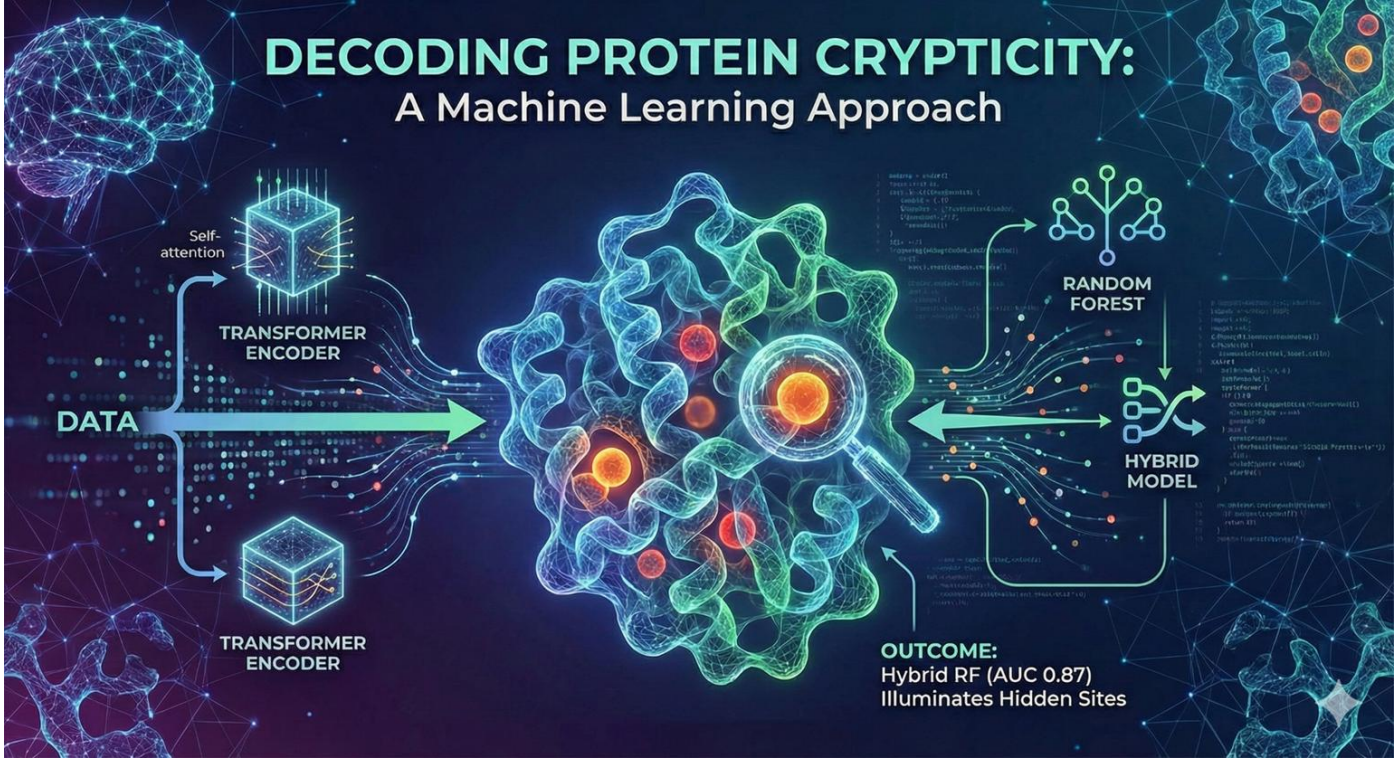


Charlie Li: Data Prep, Transformers, Feature Importance

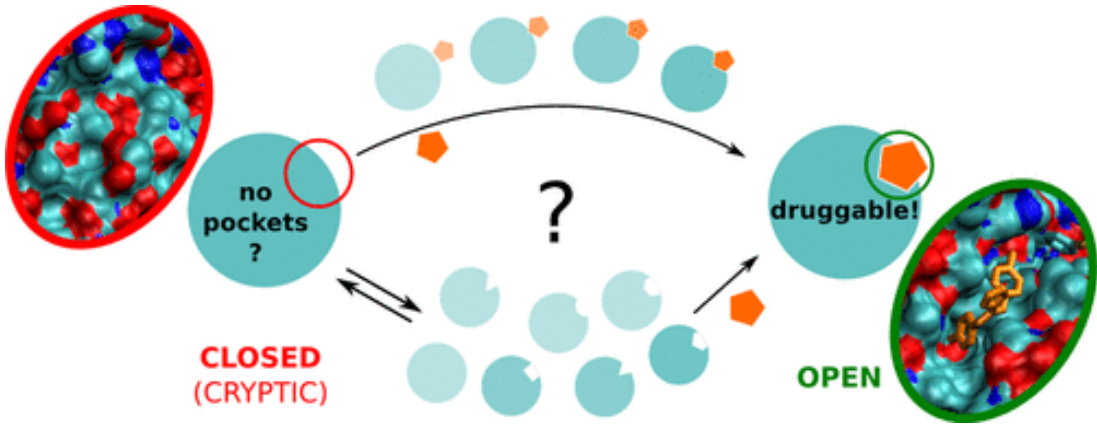
Matthew Rao: Data Prep, Single Models, Hybrid Models

Image source: Nano Banana

Contents

Cryptic Pockets

- Solvent exposure only upon ligand binding [2]
- Potential drug targets



Adapted from Oleinikovas et al., 2016

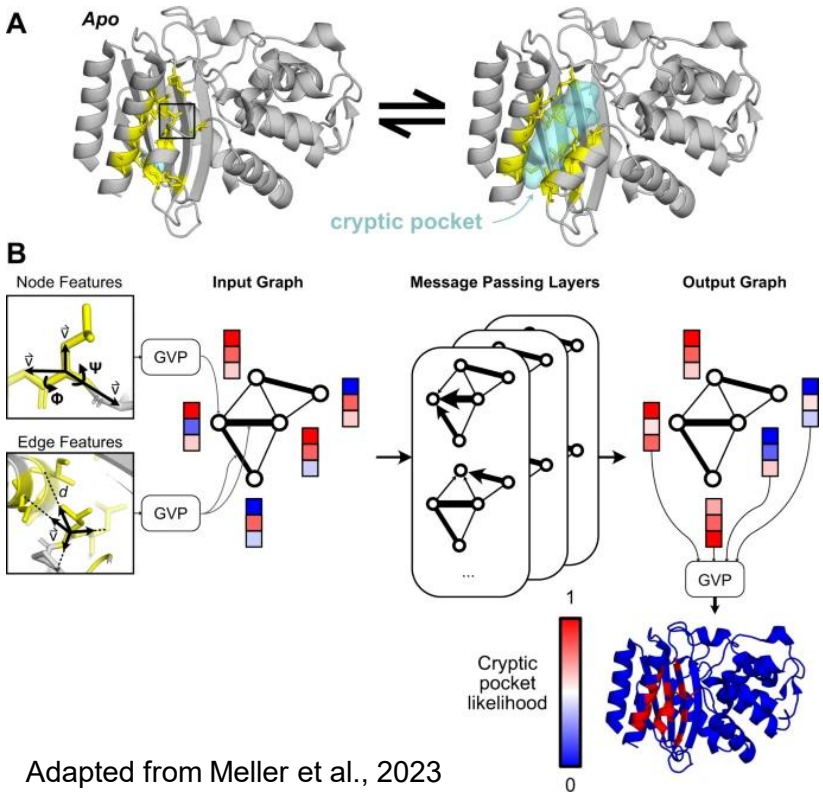
Challenges

- Dynamic property but static imaging
- Ground-state inaccessible: spontaneous closing in molecular dynamics simulation

(2) Bemelmans et al., 2025

Past Work

- **CryptoSite** [3]
  - Simulation input on-the-fly
  - Trained on known examples
  - Run time: 1 day/protein
  - ROC-AUC: 0.83
- **PocketMiner** [4]
  - Graph neural network
  - Trained on simulations
  - Fast run time
  - ROC-AUC: 0.79 and 0.83



Adapted from Meller et al., 2023

(3) Cimermancic et al, 2016.; (4) Meller et al., 2023

Issues

- Representativeness of the starting point static structure
- Left out other relevant information
  - Evolutionary conservation, post translational modification, etc

Our Project

- Residue-level Crypticity Classification
- Sequence and Multiple Feature Categories
- Hybrid Model, ROC-AUC = 0.87

