Protein Classification by Isoelectric Point Machine Learning Approach

Justyna Kowalska

Some plots are at the end of the file
(All are in results folder)

1 Introduction

Project Goal: Classification of proteins and peptides into acidic (pI ¡ 5) and non-acidic (pI ≥ 5) using machine learning methods.

Isoelectric Point (pI) is the pH value at which a molecule carries no net electrical charge or is electrically neutral on average.

Task: Building ML models for binary classification of proteins/peptides based on their amino acid sequences.

2 Data Preparation

2.1 Data Sources

- **IPC peptide**: 16,882 items
- **IPC protein**: 2,324 items
- Source: http://isoelectric.org/datasets.html

Number of proteins with pI < 5.0	674
Number of peptides with pI < 5.0	5757
Number of proteins with pI > 10.0	19
Number of peptides with pI > 10.0	3386

2.2 Preprocessing

- Labeling: pI ¡ 5 → acidic (1), pI ≥ 5 → non-acidic (0)
- Split: 60% training / 20% testing / 20% validation

Dataset	Total Items	Proteins	Peptides	Acidic
Training	1001	347 (34.67%)	654 (65.33%)	663 (66.23%)
Testing	334	119 (35.63%)	215 (64.37%)	227 (67.96%)
Validation	334	103 (30.84%)	231 (69.16%)	210 (62.87%)

Table 1: Dataset distribution across training, testing, and validation sets

3 Features

3.1 Selected Features

Feature Name	Type	Description
uid	String	MD5 hash of sequence
data_type	String	Source type (protein/peptide)
pI	Float	Isoelectric point
length	Integer	Sequence length
molecular_weight	Float	Molecular weight (Da)
label	Integer	Classification label (0/1)
aa_A, aa_C, ..., aa_Y	Integer	Count of each amino acid
hydro_total	Float	Total hydrophobicity (Kyte-Doolittle)
hydro_mean	Float	Mean hydrophobicity per residue
charge_Sillero	Float	Net charge at pH 7 (Sillero scale)
charge_EMBOSS	Float	Net charge at pH 7 (EMBOSS scale)
charge_DTASelect	Float	Net charge at pH 7 (DTASelect scale)
charge_Solomon	Float	Net charge at pH 7 (Solomon scale)
charge_Rodwell	Float	Net charge at pH 7 (Rodwell scale)

Table 2: Feature descriptions

4 Machine Learning Models

4.1 Decision Tree

4.1.1 Hyperparameter Optimization

A comprehensive grid search was performed with 5-fold cross-validation across 576 parameter combinations, totaling 2,880 model fits.

Best Parameters:

Parameter	Value
criterion	gini
max_depth	3
max_features	sqrt
min_samples_leaf	1
min_samples_split	2

Table 3: Optimal Decision Tree hyperparameters

Method	Best Parameter	CV Score	Validation Score
Depth Optimization	Depth: 1	0.997	0.997
Feature Selection	Features:1	0.997	0.997
Grid Search	Multiple params	0.997	1.000

Table 4: Decision Tree optimization methods comparison

4.1.2 Performance Results

4.1.3 Feature Importance Analysis

The Decision Tree modelidentified the most discriminative features for acidic protein classification:

Rank	Feature	Importance
1	charge Rodwell	0.9898
2	aa_R	0.0087
3	aa_N	0.0015
4-15	Other features	0.0000

Table 5: Top features by importance in Decision Tree model

The model demonstrates that **charge Rodwell** is by far the most important feature, accounting for 98.98% of the decision-making process, followed by arginine content (aa R) and asparagine content (aa N) with minimal contributions.

4.1.4 Final Model Performance

Test Set Results:

- **Test Accuracy:** 99.4%
- **Cross-validation Score:** 99.7%
- **Validation Score:** 100.0%

Class	Precision	Recall	F1-score
Non-acidic	0.98	1.00	0.99
Acidic	1.00	0.99	1.00
Weighted Avg	**0.99**	**0.99**	**0.99**

Table 6: Decision Tree classification report on test set

The Decision Tree model achieved exceptionally high performance metrics,what seems unusual to me.

4.2 Random Forest

The Random Forest model was optimized through systematic testing of multiple parameters, including number of estimators and comprehensive grid search across 648 parameter combinations (3,240 total fits with 5-fold CV).

Number of Trees Optimization:

Trees	Train Acc	Val Acc	CV Score
10	1.000	1.000	0.997
25	1.000	1.000	0.997
50	1.000	1.000	0.997
100	1.000	1.000	0.997
200	1.000	1.000	0.997
500	1.000	1.000	0.997

Table 7: Random Forest n estimators optimization

Best Parameters:

Parameter	Value
n_estimators	100
max_depth	5
max_features	sqrt
min_samples split	2
min_samples leaf	1
bootstrap	True

Table 8: Optimal Random Forest hyperparameters

4.2.1 Feature Importance Analysis

Unlike Decision Tree's single dominant feature, Random Forest shows more distributed importance:

Rank	Feature	Importance
1	charge Rodwell	0.226
2	charge Solomon	0.201
3	charge EMBOSS	0.183
4	charge Sillero	0.158
5	charge DTASelect	0.103
6	aa_E	0.036
7	length	0.026
8	aa_D	0.018

Table 9: Top Random Forest feature importance

4.2.2 Ensemble Analysis

Tree Diversity Statistics:

- Average tree depth: 2.60 ± 1.02
- Average nodes per tree: 6 ± 2
- Feature importance correlation: 0.069
- Best individual tree accuracy: 100.0%
- Worst individual tree accuracy: 94.9%

4.2.3 Final Performance

Test Set Results:

- **Test Accuracy:** 100.0%
- **Cross-validation Score:** 99.7%
- **Validation Score:** 100.0%

Class	Precision	Recall	F1-score
Non-acidic	1.00	1.00	1.00
Acidic	1.00	1.00	1.00
Weighted Avg	**1.00**	**1.00**	**1.00**

Table 10: Random Forest classification report

Random Forest achieved perfect test accuracy (100%). Suspiciously perfect performance continues to suggest potential data

4.3 K-Nearest Neighbors (KNN)

The KNN model was analyzed through systematic testing of the number of neighbors (k), distance metrics, and feature selection methods. Data was standardized for optimal KNN performance.

Metric	CV Score
Euclidean	0.993
Manhattan	0.995
Chebyshev	0.957
Minkowski	0.995

Table 11: KNN distance metric comparison

Method	Optimal Features	CV Score
SelectKBest	12 features	1.000
RFE	8 features	0.999
Baseline (all)	30 features	0.995

Table 12: Feature selection comparison for KNN

4.3.1 Best parameters

- algorithm=auto,
- metric=manhattan,
- n neighbors=7,
- weights=distance
- **Cross-validation Score:** 0.994
- **Validation Accuracy:** 0.997

4.3.2 Performance Summary

Method	CV Score	Validation Score
K Optimization (k=1)	0.993	1.000
Distance Metrics (Manhattan)	0.995	-
Feature Selection (12 features)	1.000	-
Grid Search (Final)	0.994	0.997

Table 13: KNN optimization methods summary

4.3.3 Final Model Evaluation

Class	Precision	Recall	F1-score
Non-acidic	1.00	0.99	1.00
Acidic	1.00	1.00	1.00
Weighted Avg	**1.00**	**1.00**	**1.00**

Table 14: KNN classification report on test set

Continues the pattern of suspiciously high performance across all models.

4.4 Support Vector Machine (SVM)

The SVM model was optimized through systematic testing of regularization parameter (C), gamma parameter, and kernel types. Data was standardized for optimal SVM performance.

4.4.1 Best parameters

- C=10
- gamma='scale'
- kernel='rbf'
- **Cross-validation Score:** 1.000
- **Validation Accuracy:** 1.000

4.4.2 Feature Importance Analysis (Linear SVM)

Rank	Feature	Importance
1	charge DTASelect	2.027
2	charge EMBOSS	2.024
3	charge Sillero	2.022
4	charge Rodwell	1.993
5	charge Solomon	1.992
6	aa_H	0.796
7	aa_K	0.557
8	aa_C	0.505

Table 15: Top SVM feature importance (Linear kernel)

4.4.3 Performance Summary

Method	CV Score	Validation Score
C Optimization (C=10)	1.000	-
Gamma Optimization (0.01)	0.999	-
Kernel Comparison (RBF)	1.000	-
Grid Search (Final)	1.000	1.000

Table 16: SVM optimization methods summary

4.4.4 Final Model Evaluation

Class	Precision	Recall	F1-score
Non-acidic	1.00	1.00	1.00
Acidic	1.00	1.00	1.00
Weighted Avg	**1.00**	**1.00**	**1.00**

Table 17: SVM classification report on test set

Perfect performance continues the concerning pattern across all algorithms

4.5 Deep Learning Models

4.5.1 Model Architectures and Training Results

Ten different deep learning architectures were trained and evaluated on the dataset using CPU processing. All models achieved perfect accuracy (1.0000) on the training data, demonstrating the dataset's high separability.

