Trygve Nelson

Quantitative Structure-Chromatography Relationship Portfolio

Dep 390

trygve.nelson94@gmail.com
https://github.com/trygvenelson94

Preface **3**
PredElute **4**
Pedelute Schematic **4**
PredElute Model 1: Elution w/ modifiers on weak, strong, and MMC Cation Exchange Resins **5**
Goal 5
Limitations 5
Results and Discussion 5
Leave-one-out validation plots 8
Elution with modifiers forecast example 13
Feature Importance of fractional residue surface area on models with r2 > 0.6 14
PredElute Model 2: pH-dependent elution on SP Sepharose **17**
Goal 17
Limitations 17
Results and Discussion 17
Leave one out validation plots 19
Extrapolated elution prediction 20
Elution as a function of pH and extrapolation forecast example 21
Chromortho **22**
ChromOrtho Schematic 22
Goal 23
Limitations 23
Results and Discussion 23
Model performance 24

Preface

This portfolio presents a series of sequence-driven Quantitative Structure–Chromatography Relationship (QSCR) projects aimed at improving practical chromatography workflows by recommending conditions such as resin type, pH, and mobile-phase modifiers directly from protein sequence. It is intended primarily as a demonstration of my computational biology skill set and of the kinds of proof-of-concept tools I can independently design and deliver over the course of a few months of personal time. I am relatively new to Python, and part of the purpose of this portfolio is to show the level of end-to-end functionality I can already achieve while I continue to build my coding experience.

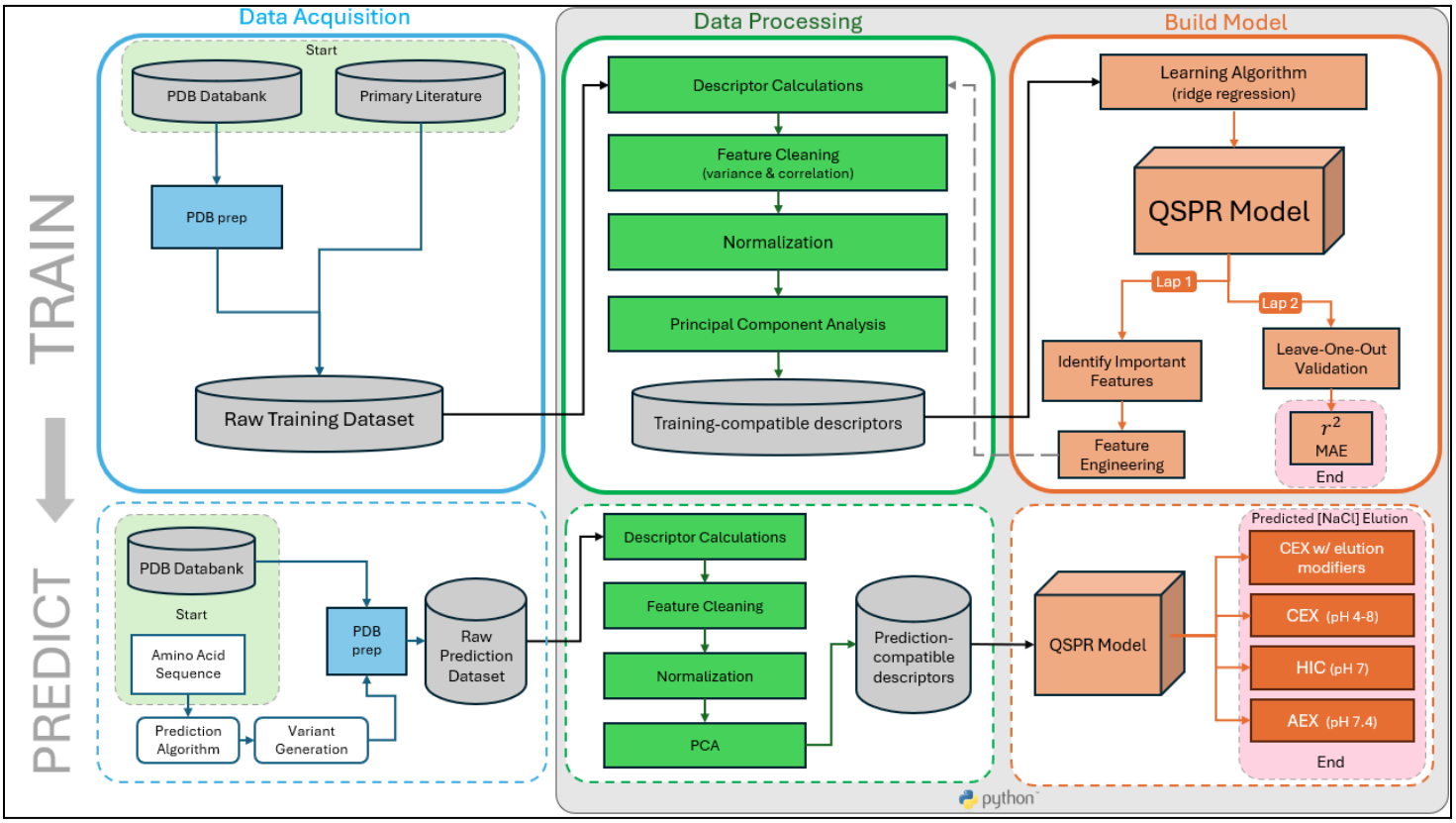
The emphasis here is less on proposing fundamentally new chromatographic theory and more on showing that established ideas from the literature can be turned into working, end-to-end pipelines: harvesting and cleaning experimental data, generating protein and ligand descriptors, engineering additional features, and training and cross-validating machine-learning models that make actionable predictions for unseen sequences. The models and architectures showcased herein should be viewed as research prototypes and starting points rather than polished, production-ready tools or novel theories.

While these models have been rigorously evaluated using cross-validation schemes designed to test generalization to unseen proteins, predictions for protein variants such as truncations, dimers, or other engineered constructs should be treated as hypotheses that still require experimental verification. In particular, extrapolated forecasts and variant-level predictions are meant to guide thinking and experiment design, not replace wet-lab confirmation. Finally, to avoid any conflicts around intellectual property or proprietary information, all input data sources and pre-built Python tools used in these projects are drawn from the public domain (academic journals, PDB, GitHub, etc...), and no internal or confidential data are included in this work.

PredElute

Pedelute Schematic

Schematic represents both PredElute Model 1 and Model 2 architecture.



PredElute Model 1: Elution w/ modifiers on weak, strong, and MMC Cation Exchange Resins

Goal

Using published data and taking inspiration Cramer's research group[1], train a QSCR model that estimates the elution concentration of a protein and related isoforms (truncations, dimers, etc.) using just the amino acid sequence as the input on weak, strong, and multimodal cation exchange resins with mobile phase elution modifiers.

Limitations

True recorded elution concentration values are not known as elution data is only presented in a visual histogram. Therefore harvesting of elution concentration data required building and incorporation of a pixel counting algorithm to estimate values. The low pixel density of the histogram plot limited the accuracy of these values resulting in some proteins having identical NaCl elution concentration as the values approach zero.

Results and Discussion

Protein elution concentrations were harvested using a home-made pixel-counting tool built in Python. Proteins were retrieved from the Protein Data Bank and missing structural information were completed with AlphaFold. Two protein pools were tested: one including all proteins tested and another with protein homologues merged into a single structure. Protein descriptors were calculated using Schrodinger and ProDes protein descriptor tools and three combinations of descriptor pools were tested for optimal performance: Schrodinger descriptors alone, ProDes descriptors alone, and both Schrodinger and ProDes descriptors combined. In an attempt to improve model performance, additional interaction descriptors were engineered, however none of the engineered features resulted in performance improvement for this example (data not shown). Descriptors were then refined by variance and feature correlation and then normalized. Finally, dimensionality reduction was conducted via principal component analysis and used to train multiple machine learning algorithms to determine the best method. A parameter grid search was implemented to find the optimal model parameters for each condition. They are shown in **Table 1**.