Contents

First, Some Quick Definitions…

- ASA = Accessible Surface Area
- RSA = Relative Solvent Accessibility = $\frac{ASA_{observed}}{\max(ASA)} \times 100\%$
- Crypticity Criterion:

max(RSA)>25 && min(RSA)<25 && max(RSA) - min(RSA) > 20

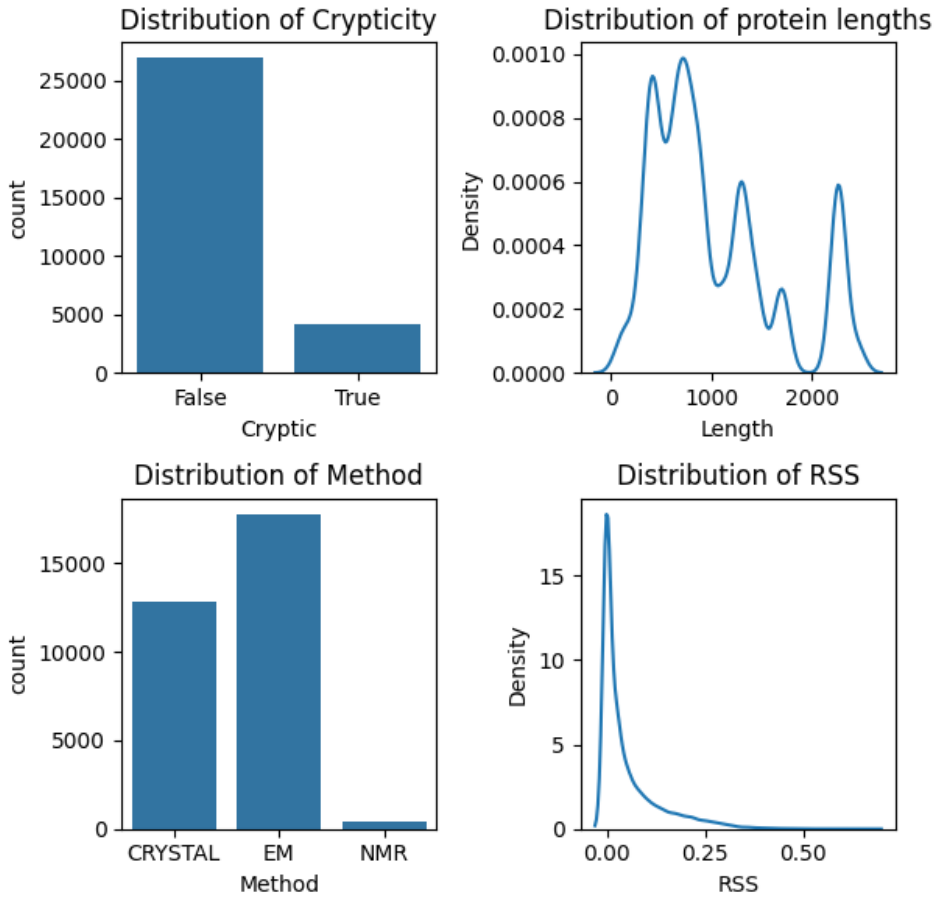
  - Why 25 and 20? Arbitrary designation.