Architecture	Layers	Parameters	Accuracy	Activation	BatchNorm	Dropout
Deep	5	51,905	1.0000	relu	True	0.3
Very_Deep	6	191,937	1.0000	relu	True	0.4
Wide	3	148,481	1.0000	relu	True	0.3
ELU_Deep	5	51,905	1.0000	elu	True	0.3
Shallow	3	4,225	1.0000	relu	True	0.2
Medium	4	14,657	1.0000	relu	True	0.3
GELU_Medium	4	14,657	1.0000	gelu	True	0.2
No_BatchNorm	4	14,209	1.0000	relu	False	0.4
High_Dropout	4	49,793	1.0000	relu	True	0.5
Ultra_Deep	8	139,265	1.0000	relu	True	0.3

Table 18: Comparison of deep learning architectures trained on the dataset

4.5.2 Architecture Analysis

Model Complexity Range:

- **Simplest:** Shallow (3 layers, 4,225 parameters)
- **Most Complex:** Very_Deep (6 layers, 191,937 parameters)
- **Deepest:** Ultra_Deep (8 layers, 139,265 parameters)

Activation Function Testing:

- **ReLU:** 7 models (standard choice)
- **ELU:** 1 model (ELU_Deep)
- **GELU:** 1 model (GELU_Medium)

Regularization Techniques:

- **Batch Normalization:** Used in 9/10 models
- **Dropout rates:** Ranging from 0.2 to 0.5
- **No BatchNorm model:** Tested impact of batch normalization

4.5.3 Best Model Selection

The **Deep** architecture was selected as the best model based on its balanced complexity and performance:

Architecture Details:

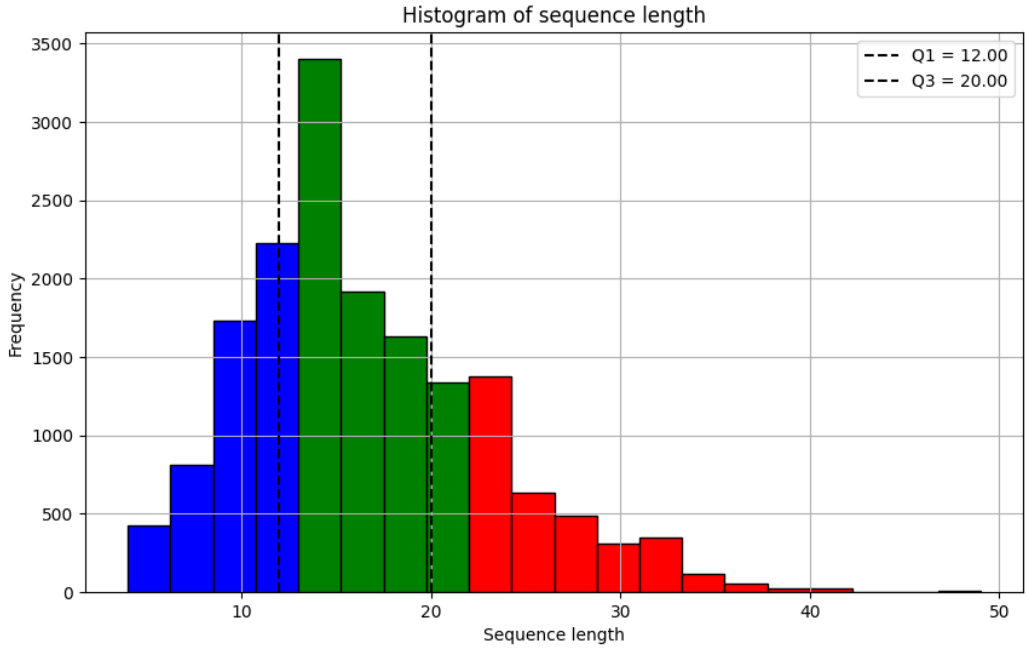
- **Configuration:** *{*'hidden dims': [256, 128, 64, 32], 'activation': 'relu', 'use batch norm': True, 'dropout rate': 0.3*}*
- **Parameters:** 51,905
- **Layers:** 5 (input + 4 hidden + output)
- **Final Accuracy:** 1.0000

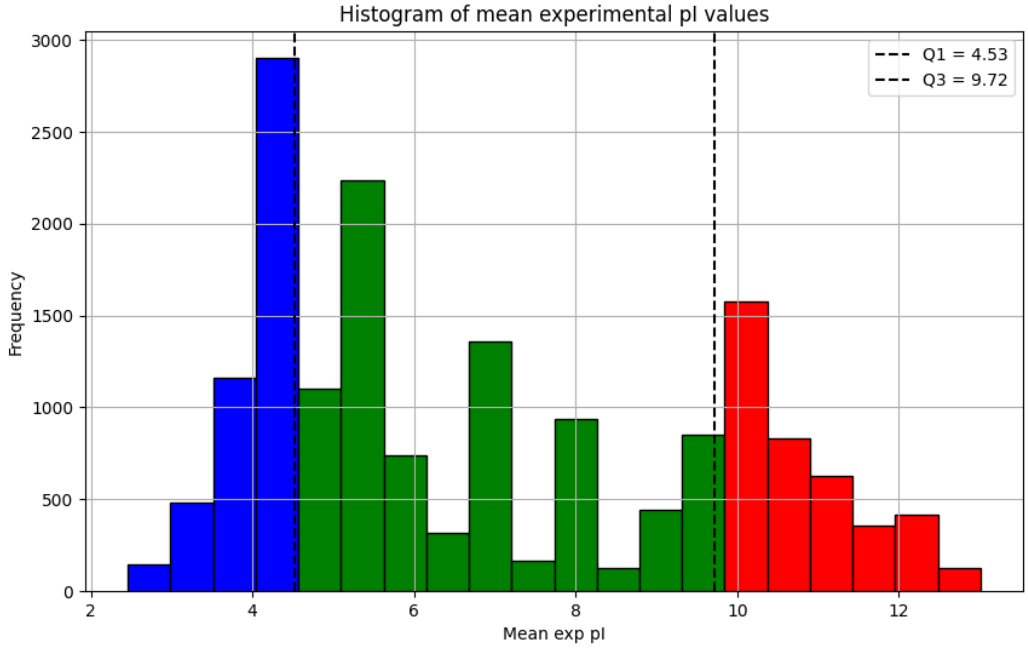
Perfect Performance Concerns:

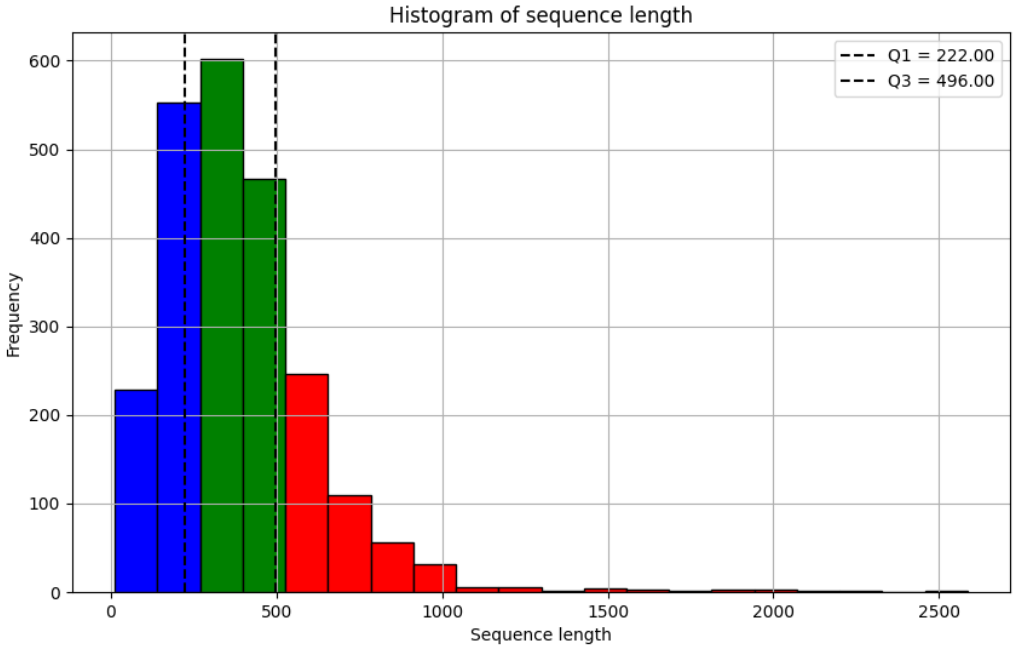
- All 10 architectures achieved identical perfect accuracy (1.0000)
- Performance independence from model complexity suggests dataset oversimplification
- Even the simplest Shallow model (4,225 parameters) achieved perfect results
- No BatchNorm model performed equally well, indicating strong feature separability