Parameter	Values included in search	# of unique values
Alpha	0.1, 1, 5, 10, 50, 100, 200	7
Number of PCA Components	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	19
Variance Threshold	0, 0.01, 0.05, 0.1	4
Correlation Threshold	0.75, 0.80, 0.85, 0.90, 0.95, 0.97, 0.99, 1.0	8
Number of Seeds	42, 43, 44, 45, 46	5
Descriptor Pool	Schrodinger, ProDes, Schrodinger + ProDes	3
Protein Pool	Full, Merged (“deduped”)	2
ML Algorithms	Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Random Forest Model, Gradient Boost	6
Total number of models tested per condition		**766,080**

Table 1: Parameter grid search values used to identify the optimal parameters for each model.

experiment	experiment_name	condition	dataset_used	n_proteins	n_features	alpha	n_components	var_thresh	corr_thresh	r2_full	r2_dedup	r2_used	mae_used
1	Arginine on CEX (pH6)	0M	full	20	103	0.1	8	0	1	0.77	0.64	0.77	0.03
		0.025M	full	20	41	0.1	9	0	0.9	0.77	0.72	0.77	0.04
		0.1M	full	19	69	0.1	14	0	0.975	0.74	0.72	0.74	0.03
2	Arginine on CEX (pH6)	0M	dedup	14	19	1	7	0.05	0.95	0.07	0.44	0.44	0.22
		0.025M	full	21	36	0.1	8	0	0.85	0.51	0.23	0.51	0.15
		0.1M	dedup	14	37	0.1	12	0.01	0.99	-0.74	0.61	0.61	0.16
3	Arginine on CEX (pH6)	0M	full	20	69	0.1	9	0	0.975	0.8	0.53	0.8	0.03
		0.1M	dedup	13	64	0.1	11	0	0.975	0.63	0.78	0.78	0.03
		0.025M	full	20	20	0.1	15	0.05	0.95	0.82	0.64	0.82	0.03
4	Arginine on CEX (pH6)	0M	dedup	14	19	1	7	0.05	0.95	0.05	0.45	0.45	0.22
		0.025M	full	21	69	0.1	15	0	0.975	0.53	-0.03	0.53	0.17
		0.1M	full	21	69	0.1	15	0	0.975	0.57	0.08	0.57	0.19
5	Arginine on CEX (pH6)	Capto MMC	full	17	43	50	3	0	0.9	-0.07	-0.49	-0.07	0.48
		SP Seph	full	17	102	0.1	13	0	1	0.75	0.42	0.75	0.05
		CM Seph	full	17	66	0.1	15	0	0.975	0.79	0.53	0.79	0.03
6	Arginine on CEX (pH6)	Capto MMC	dedup	13	18	0.1	8	0.075	0.95	-0.9	0.49	0.49	0.22
		CM Seph	full	20	69	0.1	9	0	0.975	0.77	0.56	0.77	0.03
		SP Seph	full	20	103	0.1	8	0	1	0.75	0.6	0.75	0.04
7	Arginine on CEX (pH6)	No modifier	dedup	14	19	1	7	0.05	0.95	0.07	0.45	0.45	0.21
		20% ethylene glycol	dedup	14	32	0.1	12	0	0.8	0.64	0.84	0.84	0.07
		20% propylene glycol	dedup	14	32	0.1	12	0	0.8	0.67	0.78	0.78	0.08
8	Arginine on CEX (pH6)	0.01M	full	21	103	0.1	9	0	1	0.72	0.5	0.72	0.04
		0M	full	21	69	0.1	11	0	0.975	0.85	0.66	0.85	0.02
		0.025M	full	21	21	0.1	18	0.05	0.95	0.84	0.79	0.84	0.03
9	Arginine on CEX (pH6)	0M	full	21	36	0.1	8	0	0.85	0.5	0.31	0.5	0.2
		0.01M	full	21	56	0.1	14	0	0.95	0.7	0.4	0.7	0.14
		0.025M	full	21	69	0.1	14	0	0.975	0.75	0.37	0.75	0.14

Table 2: Table of optimal parameter grid search results by model and condition. The column “dataset_used” indicates which pool yielded the best performing model was trained on where “full” indicates the best model was trained using the full protein pool and “dedup” indicates the best model was trained using the merged-homologues protein pool. “Mae_used” represents the mean average error of all five seeds. All models performed best with ProDes descriptor pool alone.

Leave-one-out validation plots

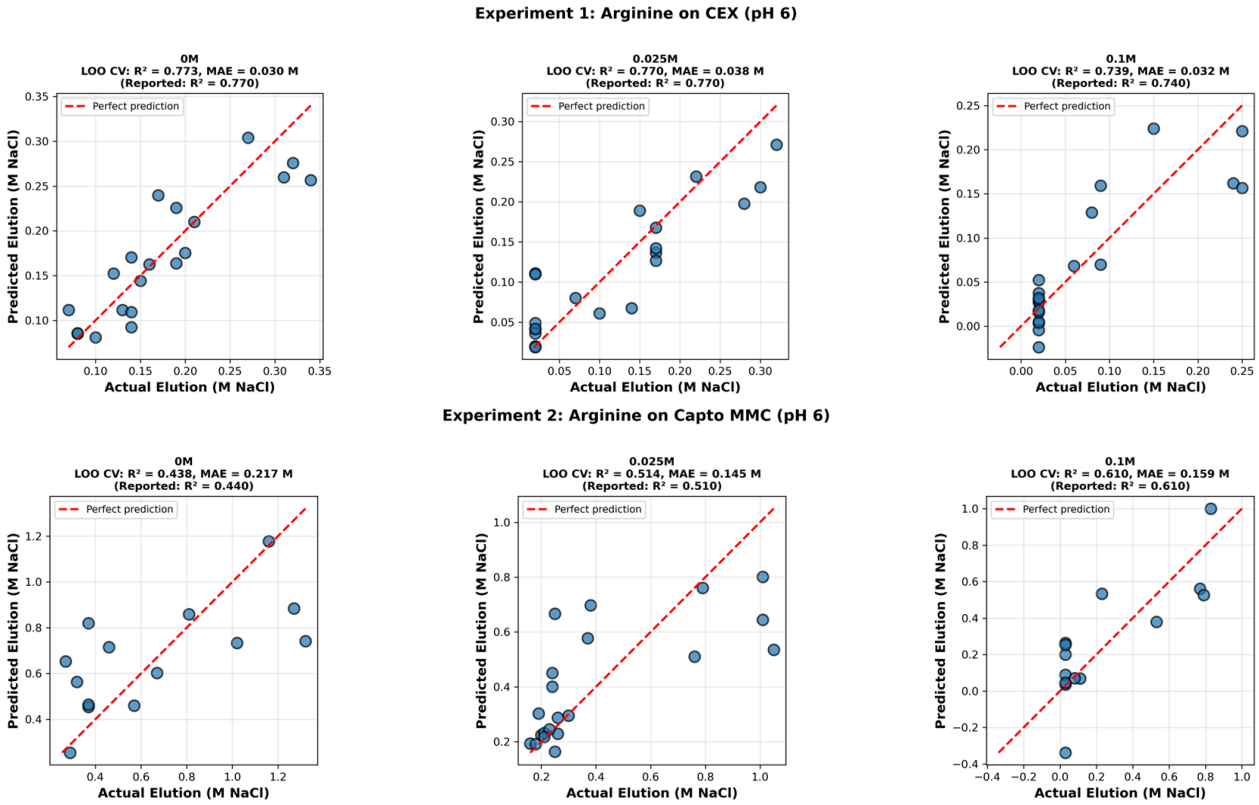


Figure 1: Leave-one-out plots of elution on MMC/CEX with arginine as mobile phase elution modifier. Limitations from the histogram resolution are apparent on the 0.1M Arginine plots exhibited as points with equal values for “Actual Elution (M NaCl)”.

