in an unknown way

- The overall similarity between experimental and uMelt curves suggests that a simple regressor should be able to predict the experimental curves from the uMelt curve

Experimental curves compared with uMelt simulated curves



The three steps



Feature number reduction

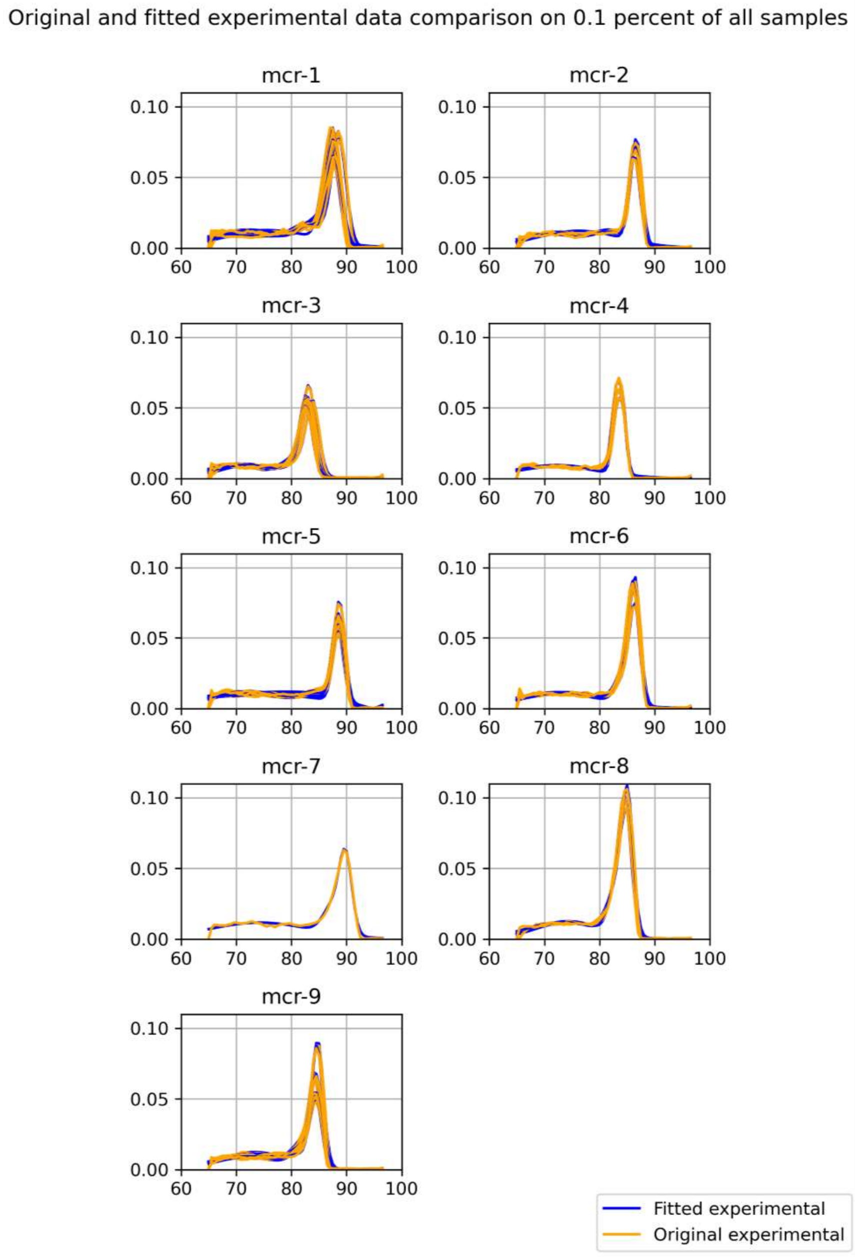
Fit melt curves to sum of three gaussian functions