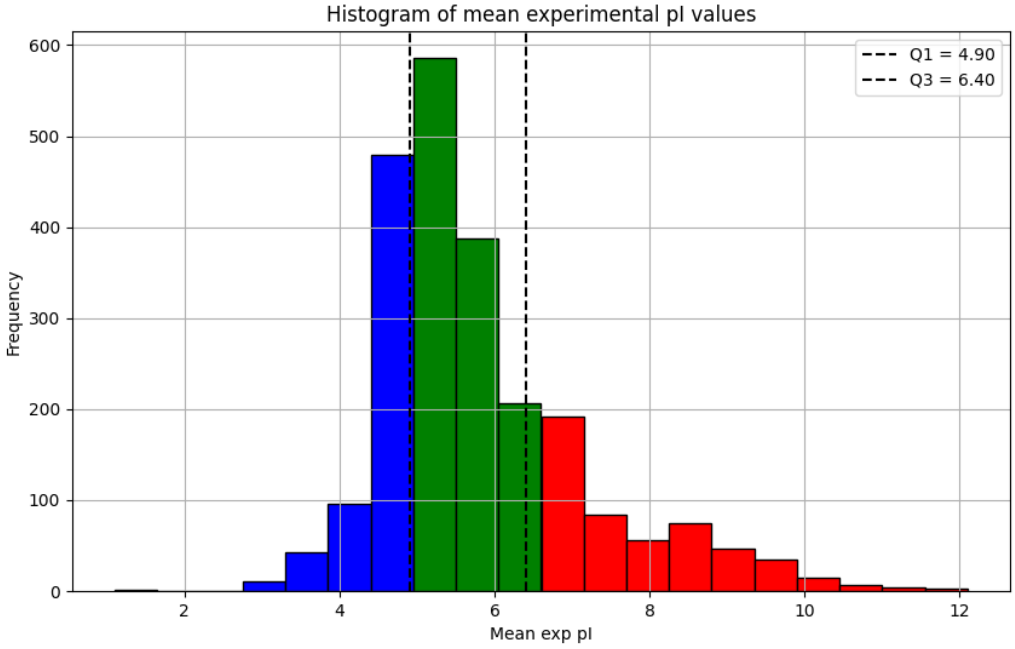
Model Robustness:

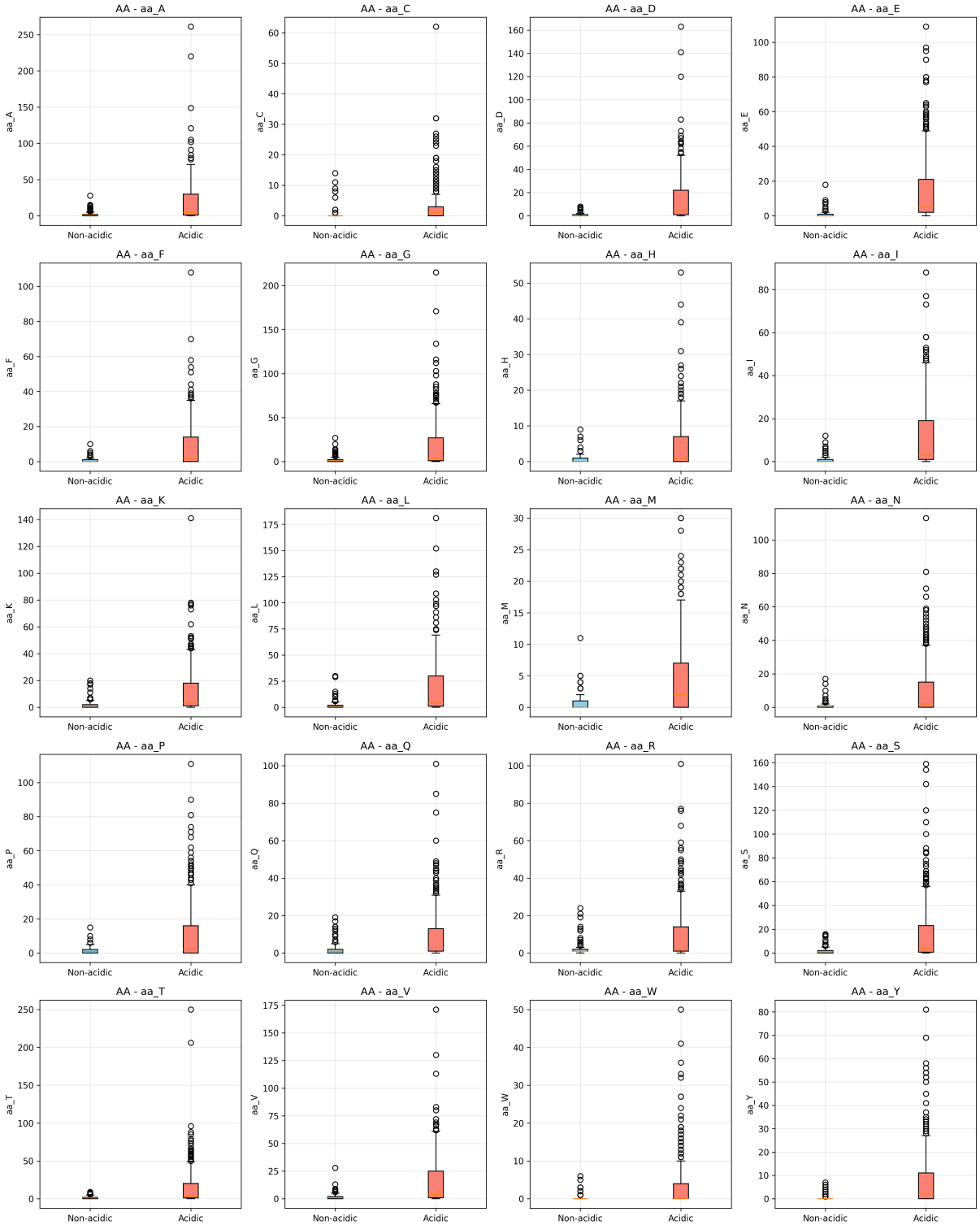
- Different activation functions (ReLU, ELU, GELU) showed no performance difference
- Varying dropout rates (0.2-0.5) had no impact on final accuracy
- Architecture depth (3-8 layers) did not affect classification performance