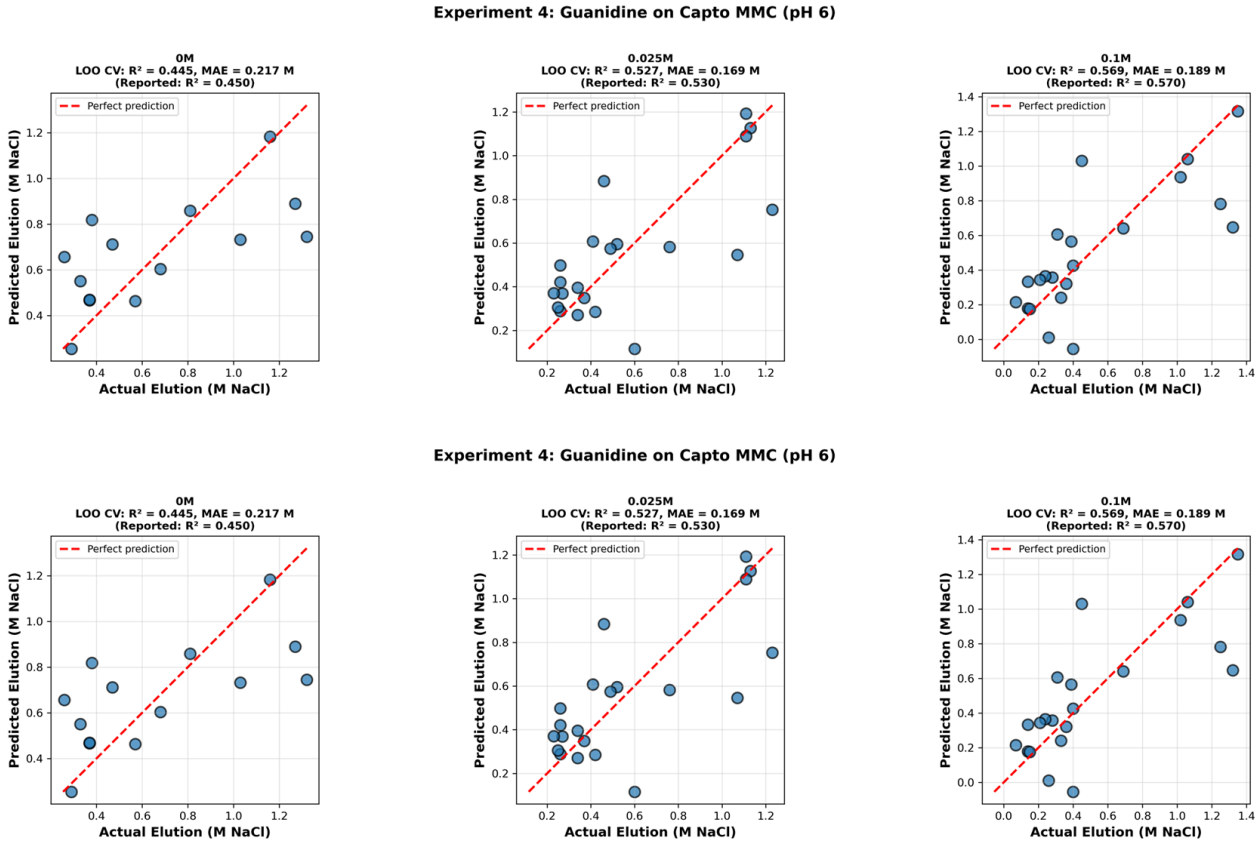


Figure 2: Leave-one-out plots of elution on MMC/CEX with guanidine as mobile phase elution modifier. Limitations from the histogram resolution are less apparent in these experiments.

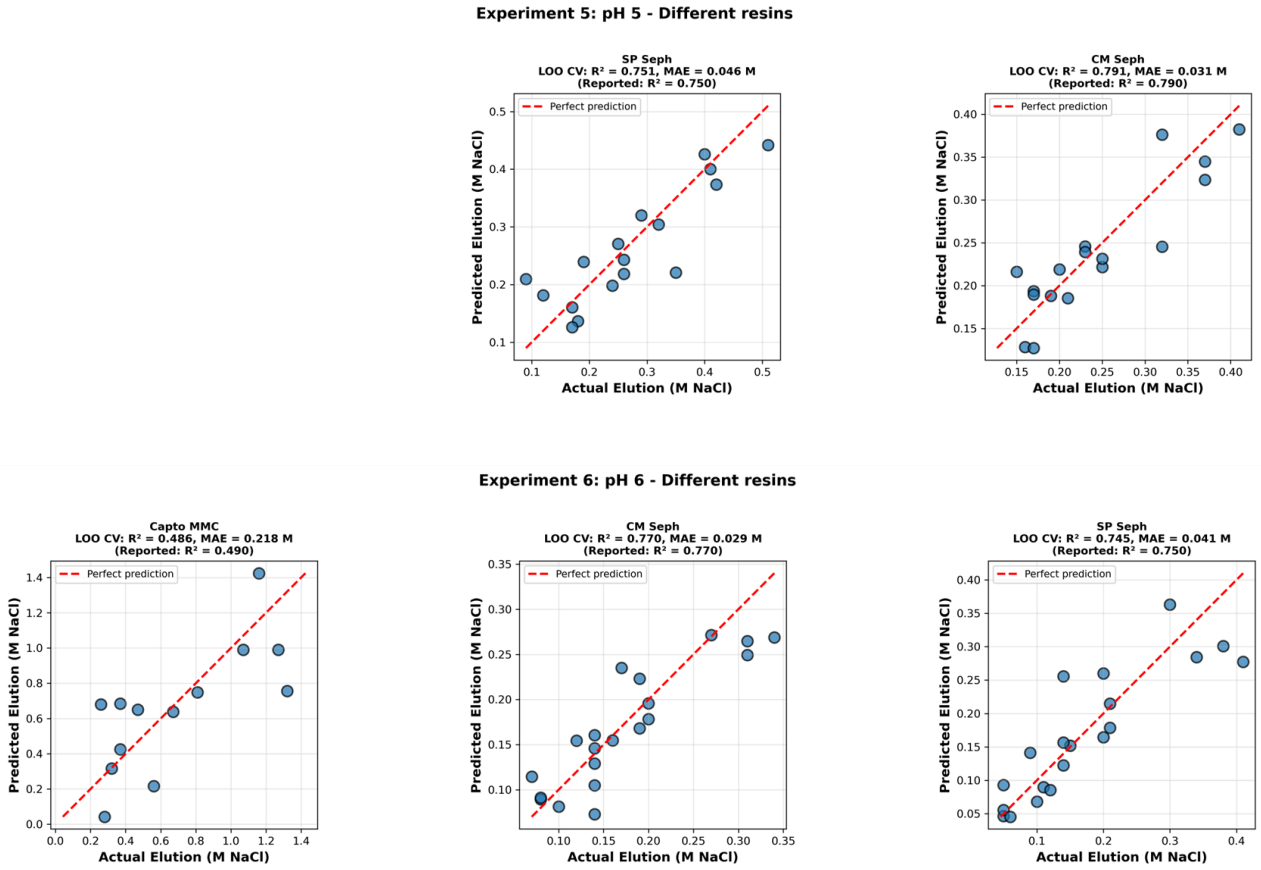


Figure 3: Leave-one-out plots of elution on MMC/CEX with no mobile phase elution modifier at pH 5 and 6. The plot for Experiment 5 Capto MMC is not shown because the r^2 for this model was below zero.