Data preprocessing

Standardize and preprocess the data

Random forest

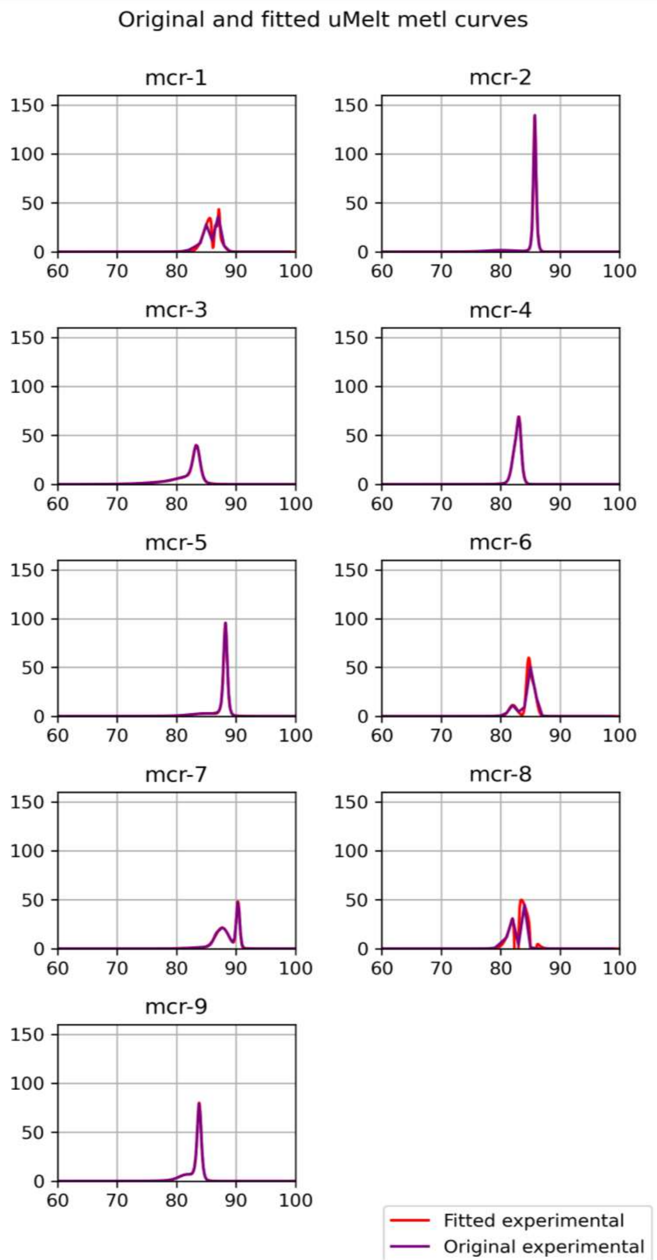
Tune and train random forest



Original and fitted experimental data comparison on 0.1 percent of all samples

Fitting Experimental and Simulated Curves to Gaussian Functions

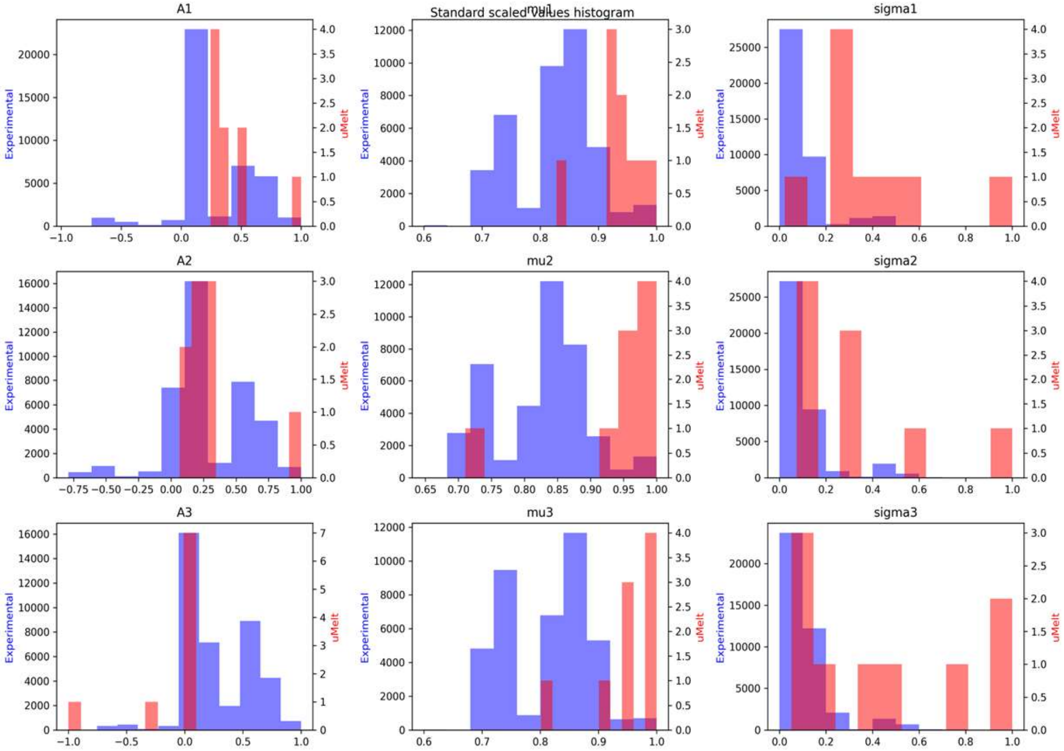
- All curves are fitted to the sum of 3 gaussians to reduce the number of features (40 to 160 depending on the resolution to 9)
- The features are ordered so that the first gaussian has the largest amplitude and the last one the smallest
- The fittings are very satisfactory
- The uMelt curves for targets mcr-1, mcr-6 and mcr-8 are limited to a lower resolution (1 point per degree) since the amplicons are longer than 500 bases.