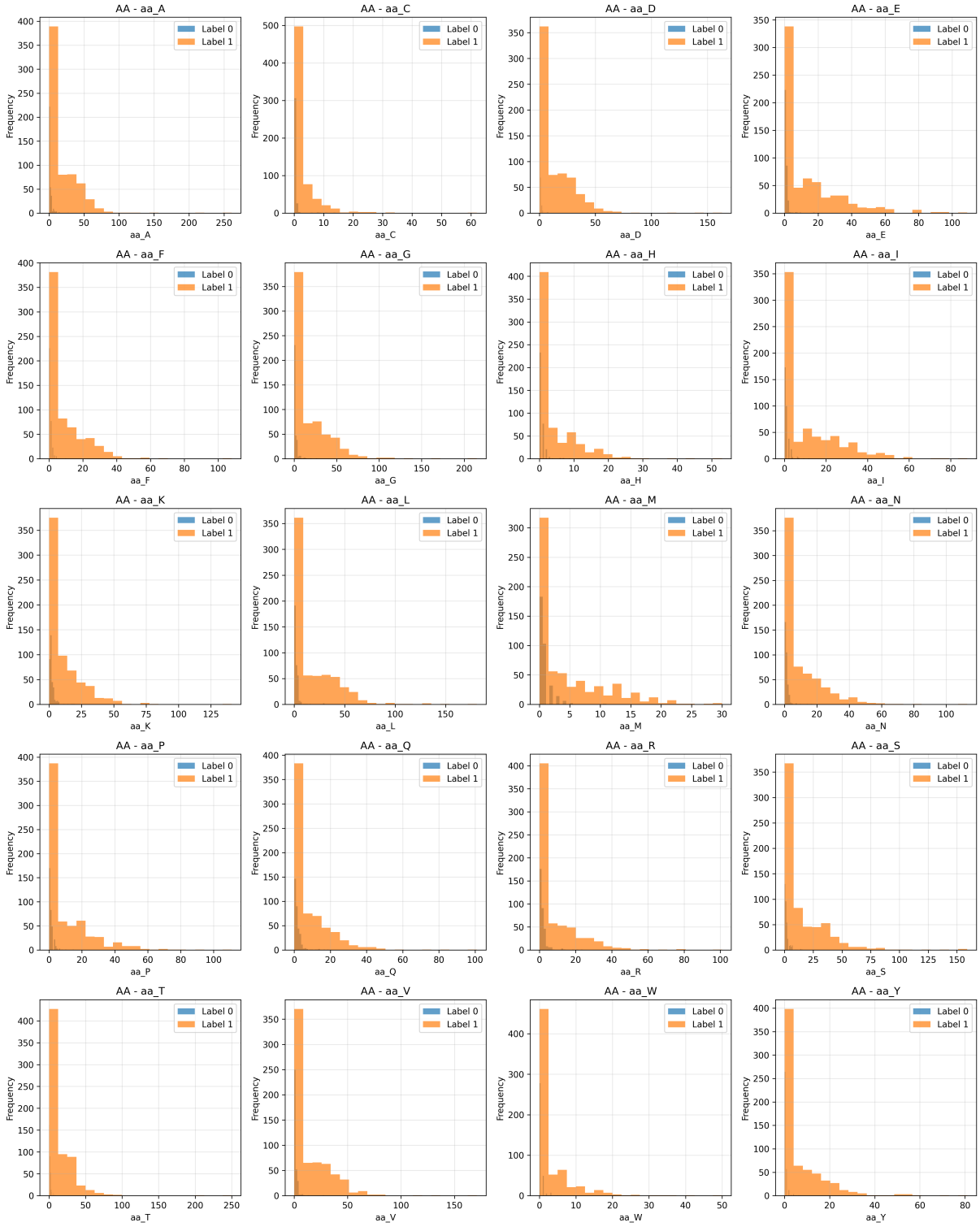

Histogram of sequence length
Q1 = 12.00
Q3 = 20.00
Frequency
Sequence length
0
500
1000
1500
2000
2500
3000
3500
10
20
30
40
50

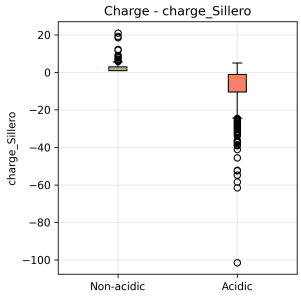

Histogram of mean experimental pI values
Q1 = 4.53
Q3 = 9.72
Frequency
Mean exp pI
0
500
1000
1500
2000
2500
3000
2
4
6
8
10
12

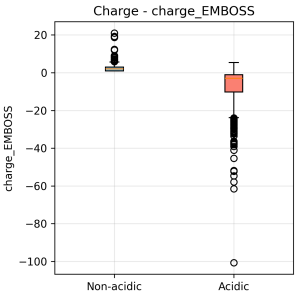

Histogram of sequence length
Q1 = 222.00
Q3 = 496.00
Frequency
Sequence length
0
100
200
300
400
500
600
0
500
1000
1500
2000
2500

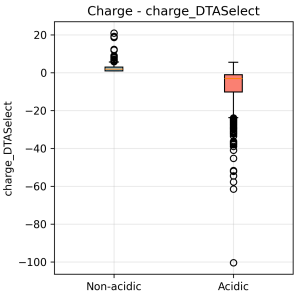

Histogram of mean experimental pI values
Q1 = 4.90
Q3 = 6.40
Frequency
Mean exp pI
0
100
200
300
400
500
600
2
4
6
8
10
12

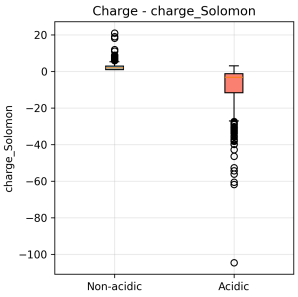

AA - aa_A
aa_A
AA - aa_C
aa_C
AA - aa_D
aa_D
AA - aa_E
aa_E
AA - aa_F
aa_F
AA - aa_G
aa_G
AA - aa_H
aa_H
AA - aa_I
aa_I
AA - aa_K
aa_K
AA - aa_L
aa_L
AA - aa_M
aa_M
AA - aa_N
aa_N
AA - aa_P
aa_P
AA - aa_Q
aa_Q
AA - aa_R
aa_R
AA - aa_S
aa_S
AA - aa_T
aa_T
AA - aa_V
aa_V
AA - aa_W
aa_W
AA - aa_Y
aa_Y
Non-acidic
Acidic

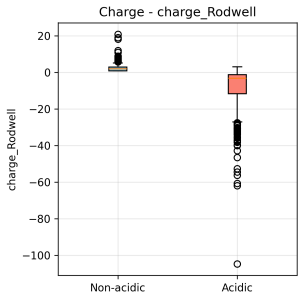

AA - aa_A
AA - aa_C
AA - aa_D
AA - aa_E
AA - aa_F
AA - aa_G
AA - aa_H
AA - aa_I
AA - aa_K
AA - aa_L
AA - aa_M
AA - aa_N
AA - aa_P
AA - aa_Q
AA - aa_R
AA - aa_S
AA - aa_T
AA - aa_V
AA - aa_W
AA - aa_Y
Label 0
Label 1
Frequency