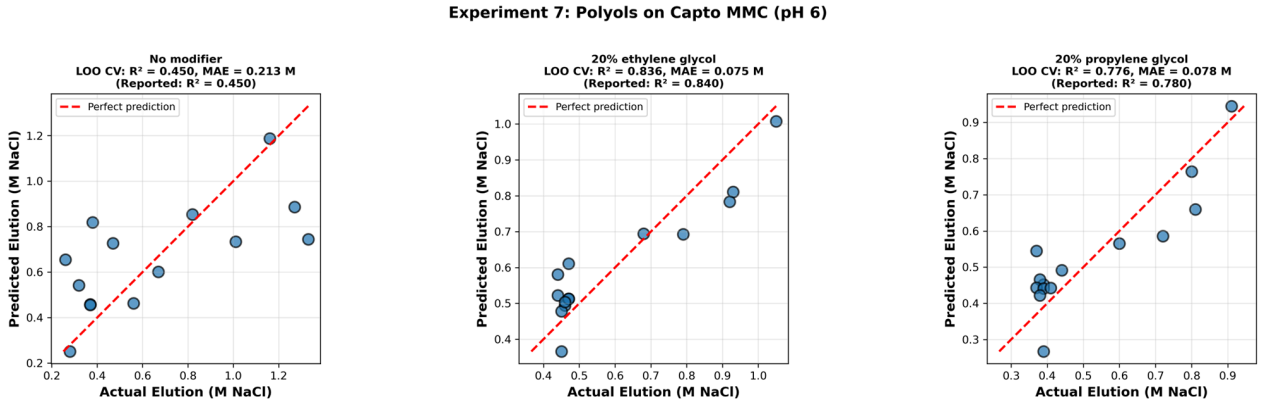


Figure 4: Leave-one-out plots of elution on MMC/CEX with ethylene glycol and polyethylene glycol as mobile phase elution modifier. Limitations from the histogram resolution are again apparent on the poly-ol plots.

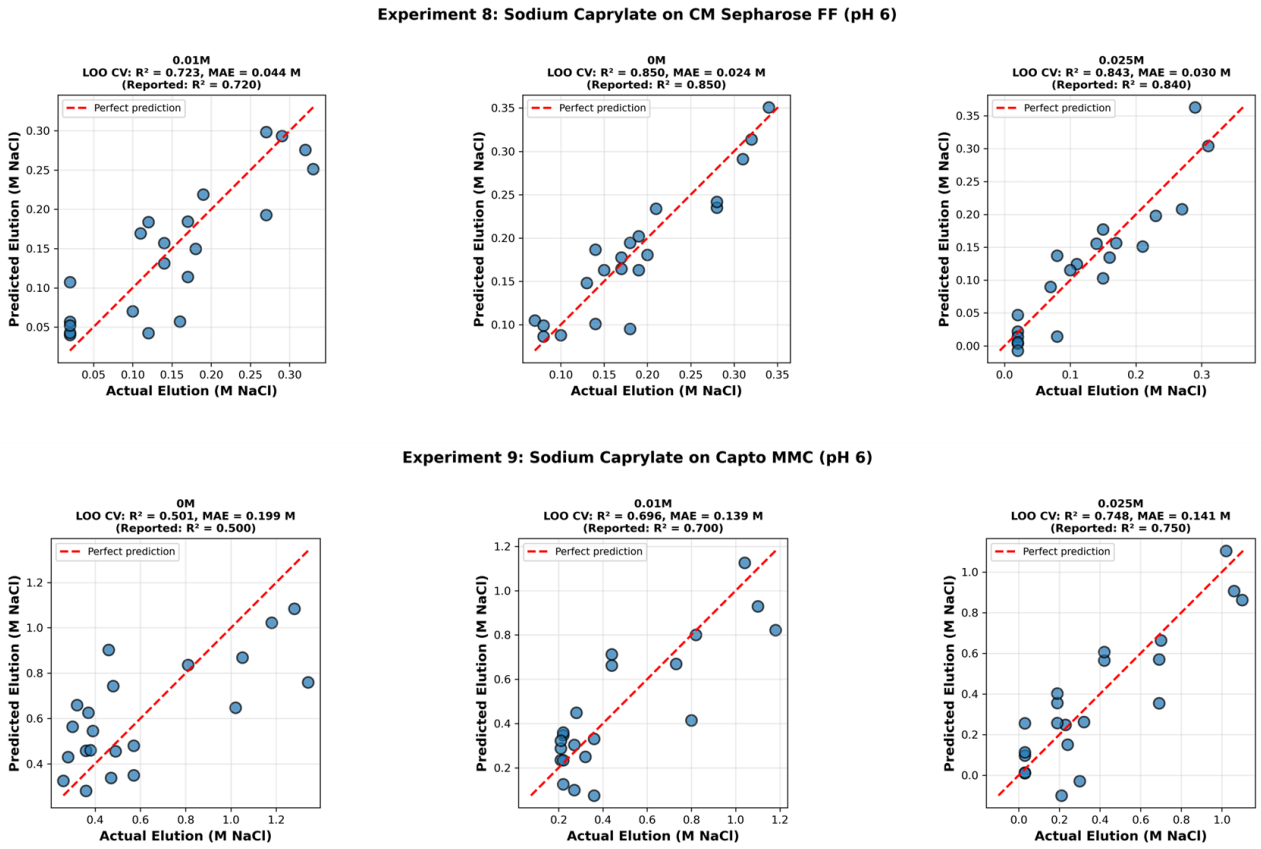
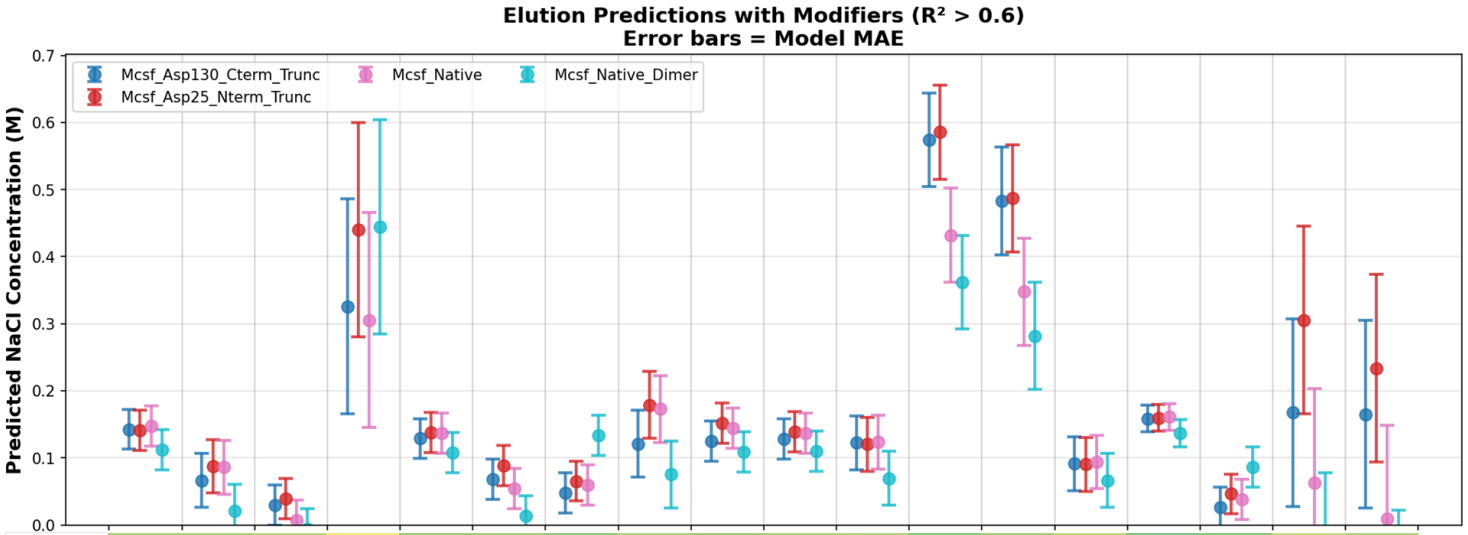


Figure 5: Leave-one-out plots of elution on MMC/CEX with sodium caprylate as mobile phase elution modifier. Limitations from the histogram resolution are less apparent but still present.