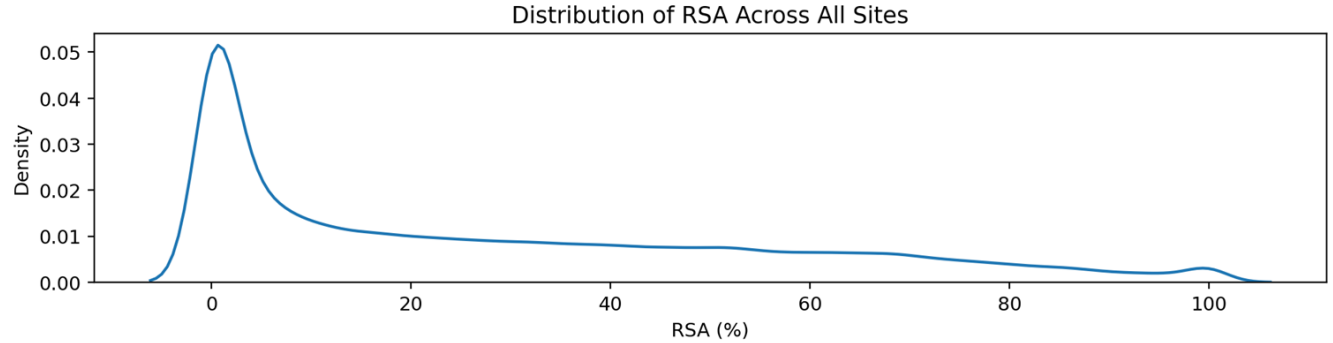
Data Sources

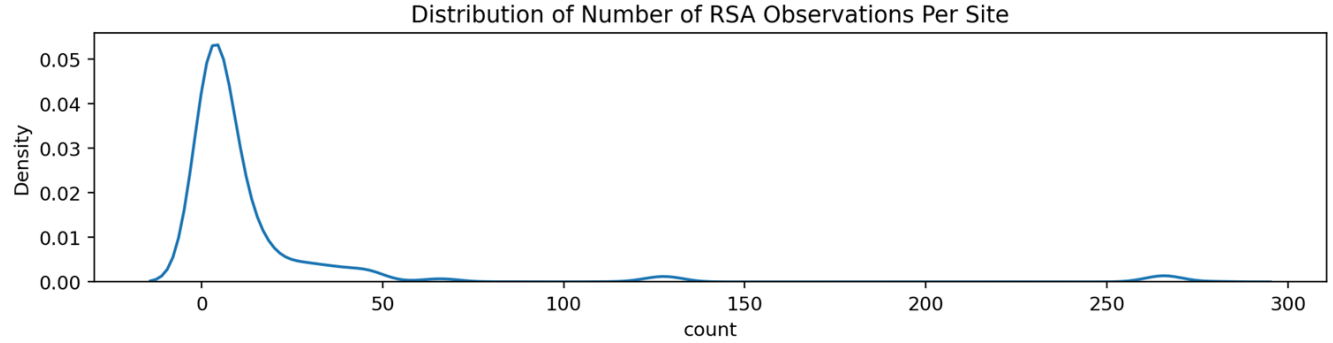
- **Protein Data Bank** (PDB) [5]: Sequence, Method, Annotations, residue RSA (calculated through a developed pipeline)
- **Uniprot** [6]: Protein names
- Amino Acid Chemical Properties
- **Aminode** [7]: Relative Substitution Score (RSS)
- **PhosphoSite** [8]: Post Translational Modifications (No Match)

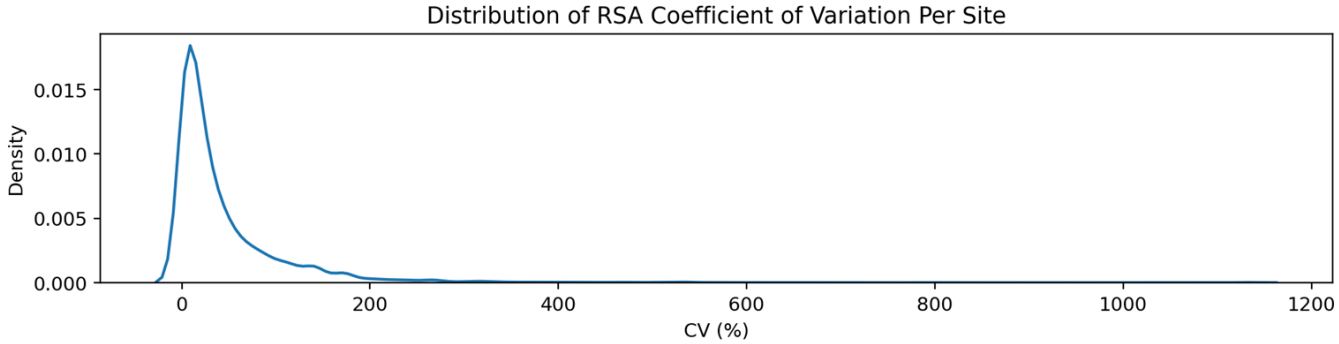
- CSV Table
- 500,000 Observations
- After aggregation: 31,080 rows
- After dropping RSS mismatches: 27,844 rows