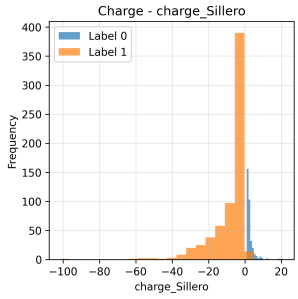

Charge - charge_Sillero
charge_Sillero
Non-acidic
Acidic

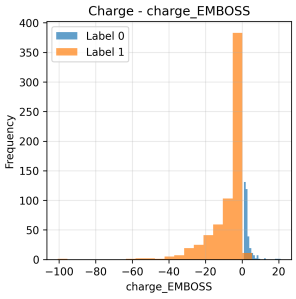

Charge - charge_EMBOSS
charge_EMBOSS
Non-acidic
Acidic

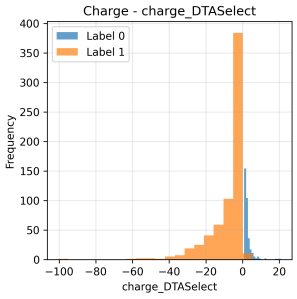

Charge - charge_DTASelect
charge_DTASelect
Non-acidic
Acidic

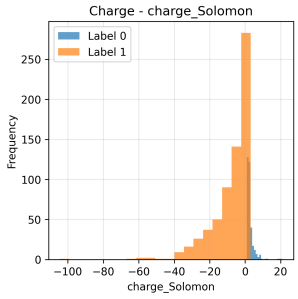

Charge - charge_Solomon
charge_Solomon
Non-acidic
Acidic

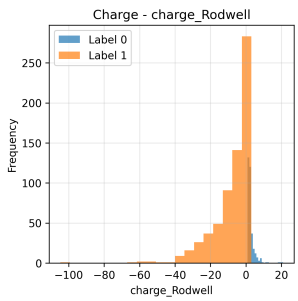

Charge - charge_Rodwell
charge_Rodwell
Non-acidic
Acidic

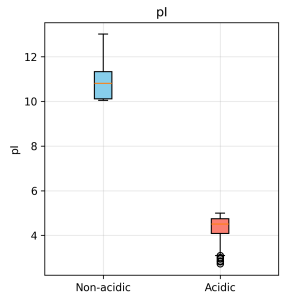

Charge - charge_Sillero
Label 0
Label 1
Frequency
charge_Sillero

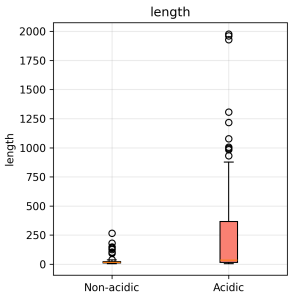

Charge - charge_EMBOSS
Label 0
Label 1
Frequency
charge_EMBOSS

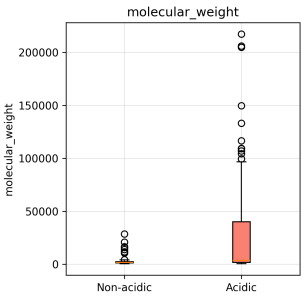

Charge - charge_DTASelect
Label 0
Label 1
Frequency
charge_DTASelect

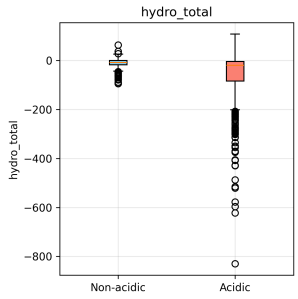

Charge - charge_Solomon
Label 0
Label 1
Frequency
charge_Solomon

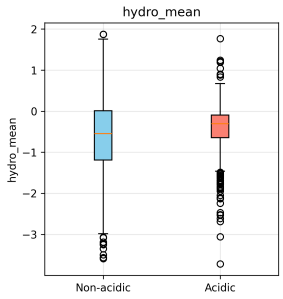

Charge - charge_Rodwell
Label 0
Label 1
Frequency
charge_Rodwell

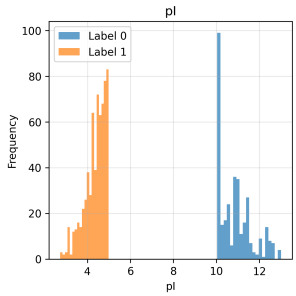

pI
pI
Non-acidic
Acidic

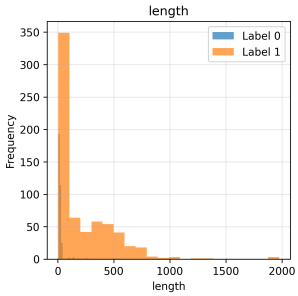

length
length
Non-acidic
Acidic

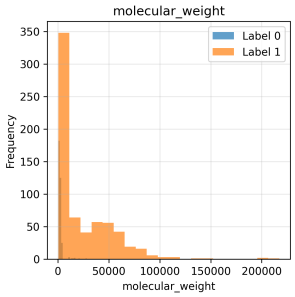

molecular_weight
molecular_weight
Non-acidic
Acidic

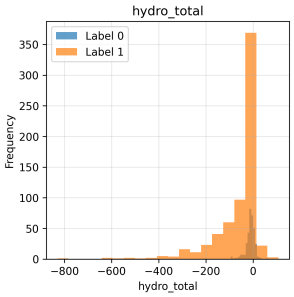

hydro_total
hydro_total
Non-acidic
Acidic

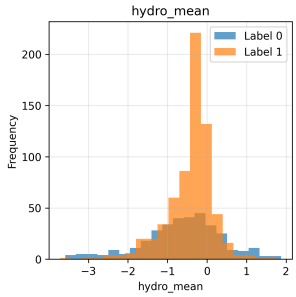

hydro_mean
hydro_mean
Non-acidic
Acidic

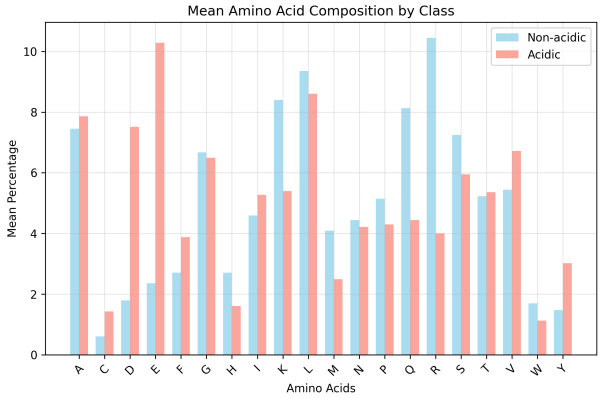

pI
Label 0