Original and fitted uMelt metl curves

Preprocessing the data

- A max abs scaler is applied to all the previously extracted features (divide by maximum absolute value)
    - 3 amplitudes (A1, A2, A3)
    - 3 means (mu1, mu2, mu3)
    - 3 variances (sigma1, sigma2, sigma3)
- One scaler is used for input data, one for the output data.
- Amplicon length and GC content added to uMelt features
- 11 features of the uMelt curves as Input and 9 features of the experimental curves as output of the MLP



Tuning and training the MLP regressor

- Two strategies were tested:

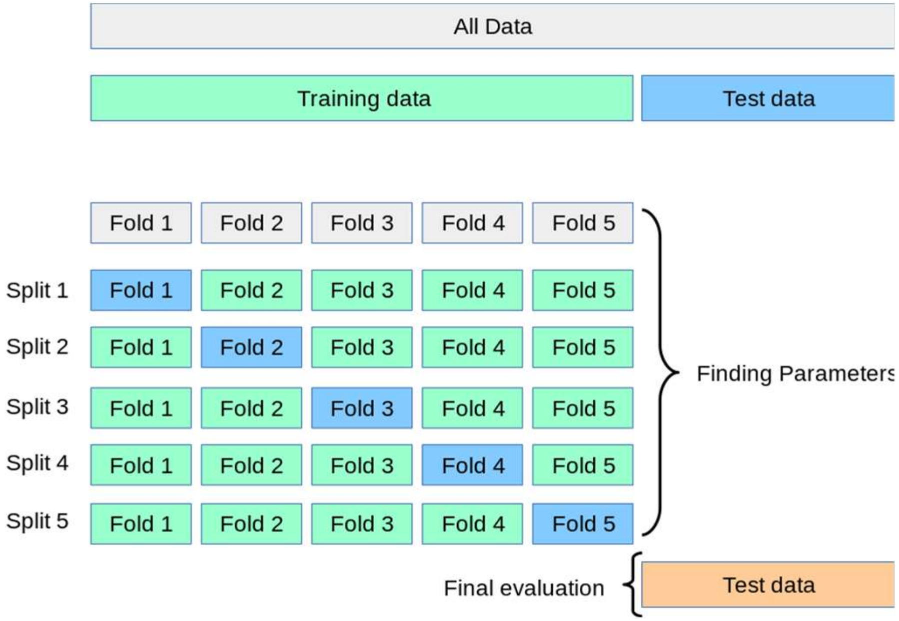
  1. The test data is comprised of one target.

  The best parameters for the training data is estimated and then evaluated on the data set that the random forest has never seen.

  2. The test data is taken randomly as 20% of the overall data (usually from all targets), used to check if MLP is learning correctly

  The best parameters for the training data is estimated and then evaluated on the testing data set.

- Parameters tuned:
  - Number of estimators: number of trees
  - Max depth: maximum depth of the tree
  - Max features: The number of features to consider when looking for the best split