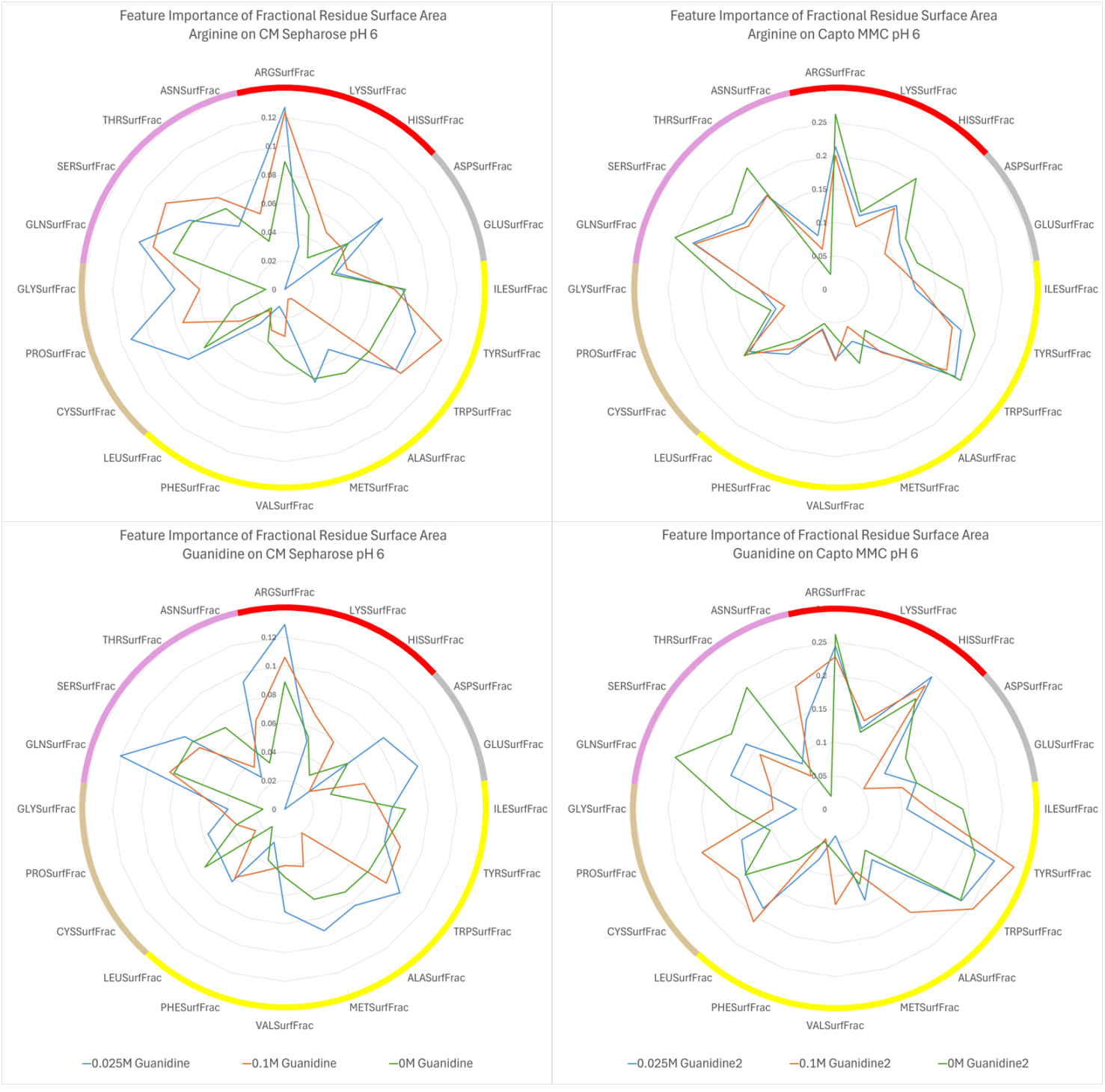
Elution with modifiers forecast example



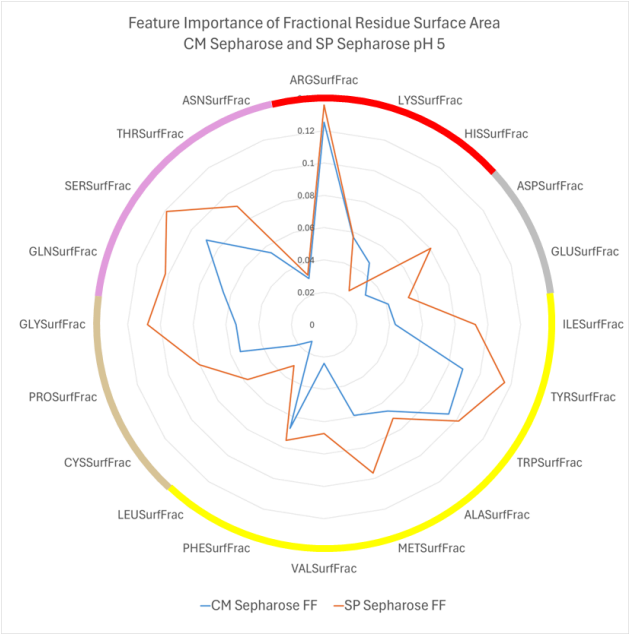
r^2	0.77	0.77	0.74	0.61	0.8	0.78	0.82	0.75	0.79	0.77	0.75	0.84	0.78	0.72	0.85	0.84	0.7	0.75
MAE (+/- error)	0.03	0.04	0.03	0.16	0.3	0.3	0.3	0.5	0.3	0.3	0.4	0.7	0.8	0.4	0.2	0.3	0.14	0.14
pH	pH 6	pH 6	pH 6	pH 6	pH 6	pH 6	pH 6	pH 5	pH 5	pH 6	pH 6	pH 6	pH 6	pH 6	pH 6	pH 6	pH 6	pH 6
modifier	Arginine	Arginine	Arginine	Arginine	Guanidine	Guanidine	Guanidine	NA	NA	NA	NA	ethylene glycol	propylene glycol	Sodium Caprylate	Sodium Caprylate	Sodium Caprylate	Sodium Caprylate	Sodium Caprylate
Modifier concentration	0 M	0.025 M	0.1 M	0.1 M	0 M	0.025 M	0.1 M	NA	NA	NA	NA	20%	20%	0 M	0.01 M	0.025 M	0.01 M	0.025 M
column	CM Seph	CM Seph	CM Seph	Capto MMC	CM Seph	CM Seph	CM Seph	CM Seph	SP Seph	CM Seph	SP Seph	Capto MMC	Capto MMC	CM Seph	CM Seph	CM Seph	Capto MMC	Capto MMC

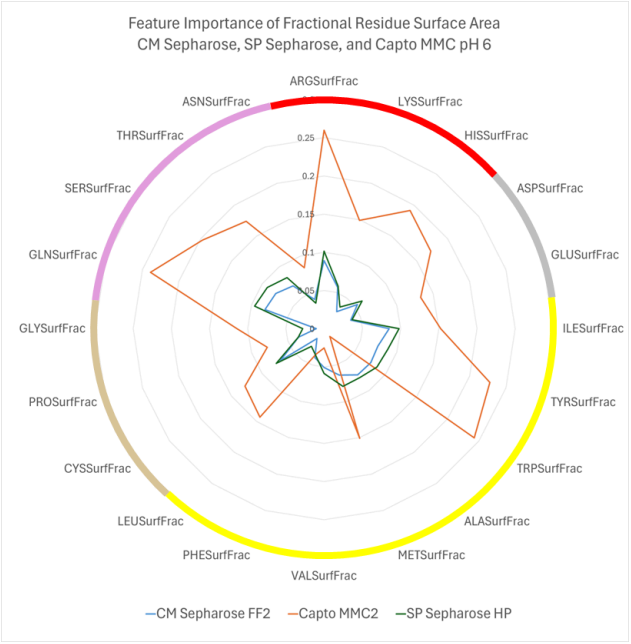
Figure 6: Example of elution forecast plot for a unique M-CSF generated in alphafold. Only models with $r^2 > 0.60$ are shown. Truncated variants were identified using Schrodinger's reactive residue identification tool and determined by the closest proteolytic site to each terminus. The dimer was generated using Schrodinger's protein-protein docking tool with the lowest-energy dimer selected for this example.