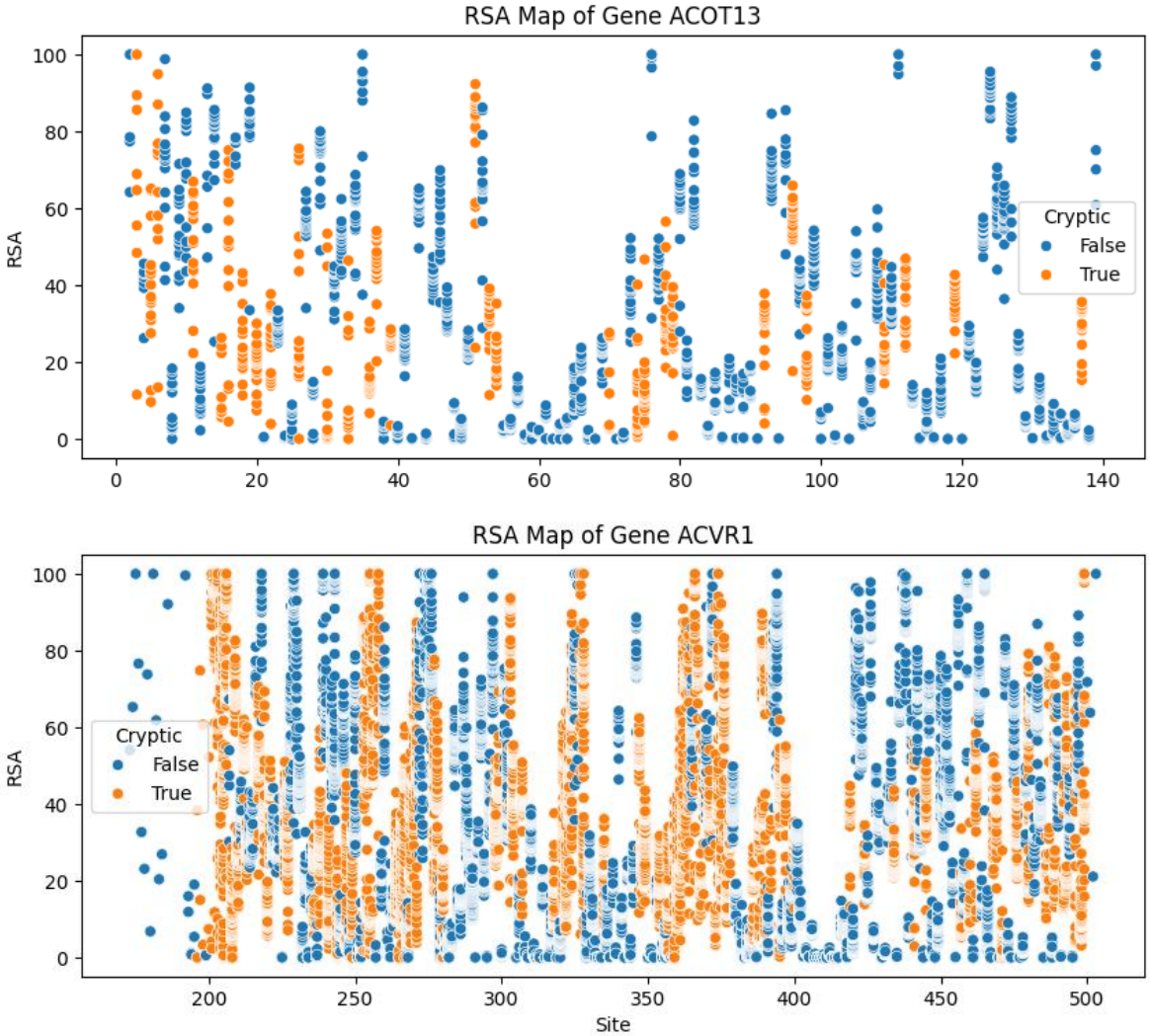
(5) Berman et al., 2000; (6) Bateman et al., 2024;. (7) Chang et al., 2018; (8) Hornbeck et al., 2014


Distribution of Crypticity
count
25000
20000
15000
10000
5000
0
False
True
Cryptic
Distribution of protein lengths
Density
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
1000
2000
Length
Distribution of Method
count
15000
10000
5000
0
CRYSTAL
EM
NMR
Method
Distribution of RSS
Density
15
10
5
0
0.00
0.25
0.50
RSS