All Data

Training data

Test data

Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Finding Parameters

Final evaluation

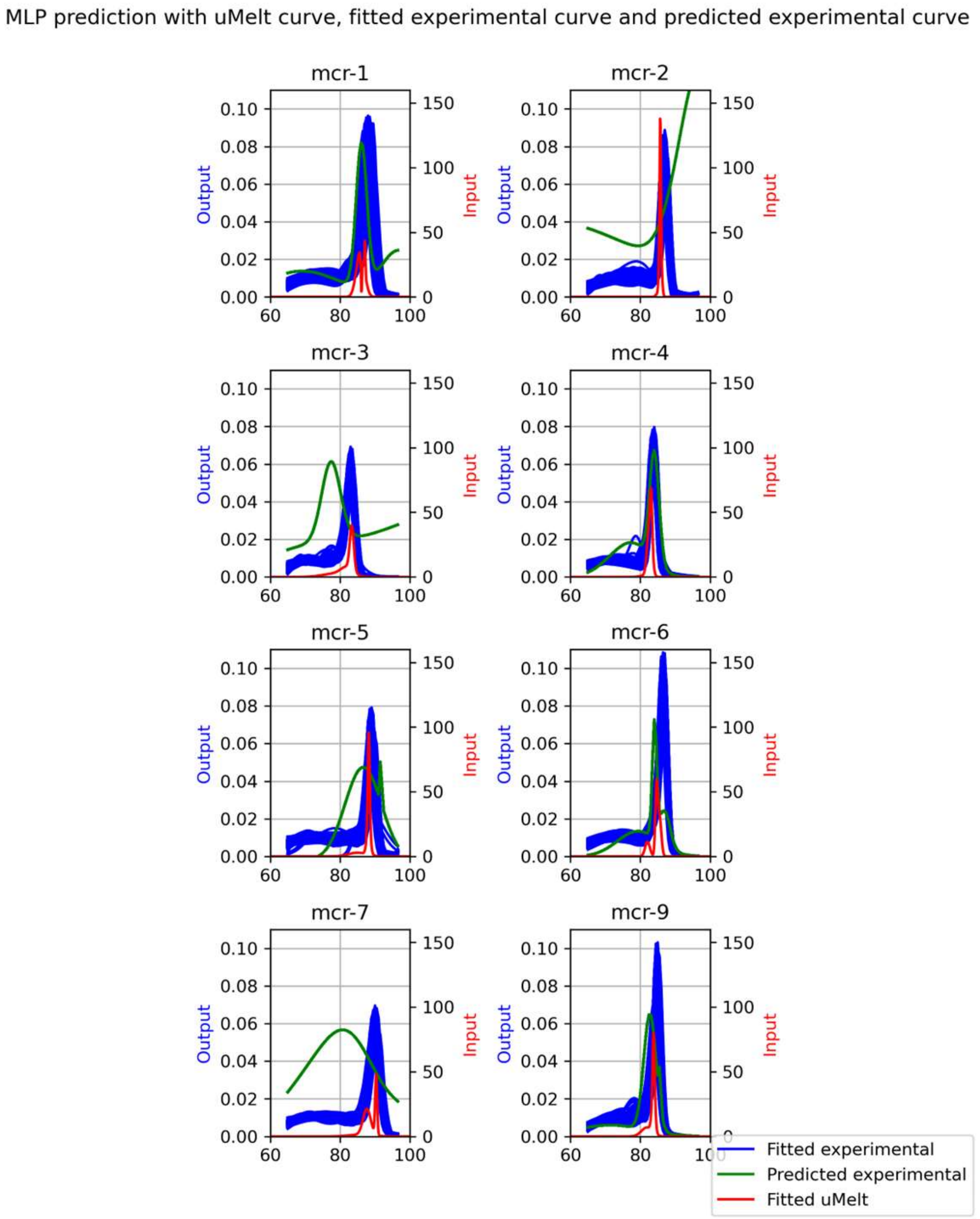
Test data

Visualization of the random forest tuning

https://scikit-learn.org/stable/modules/cross_validation.html

Strategy 1

The test data is comprised of one target, i.e no data leakage