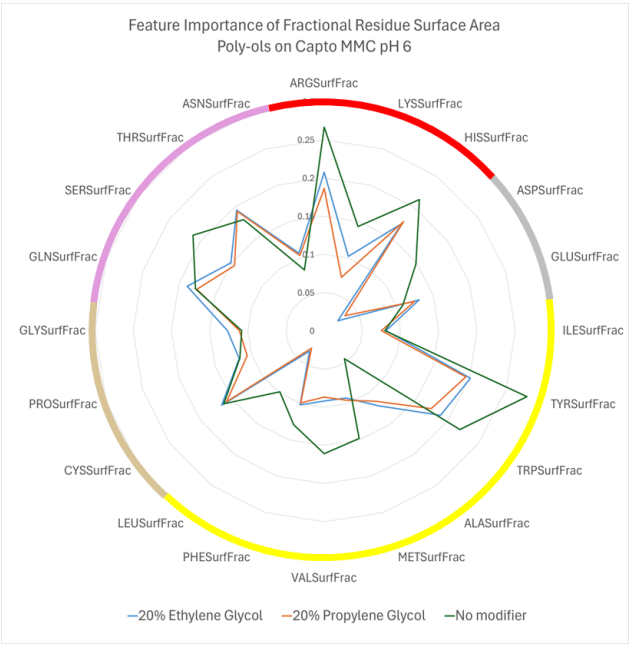
Feature Importance of fractional residue surface area on models with $r^2 > 0.6$



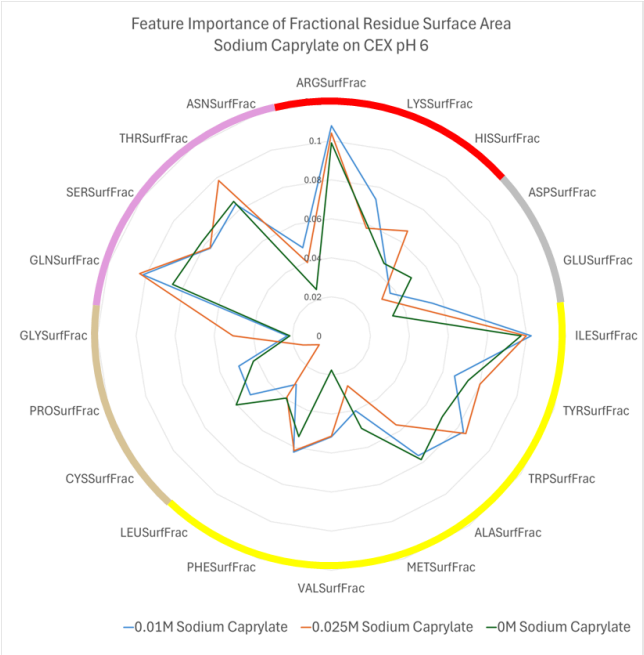
positive
negative
hydrophobic
other
polar uncharged

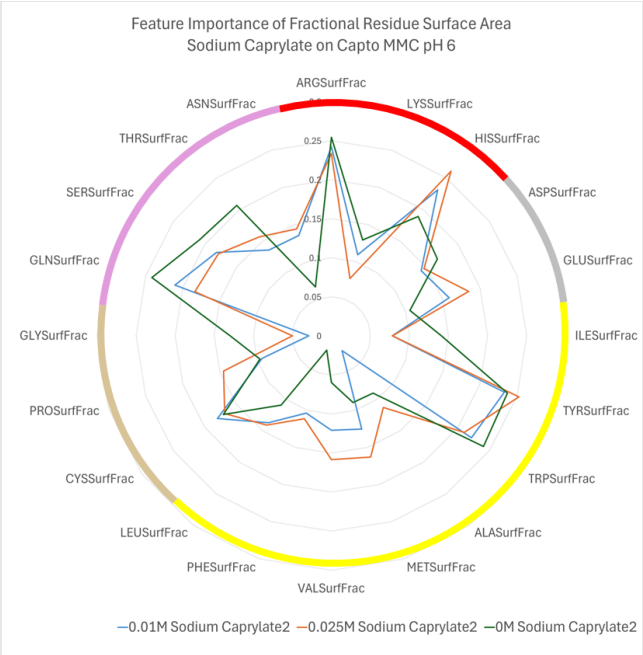






positive
negative
hydrophobic
other
polar uncharged





positive
negative
hydrophobic
other
polar uncharged

PredElute Model 2: pH-dependent elution on SP Sepharose

Goal

Using published data and taking inspiration from Cramer's research group[2], train a QSCR model that estimates the elution concentration of a protein and related isoforms (truncations, dimers, etc.) using just the amino acid sequence as the input SP Sepharose resin as a function of pH.

Limitations

Three of the proteins in the training pool are missing elution data points so they cannot be included in the training model. Additionally, lactoferrin seems to be a strong outlier and may be removed in later versions if model performance improves without a significant compromise.

Results and Discussion

Protein elution concentrations were calculated by converting retention volumes into molar NaCl concentrations based on the methods described in the paper. Proteins were retrieved from the Protein Data Bank and missing structural information were completed with AlphaFold. No protein homologues were present in this pool so pdb merging was not required. Protein descriptors were calculated using Schrodinger and ProDes protein descriptor tools and three combinations of descriptor pools were tested for optimal performance: Schrodinger descriptors alone, ProDes descriptors alone, and both Schrodinger and ProDes descriptors combined. In an attempt to improve model performance, additional interaction descriptors were engineered which resulted in a ~3% increase in model performance after incorporation. Descriptors were then refined by variance and feature correlation and then normalized. Finally, dimensionality reduction was conducted via principal component analysis and used to train multiple machine learning algorithms to determine the best method. A parameter grid search was implemented to find the optimal model parameters for each condition. They are shown in **Table 1**.

Parameter	Values included in search	# of unique values
Alpha	0.1, 1, 5, 10, 50, 100, 200	7
Number of PCA Components	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	14
Variance Threshold	0, 0.01, 0.05, 0.1	4
Correlation Threshold	0.75, 0.80, 0.85, 0.90, 0.95, 0.97, 0.99, 1.0	8
Number of Seeds	42, 43, 44, 45, 46	5
Descriptor Pool	Schrodinger, ProDes, Schrodinger + ProDes	3
ML Algorithms	Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Random Forest Model, Gradient Boost	6
Total number of models tested per condition		**282,240**

Table 3: Parameter grid search values used to identify the optimal parameters for each model.

After feature importance analysis and retraining, the following engineered features were found to improve the model performance by ~3%.

Interaction Name	**Formula**
charge_balance	sum(ARG,LYS,HIS)/sum(ASP,GLU)
arg_lys_ratio	ARGSurfFrac/LYSSurfFrac
aromatic_sum	sum(TRP,TYR,PHE)
trp_tyr_interaction	TRPSurfFrac*TYRSurfFrac
aromatic_hydrophobic	(TRPSurfFrac+TYRSurfFrac)*SurfMhpMean
aliphatic_cluster	ILESurfFrac*LEUSurfFrac*VALSurfFrac
polar_positive	(SERSurfFrac+THRSurfFrac)*(ARGSurfFrac+LYSSurfFrac)
gln_charge	GLNSurfFrac*SurfEpPosSumAverage
shape_arg	Shape max*ARGSurfFrac
pocket_charge	Shape min*SurfEpPosSumAverage
pi_charge	Isoelectric point*SurfEpPosSumAverage

Table 4: Engineered feature interaction descriptors and their respective formulas.

Leave one out validation plots

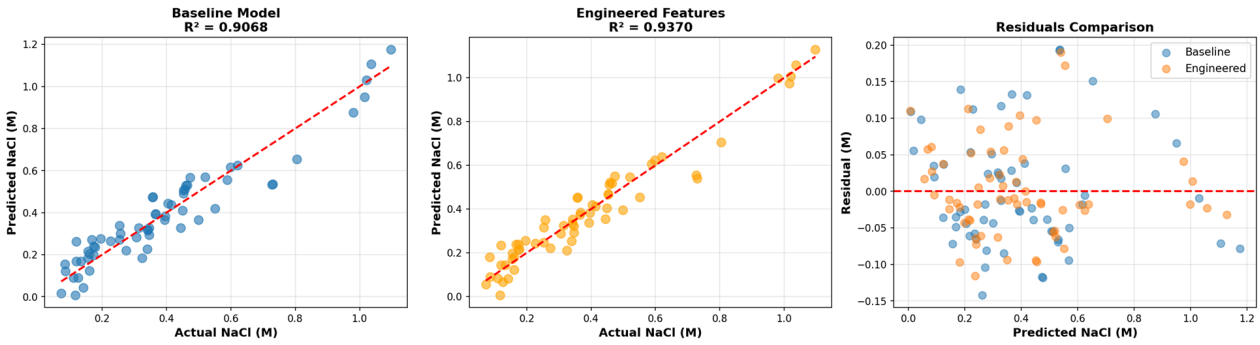


Figure 7: Results of leave-one-out validation before (blue) and after (orange) incorporation of engineered features defined in **Table 4**. In this example the engineered feature descriptors resulted in ~3% improvement in model performance (R^2 = 0.9068 to R^2 = 0.9370).