Label 1
Frequency
pI

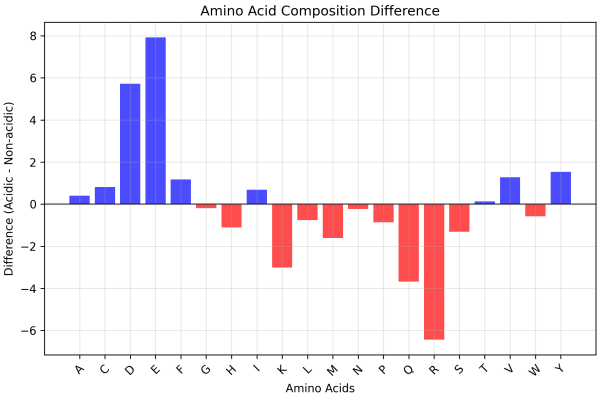

length
Label 0
Label 1
Frequency
length

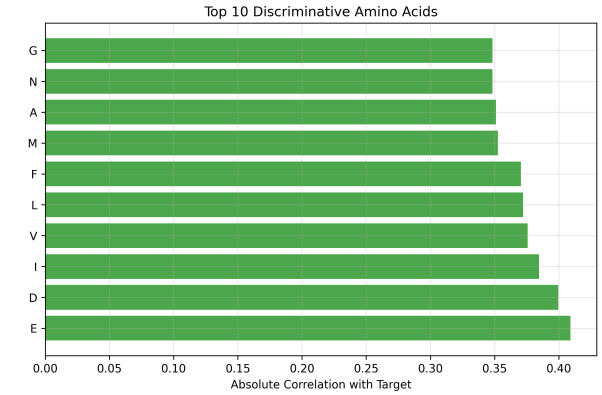

molecular_weight
Label 0
Label 1
Frequency
molecular_weight

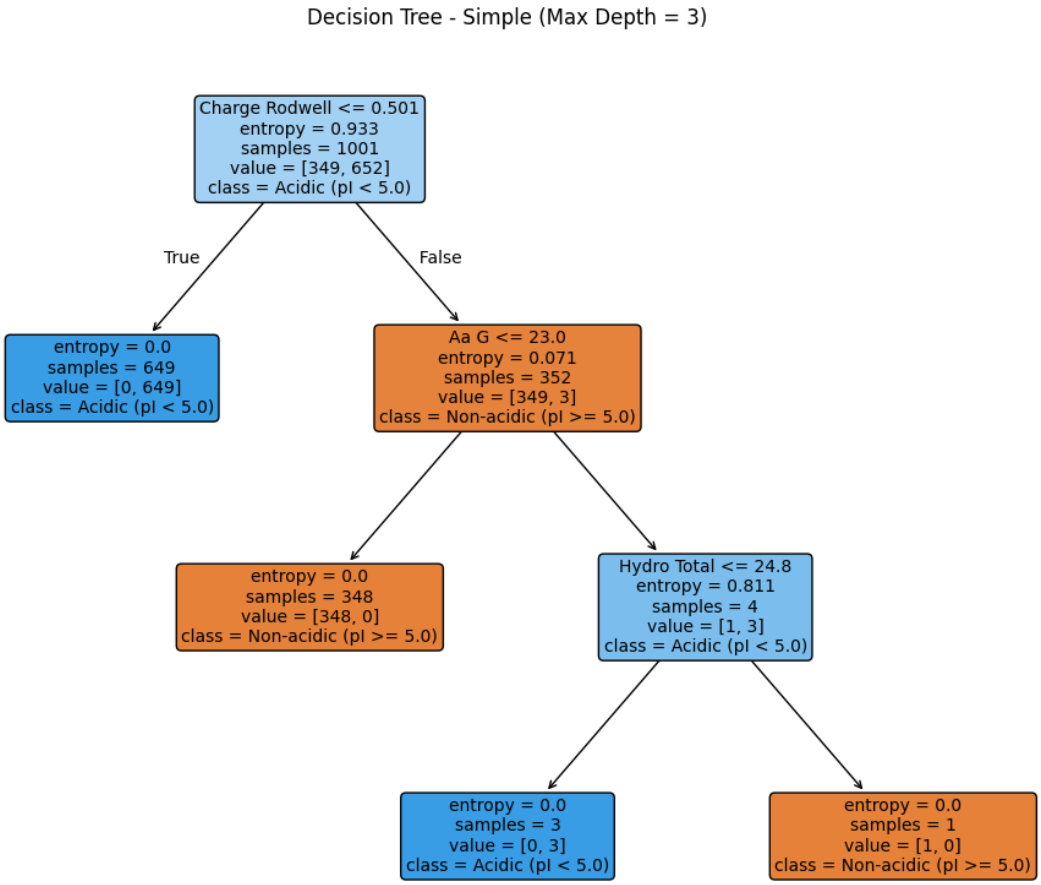

hydro_total
Label 0
Label 1
Frequency
hydro_total

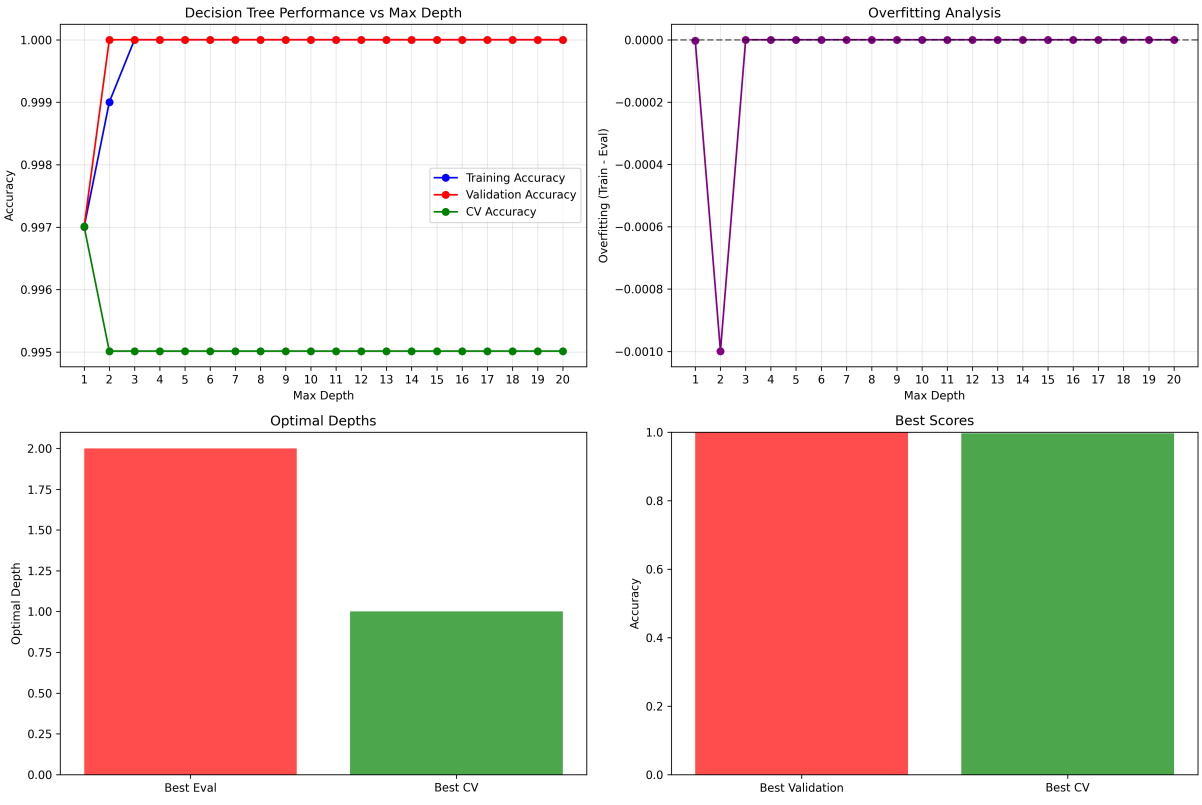

hydro_mean
Label 0
Label 1
Frequency
hydro_mean

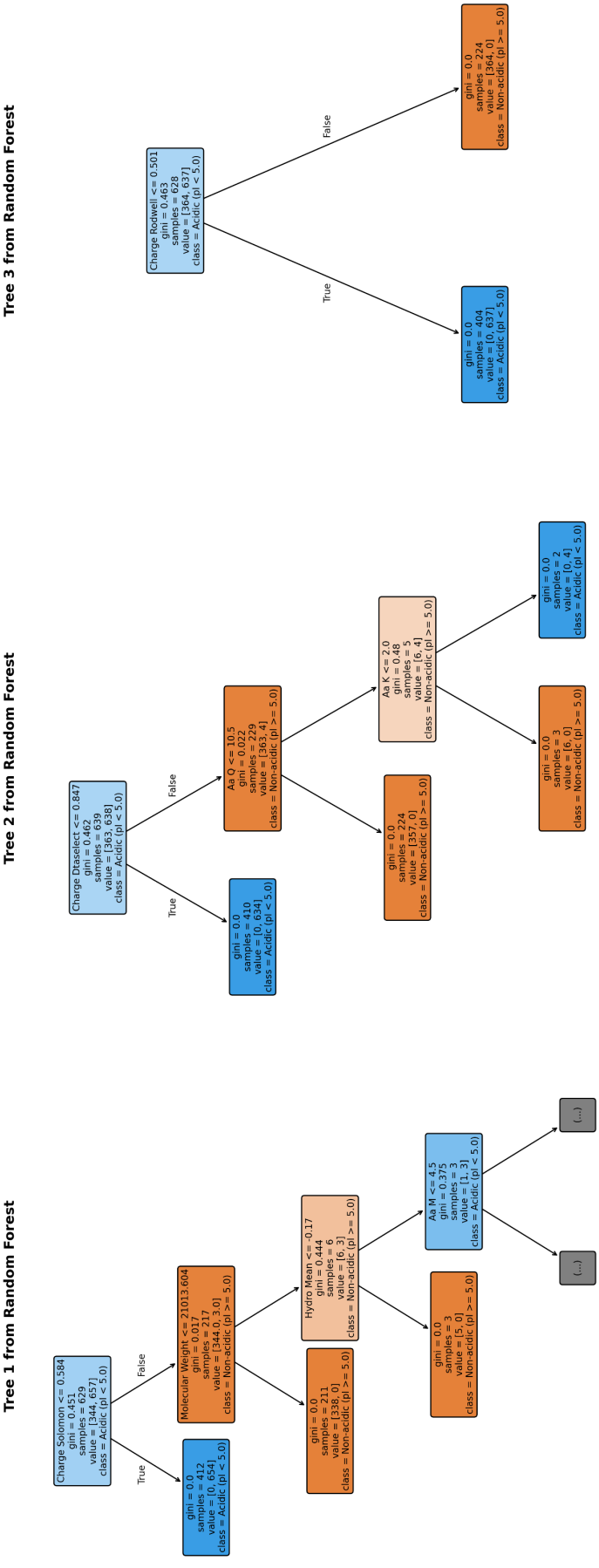

Mean Amino Acid Composition by Class
Non-acidic
Acidic
Mean Percentage
A C D E F G H I K L M N P Q R S T V W Y
Amino Acids

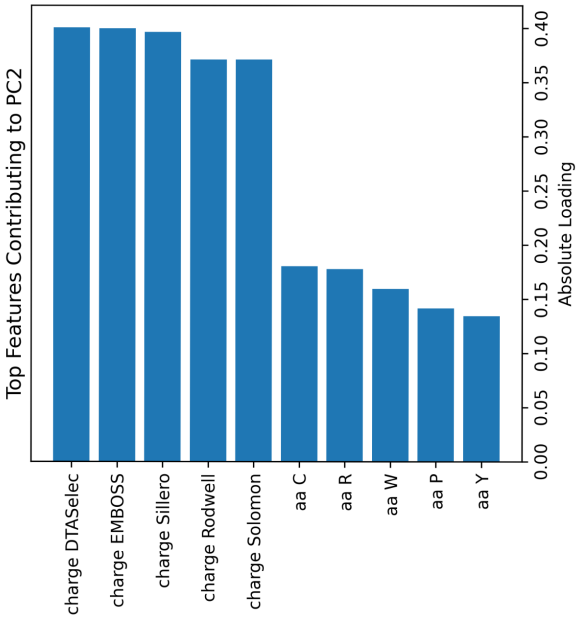

Amino Acid Composition Difference
Difference (Acidic - Non-acidic)
A C D E F G H I K L M N P Q R S T V W Y
Amino Acids

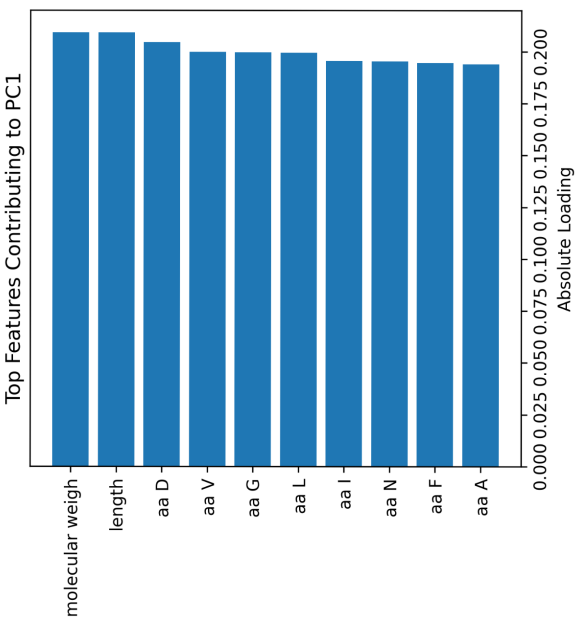

Top 10 Discriminative Amino Acids
G
N
A
M
F
L
V
I
D
E
Absolute Correlation with Target

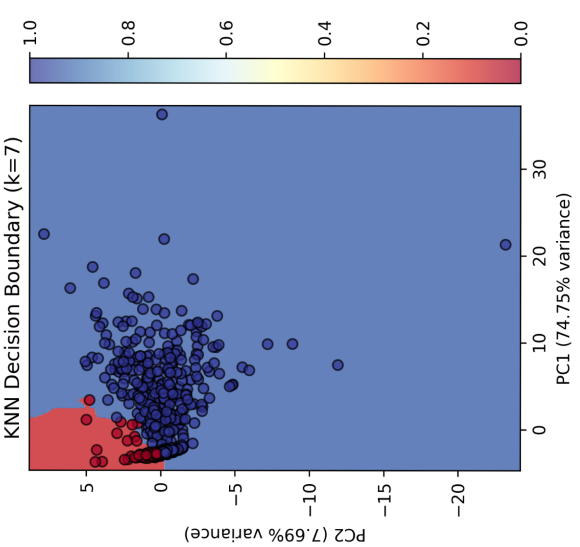

Decision Tree - Simple (Max Depth = 3)
Charge Rodwell <= 0.501
entropy = 0.933
samples = 1001
value = [349, 652]
class = Acidic (pI < 5.0)
True
False
entropy = 0.0
samples = 649
value = [0, 649]
class = Acidic (pI < 5.0)
Aa G <= 23.0
entropy = 0.071
samples = 352
value = [349, 3]
class = Non-acidic (pI >= 5.0)
entropy = 0.0
samples = 348
value = [348, 0]
class = Non-acidic (pI >= 5.0)
Hydro Total <= 24.8
entropy = 0.811
samples = 4
value = [1, 3]
class = Acidic (pI < 5.0)
entropy = 0.0
samples = 3
value = [0, 3]
class = Acidic (pI < 5.0)
entropy = 0.0
samples = 1
value = [1, 0]
class = Non-acidic (pI >= 5.0)