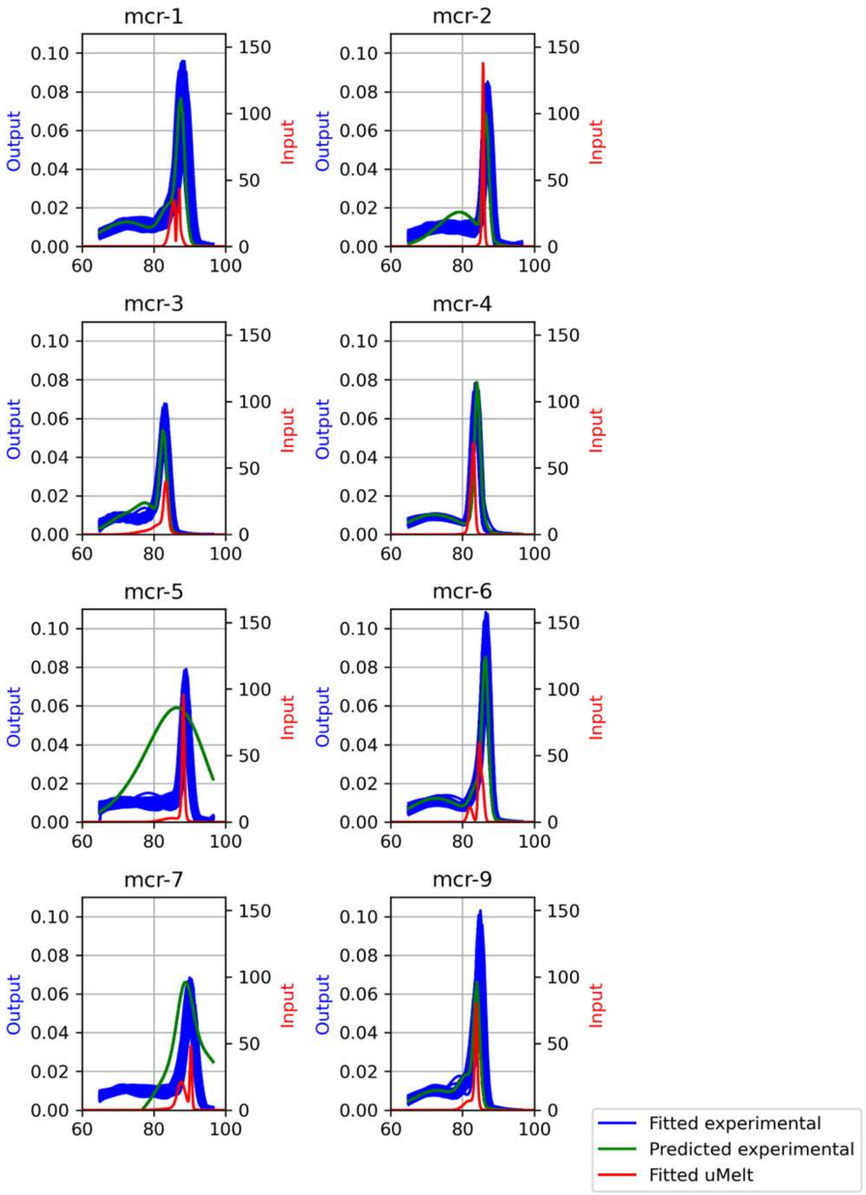


Strategy 2

The test data is taken randomly as 20% of the overall data (usually from all targets)

Remark: mcr-5 and mcr-7 shows a problem with the MLP's learning

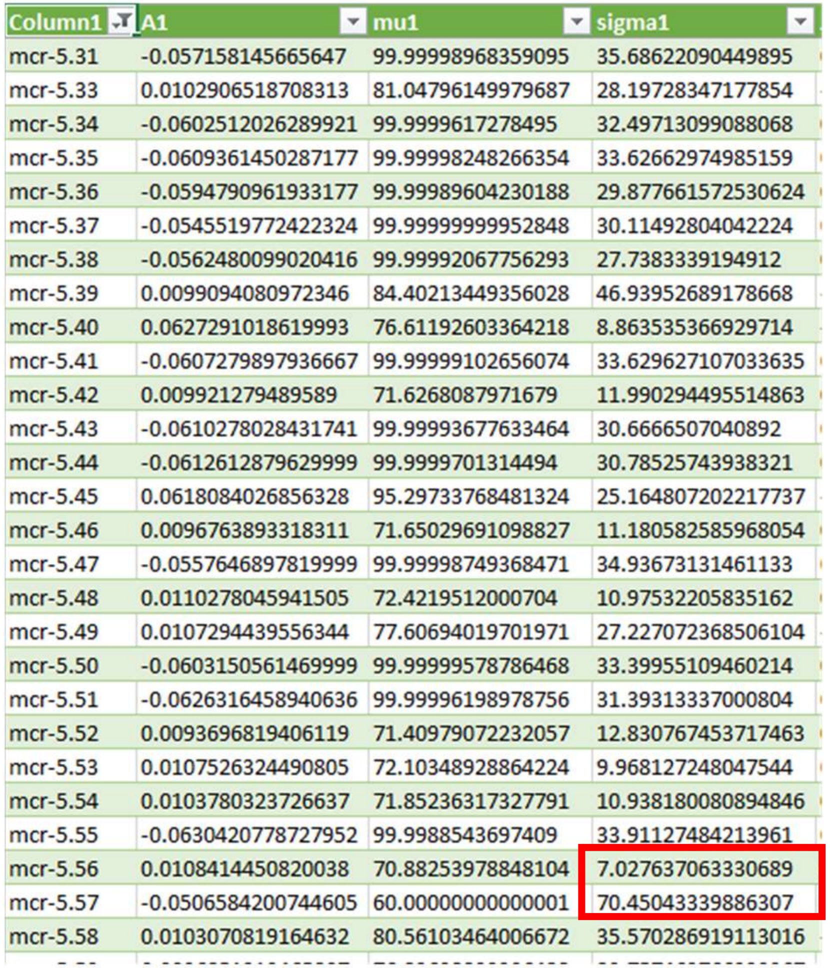
Random forest prediction with uMelt curve, fitted experimental curve and predicted experimental curve