Extrapolated elution prediction

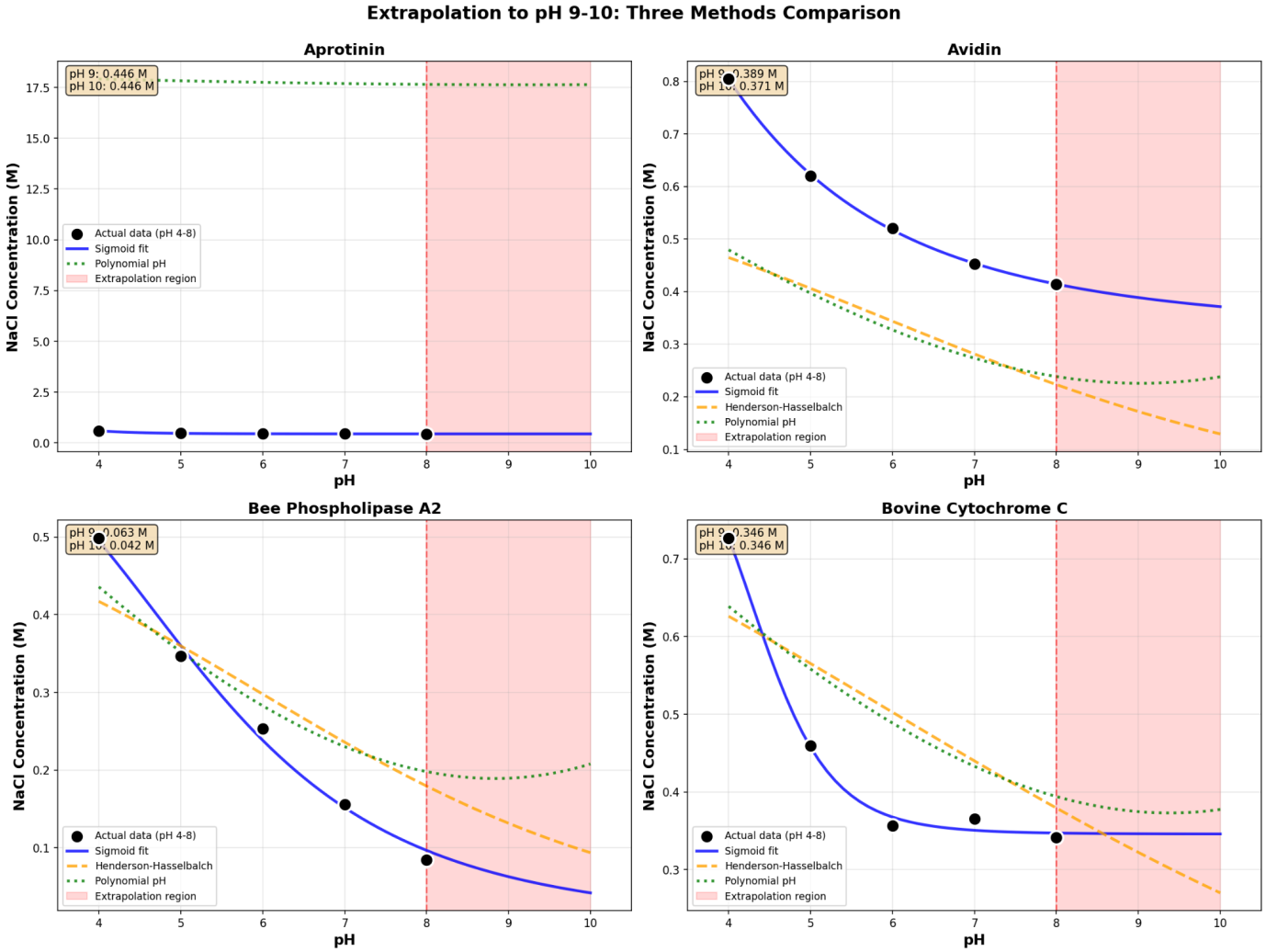


Figure 8: Analysis of various fit and extrapolate methods on four randomly selected proteins to visually evaluate which method yields the most likely elution curve extrapolated out to pH 10. Note that this extrapolation approach has yet to be experimentally validated and cannot be statistically validated with the current dataset.