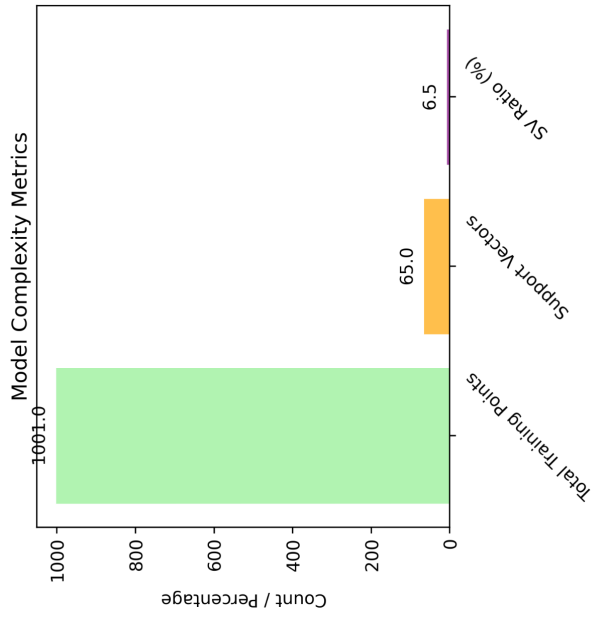

Decision Tree Performance vs Max Depth
Accuracy
Max Depth
Training Accuracy
Validation Accuracy
CV Accuracy
Overfitting Analysis
Overfitting (Train - Eval)
Max Depth
Optimal Depths
Optimal Depth
Best Eval
Best CV
Best Scores
Accuracy
Best Validation
Best CV

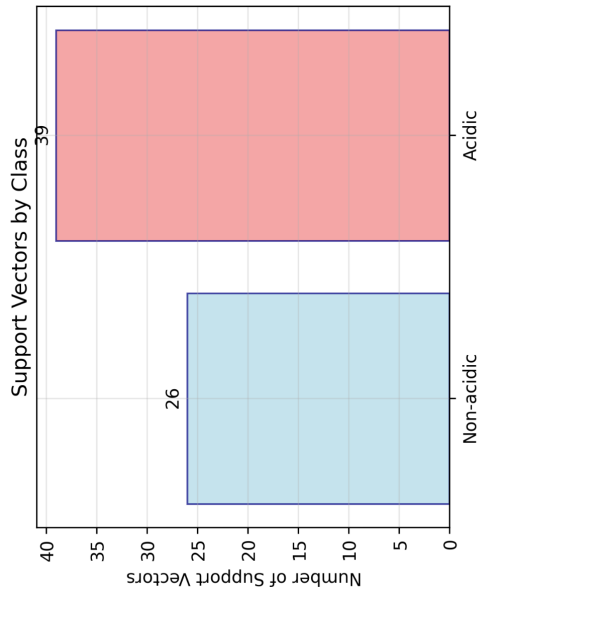

Tree 1 from Random Forest
Charge Solomon <= 0.584
gini = 0.451
samples = 629
value = [344, 657]
class = Acidic (pI < 5.0)
True
False
gini = 0.0
samples = 412
value = [0, 654]
class = Acidic (pI < 5.0)
Molecular Weight <= 21013.604
gini = 0.017
samples = 217
value = [344.0, 3.0]
class = Non-acidic (pI >= 5.0)
gini = 0.0
samples = 211
value = [338, 0]
class = Non-acidic (pI >= 5.0)
Hydro Mean <= -0.17
gini = 0.444
samples = 6
value = [6, 3]
class = Non-acidic (pI >= 5.0)
gini = 0.0
samples = 3
value = [5, 0]
class = Non-acidic (pI >= 5.0)
Aa M <= 4.5
gini = 0.375
samples = 3
value = [1, 3]
class = Acidic (pI < 5.0)
(...)
(...)
Tree 2 from Random Forest
Charge Diaselect <= 0.847
gini = 0.462
samples = 639
value = [363, 638]
class = Acidic (pI < 5.0)
True
False
gini = 0.0
samples = 410
value = [0, 634]
class = Acidic (pI < 5.0)
Aa Q <= 10.5
gini = 0.022
samples = 229
value = [363, 4]
class = Non-acidic (pI >= 5.0)
gini = 0.0
samples = 224
value = [357, 0]
class = Non-acidic (pI >= 5.0)
Aa K <= 2.0
gini = 0.48
samples = 5
value = [6, 4]
class = Non-acidic (pI >= 5.0)
gini = 0.0
samples = 3
value = [6, 0]
class = Non-acidic (pI >= 5.0)
gini = 0.0
samples = 2
value = [0, 4]
class = Acidic (pI < 5.0)
Tree 3 from Random Forest
Charge Rodwell <= 0.501
gini = 0.463
samples = 628
value = [364, 637]
class = Acidic (pI < 5.0)
True
False
gini = 0.0
samples = 404
value = [0, 637]
class = Acidic (pI < 5.0)
gini = 0.0
samples = 224
value = [364, 0]
class = Non-acidic (pI >= 5.0)

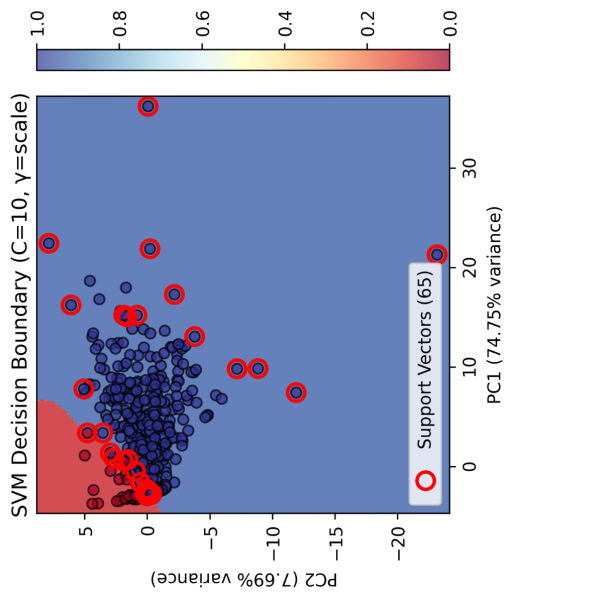

Top Features Contributing to PC2
charge DTASelec
charge EMBOSS
charge Sillero
charge Rodwell
charge Solomon
aa C
aa R
aa W
aa P
aa Y
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Absolute Loading