Identified issue

- Example shown on the right:
  - Variability of sigma is too big
  - Remark: this is due to using the amplitude of the gaussians as feature arranging method

- Try other feature arranging methods:
  - By gaussian mean and gaussian variance



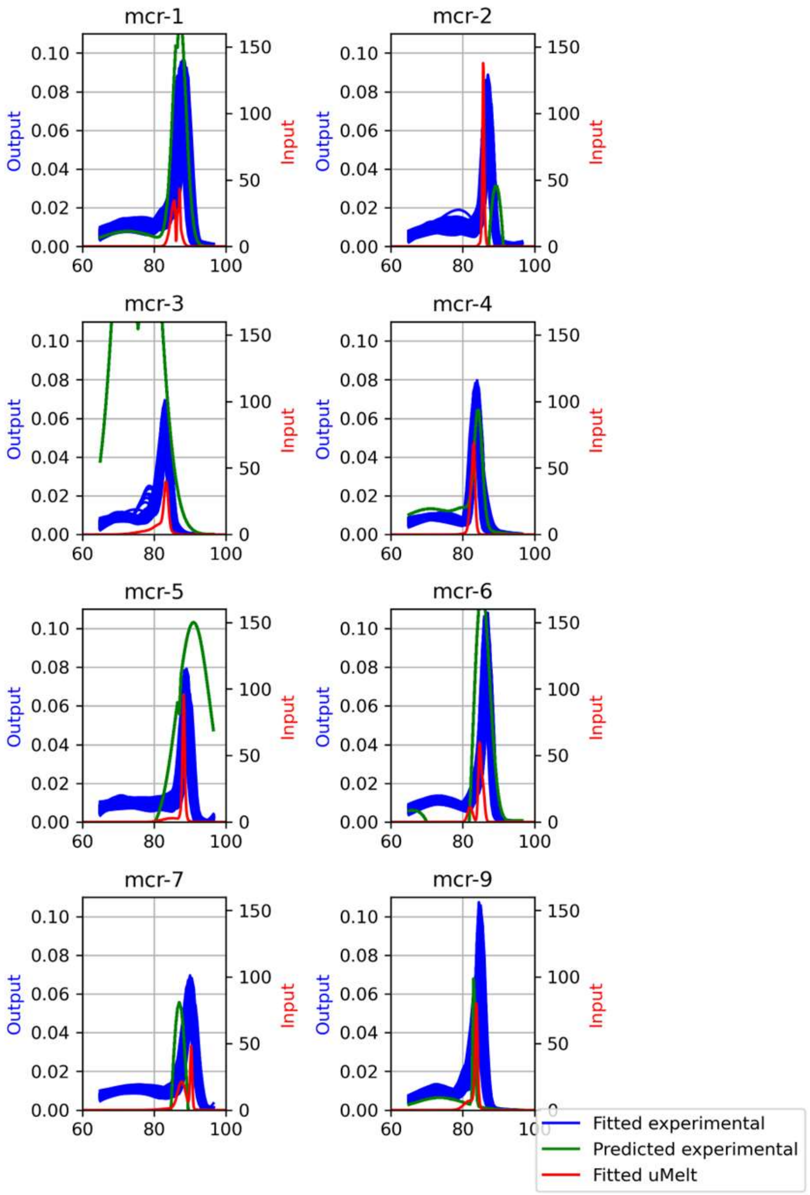
Column1	A1	mu1	sigma1
mcr-5.31	-0.057158145665647	99.99998968359095	35.68622090449895
mcr-5.33	0.0102906518708313	81.04796149979687	28.19728347177854
mcr-5.34	-0.0602512026289921	99.9999617278495	32.49713099088068
mcr-5.35	-0.0609361450287177	99.99998248266354	33.62662974985159
mcr-5.36	-0.0594790961933177	99.99989604230188	29.877661572530624
mcr-5.37	-0.0545519772422324	99.99999999952848	30.11492804042224
mcr-5.38	-0.0562480099020416	99.99992067756293	27.7383339194912
mcr-5.39	0.0099094080972346	84.40213449356028	46.93952689178668
mcr-5.40	0.0627291018619993	76.61192603364218	8.863535366929714
mcr-5.41	-0.0607279897936667	99.99999102656074	33.629627107033635
mcr-5.42	0.009921279489589	71.6268087971679	11.990294495514863
mcr-5.43	-0.0610278028431741	99.99993677633464	30.6666507040892
mcr-5.44	-0.0612612879629999	99.9999701314494	30.78525743938321
mcr-5.45	0.0618084026856328	95.29733768481324	25.164807202217737
mcr-5.46	0.0096763893318311	71.65029691098827	11.180582585968054
mcr-5.47	-0.0557646897819999	99.99998749368471	34.93673131461133
mcr-5.48	0.0110278045941505	72.4219512000704	10.97532205835162
mcr-5.49	0.0107294439556344	77.60694019701971	27.227072368506104
mcr-5.50	-0.0603150561469999	99.99999578786468	33.39955109460214
mcr-5.51	-0.0626316458940636	99.99996198978756	31.39313337000804
mcr-5.52	0.0093696819406119	71.40979072232057	12.830767453717463
mcr-5.53	0.0107526324490805	72.10348928864224	9.968127248047544
mcr-5.54	0.0103780323726637	71.85236317327791	10.938180080894846
mcr-5.55	-0.0630420778727952	99.9988543697409	33.91127484213961
mcr-5.56	0.0108414450820038	70.88253978848104	7.027637063330689
mcr-5.57	-0.0506584200744605	60.00000000000001	70.45043339886307
mcr-5.58	0.0103070819164632	80.56103464006672	35.570286919113016