Elution as a function of pH and extrapolation forecast example

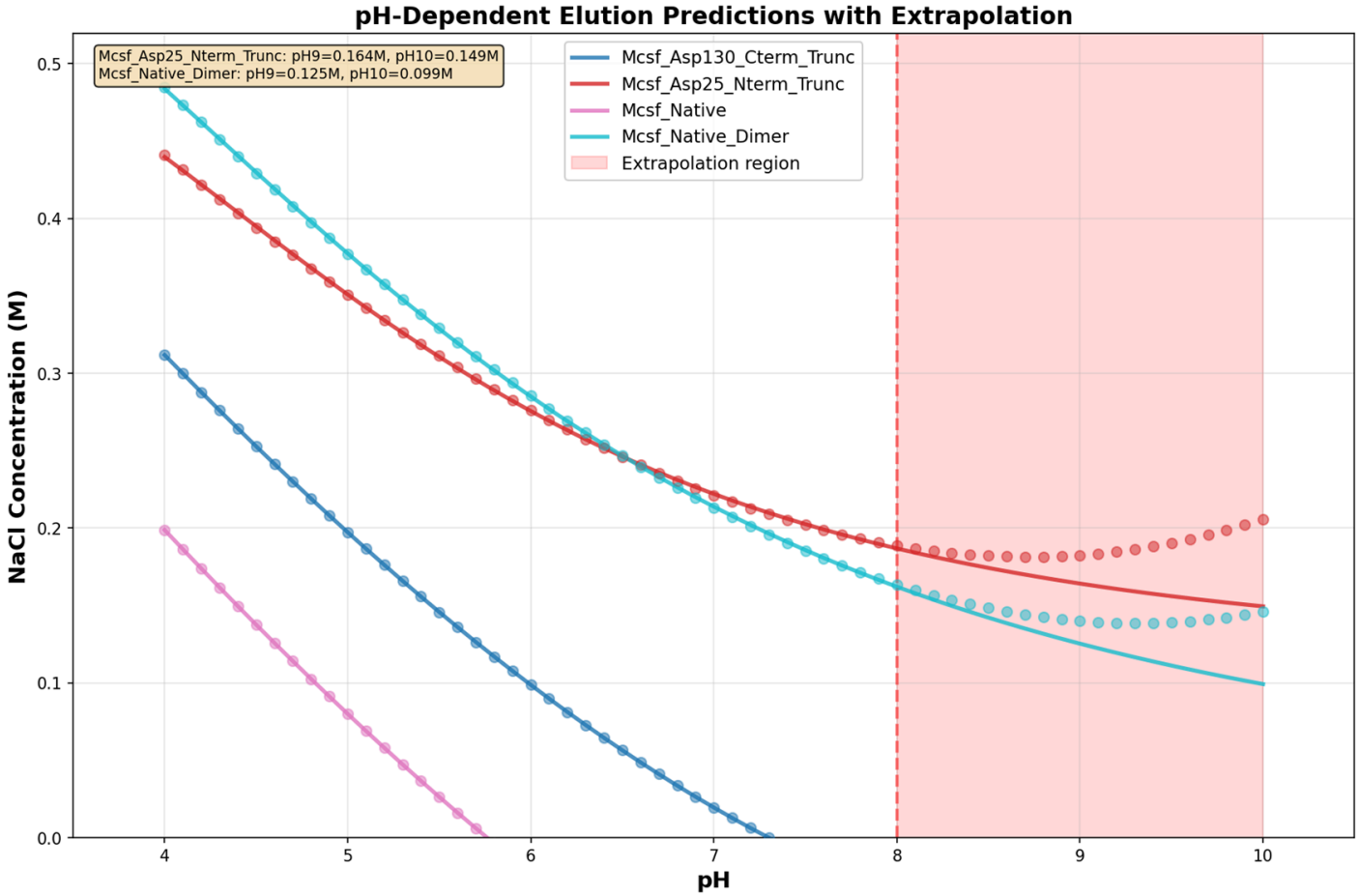
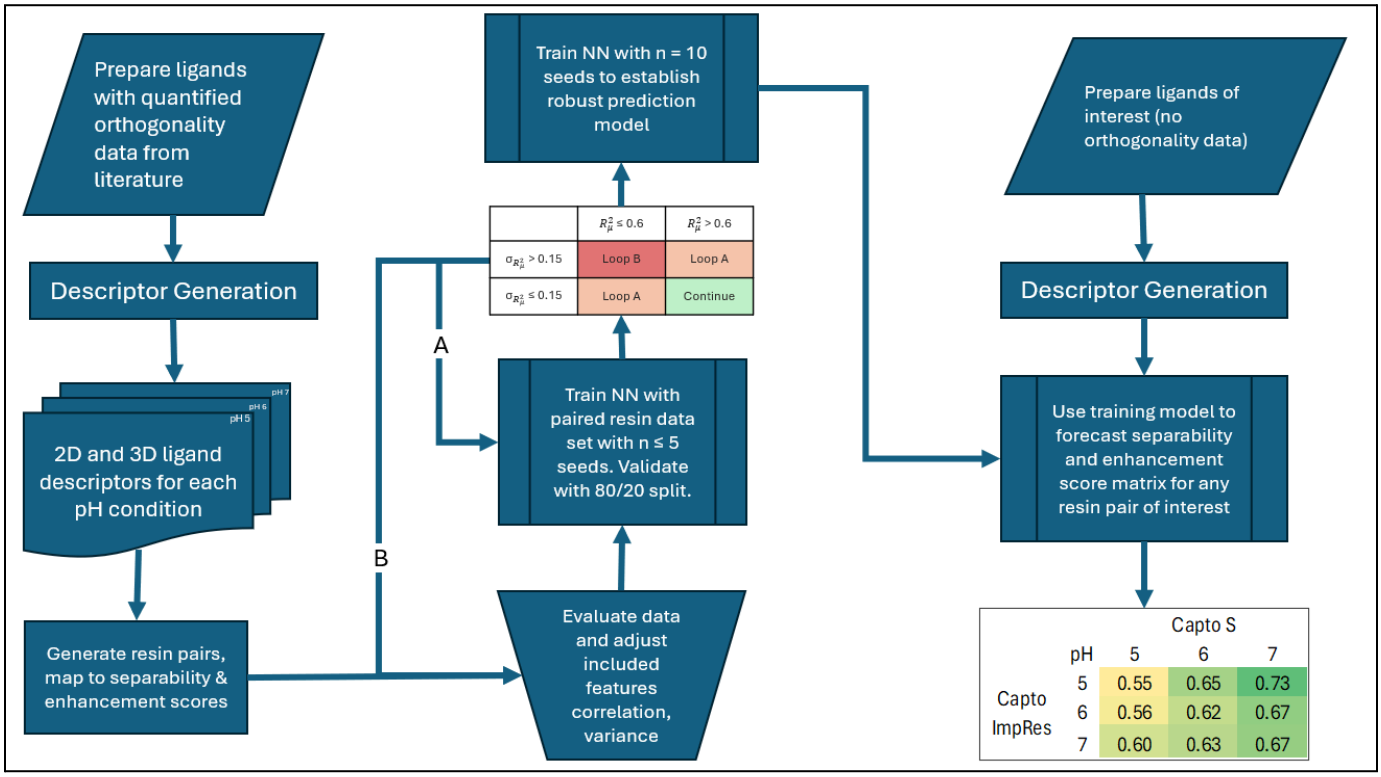


Figure 9: Example of forecasted elution curve for M-CSF and related variants. Red shaded area indicates projected elution curves extrapolated using a two-stage sigmoid curve (solid line) fit to the elution curve from pH 4 to pH 8, and predicted elution values using pH > 8 to pH 10 as a feature. Values in the red shaded area are not yet validated.

Chromortho

ChromOrtho Schematic



Goal

Inspired by published work and using data from Cramer's research group[4,5], train a machine learning algorithm that estimates the Separability Score and Enhancement factor of an unseen IEX chromatography resin pair using chemi-descriptors calculated from ligands represented in SMILES format.

Limitations

Capturing the multimodal chromatography ligand property space has proven to be quite challenging. Despite over 200+ unique resin+pH x resin+pH datapoints, the sample size is still too small to capture the complexity of CEX/AEX and IEX/MMC properties. Some ligand structures are proprietary and educated guesses had to be made based on details reported by the company (i.e. pH range, stong/weak IEX resin, tenticular ligand structure, etc.)

Results and Discussion

A number of training pool variants were made in an effort to identify the most optimal training method. First, three different SMILES pools were generated: one with the exact structure as reported by the manufacturer, a second with a normalized linker structure, and a third with just the resin functional group. Second, both pH-dependent and ph-independent descriptors were calculated for model training. Finally, homemade descriptors demonstrated by Cramer's research group[4] to accurately define the MMC ligand property space in a QSAR model were incorporated into the model and exhibited slight improvement. While many machine learning algorithms were tried in this model, a siamese neural network (SNN) proved to be the most effective. This is likely due to the inherent architecture of SNN in which model weights are shared across training towers, allowing the model to share descriptor interaction behavior while still training CEX and AEX models independently. Still, none of the models achieved $r^2 > 0.4$ when tested with leave-one-out validation (data not shown).

Model performance

Model	Test R^2	R^2 std	Validation Method	Script Name
Linear Regression	0.24	N/A	Single train/test split	compare_models_baseline.py
Ridge Regression	0.26	N/A	Single train/test split	compare_models_baseline.py
Lasso Regression	0.44	N/A	Single train/test split	compare_models_baseline.py
SVR (RBF kernel)	0.34	N/A	Single train/test split	compare_models_baseline.py
XGBoost	0.27	N/A	Single train/test split	compare_models_baseline.py
PCA + Ridge	0.15	N/A	Single train/test split	compare_models_baseline.py
k-NN Similarity	0.53	N/A	Single train/test split	compare_models_baseline.py
Siamese NN (symmetric)	N/A	N/A	Per-query inference	compare_similarity_vs_nn.py
Siamese NN (asymetric)	0.8	± 0.05	Random CV (5 seeds)	train_twohead_mtl.py (no --ab_swap)
Siamese NN (asymetric	0.8	± 0.04	Random CV (5 seeds)	train_twohead_mtl.py (with --ab_swap)

Table 5: Best performance of various machine learning models using alternative validation methods.

Works Cited

1. Holstein MA, Parimal S, McCallum SA, Cramer SM. Mobile phase modifier effects in multimodal cation exchange chromatography. Biotechnology and Bioengineering. 2011 Sep 9;109(1):176–86.
2. Yang T, Sundling MC, Freed AS, Breneman CM, Cramer SM. Prediction of pH-Dependent Chromatographic Behavior in Ion-Exchange Systems. Analytical Chemistry. 2007 Nov 3;79(23):8927–39.
3. Bilodeau CL, Vecchiarello NA, Altern S, Cramer SM. Quantifying orthogonality and separability: A method for optimizing resin selection and design. Journal of Chromatography A. 2020 Sep;1628:461429.
4. Woo JA, Chen H, Snyder MA, Chai Y, Frost RG, Cramer SM. Defining the property space for chromatographic ligands from a homologous series of mixed-mode ligands. Journal of Chromatography A. 2015 Aug;1407:58–68.