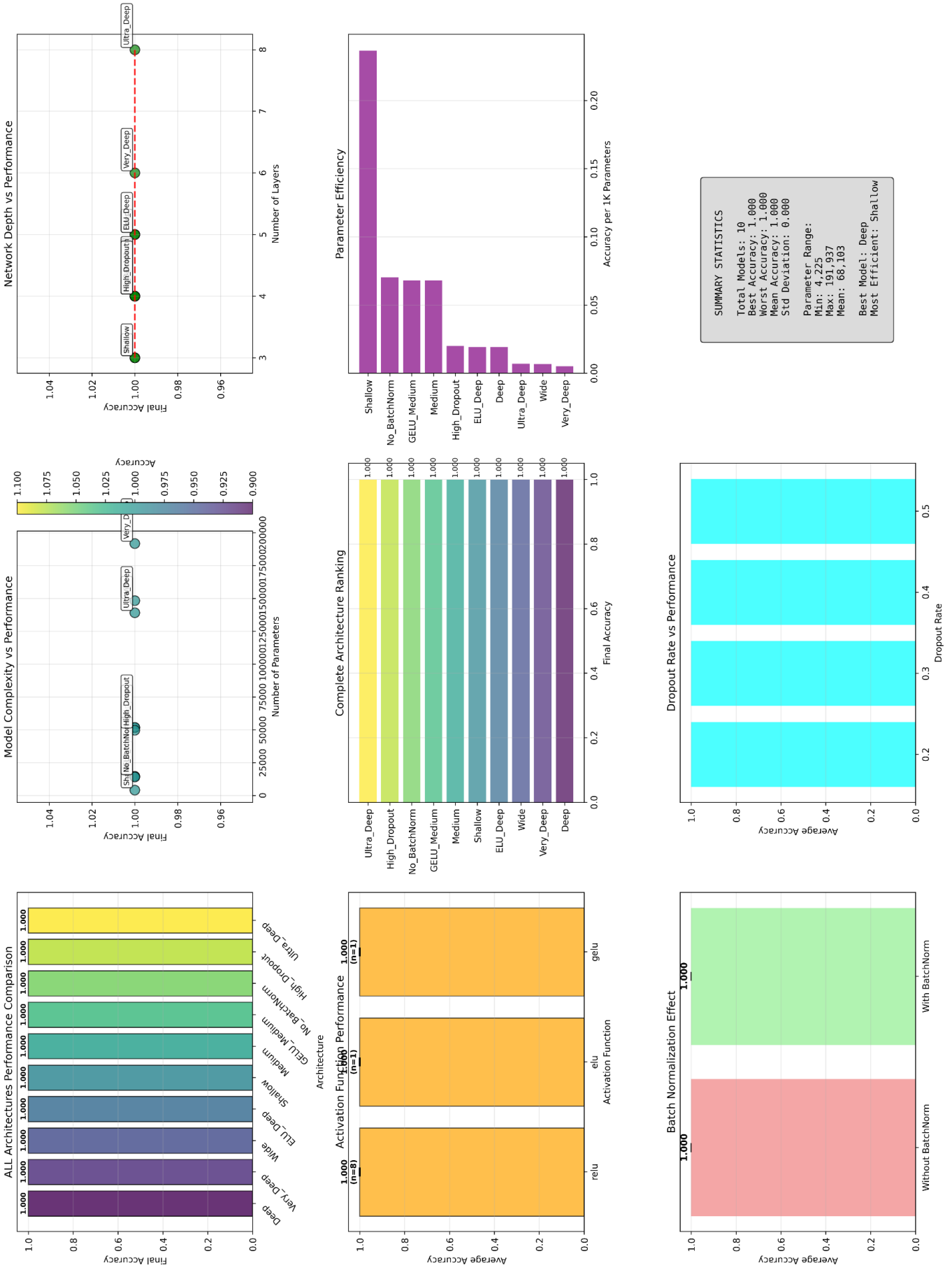

Top Features Contributing to PC1
molecular weigh
length
aa D
aa V
aa G
aa L
aa I
aa N
aa F
aa A
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Absolute Loading

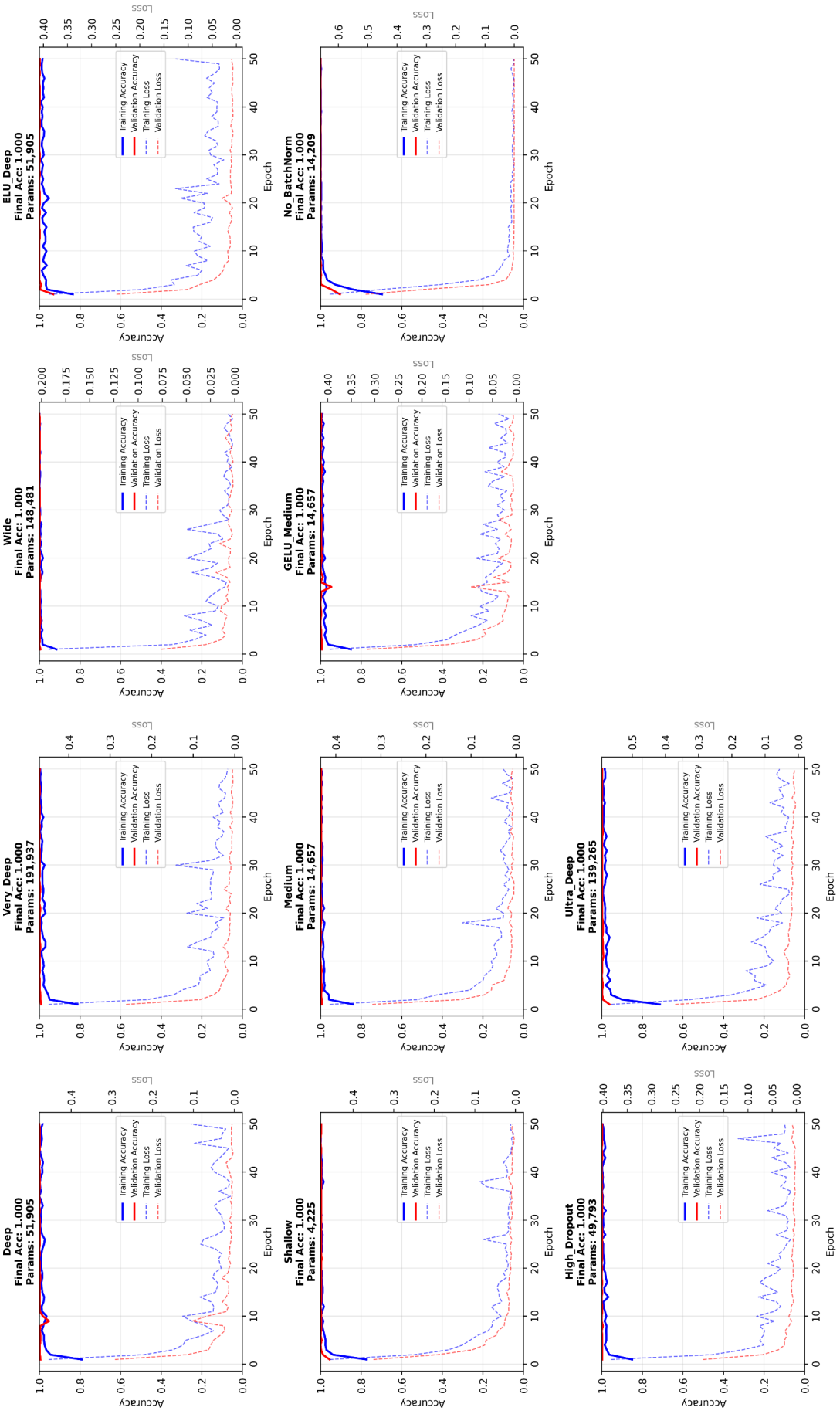

KNN Decision Boundary (k=7)
PC2 (7.69% variance)
PC1 (74.75% variance)
1.0 0.8 0.6 0.4 0.2 0.0


Model Complexity Metrics
1001.0
65.0
6.5
Total Training Points
Support Vectors
SV Ratio (%)
Count / Percentage
0 200 400 600 800 1000


Support Vectors by Class
39
26
Acidic
Non-acidic
Number of Support Vectors
0 5 10 15 20 25 30 35 40


SVM Decision Boundary (C=10, γ=scale)
Support Vectors (65)
PC2 (7.69% variance)
PC1 (74.75% variance)
1.0 0.8 0.6 0.4 0.2 0.0


Network Depth vs Performance
Final Accuracy
Number of Layers
Shallow
High_Dropout
ELU_Deep
Very_Deep
Ultra_Deep
Parameter Efficiency
Accuracy per 1K Parameters
Shallow
No_BatchNorm
GELU_Medium
Medium
High_Dropout
ELU_Deep
Deep
Ultra_Deep
Wide
Very_Deep
SUMMARY STATISTICS
Total Models: 10
Best Accuracy: 1.000
Worst Accuracy: 1.000
Mean Accuracy: 1.000
Std Deviation: 0.000
Parameter Range:
Min: 4,225
Max: 191,937
Mean: 68,103
Best Model: Deep
Most Efficient: Shallow
Model Complexity vs Performance
Final Accuracy
Number of Parameters
Complete Architecture Ranking
Final Accuracy
Ultra_Deep
High_Dropout
No_BatchNorm
GELU_Medium
Medium
Shallow
ELU_Deep
Wide
Very_Deep
Deep
Dropout Rate vs Performance
Average Accuracy
Dropout Rate
ALL Architectures Performance Comparison
Final Accuracy
Architecture
Deep
Very_Deep
Wide
ELU_Deep
Shallow
Medium
GELU_Medium
No_BatchNorm
High_Dropout
Ultra_Deep
Activation Function Performance
Average Accuracy
Activation Function
relu
elu
gelu
Batch Normalization Effect
Average Accuracy
Without BatchNorm
With BatchNorm


ELU_Deep
Final Acc: 1.000
Params: 51,905
No_BatchNorm
Final Acc: 1.000
Params: 14,209
Wide
Final Acc: 1.000
Params: 148,481
GELU_Medium
Final Acc: 1.000
Params: 14,657
Very_Deep
Final Acc: 1.000
Params: 191,937
Medium
Final Acc: 1.000
Params: 14,657
Ultra_Deep
Final Acc: 1.000
Params: 139,265
Deep
Final Acc: 1.000
Params: 51,905
Shallow
Final Acc: 1.000
Params: 4,225
High_Dropout
Final Acc: 1.000
Params: 49,793
Training Accuracy
Validation Accuracy
Training Loss
Validation Loss
Accuracy
Loss
Epoch