Strategy 1 variance arranged features

The test data is comprised of one target, i.e no data leakage

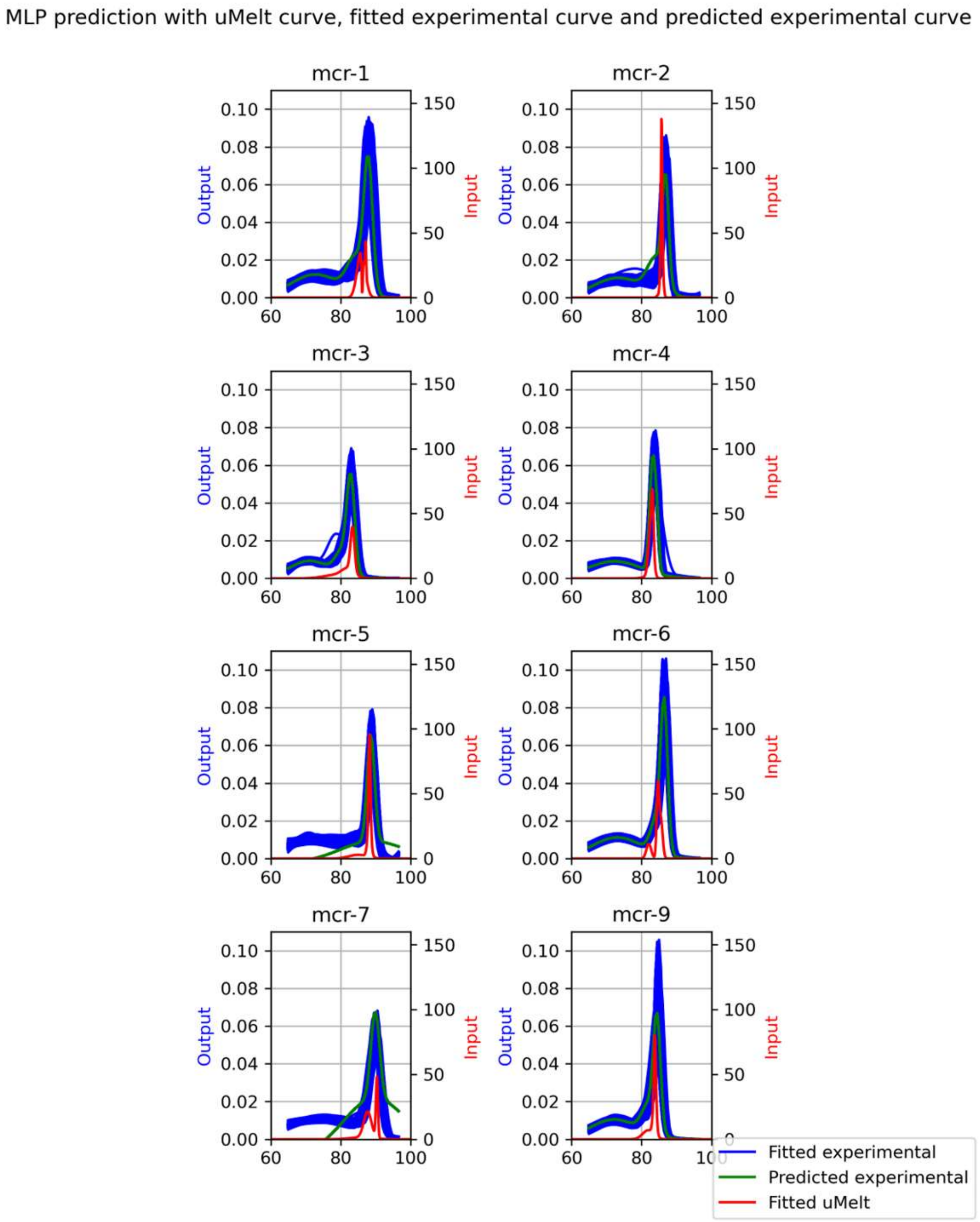
Remark : Variance arrangement gives the best result with data leakage (strategy 2)
Result without data leakage still very lacking

MLP prediction with uMelt curve, fitted experimental curve and predicted experimental curve



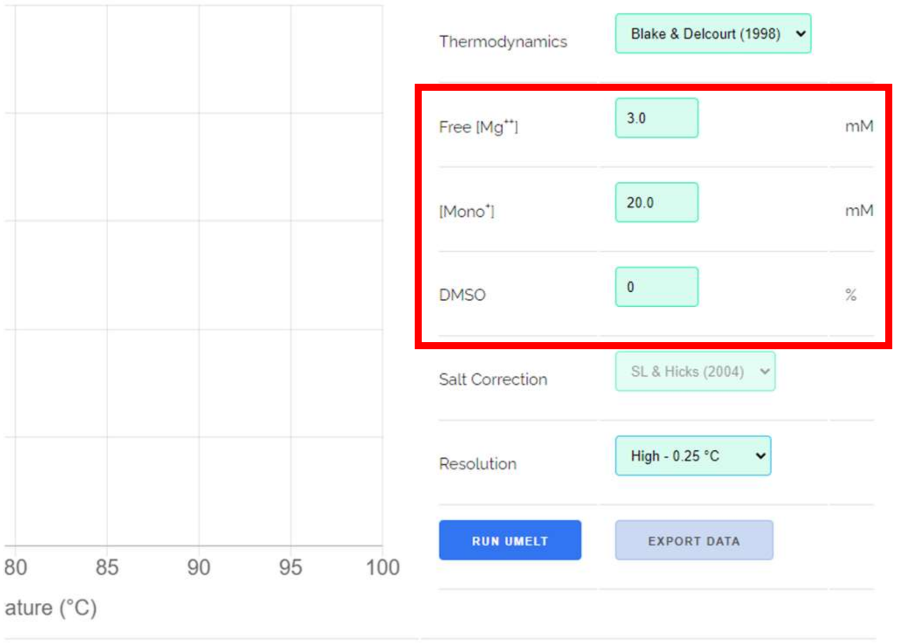
Strategy 2 variance arranged features

The test data is taken randomly as 20% of the overall data (usually from all targets)



Next step

- Introduce variability in uMelt prediction
  - Free [Mg+]
  - [Mono+]
  - DMSO
- Separate the experimental data by unique pannel and experiment id combination (exp setting id)
  - Reasoning: each different pannel and experiment id combination refers to one specific experimental setting
- Match the most similar uMelt prediction with the group of exp setting id for training



Thermodynamics: Blake & Delcourt (1998)

Free $[Mg^{++}]$: 3.0 mM

$[Mono^{+}]$: 20.0 mM

DMSO: 0 %

Salt Correction: SL & Hicks (2004)

Resolution: High - 0.25 °C

uMelt user interface (cropped)

Next step (in parallel)

- Simplify the model:
  - Use one gaussian as fitting function
  - Fitting the original melt curve to a sigmoid then either use these features or derive them and use these

- Fix uMelt resolution problem
  - uMelt only provide low resolution curves for sequences above 500 bp, (1 point per degree), thus creating weird double peaks (mcr-1, mcr-6 and mcr-8)
  - Use sigmoid to fit the original MC curve