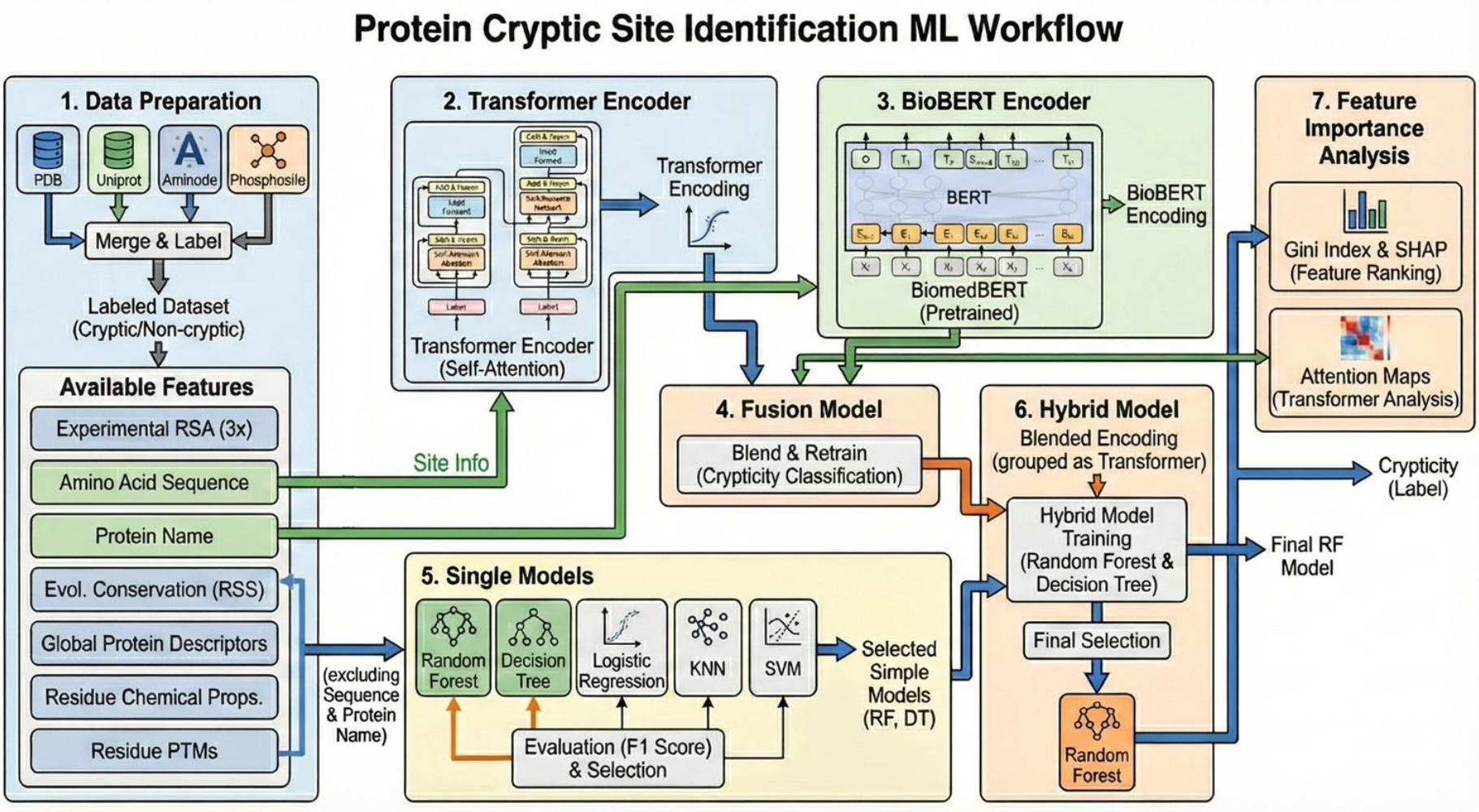

Distribution of RSA Across All Sites
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
RSA (%)


Distribution of Number of RSA Observations Per Site
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
count


Distribution of RSA Coefficient of Variation Per Site
Density
0.015
0.010
0.005
0.000
0
200
400
600
800
1000
1200
CV (%)


RSA Map of Gene ACOT13
RSA
Cryptic
False
True
RSA Map of Gene ACVR1
RSA
Cryptic
False
True
Site

Contents



Made With Nano Banana

a. Transformer Encoder

```python
def forward(self, x, positions):
    h = self.emb(x) + self.pe.unsqueeze(0)
    key_padding = (x == self.padding_idx)
    h2, attn = self.mha(h, h, h, key_padding_mask=key_padding, need_weights=True)
    h = self.ln1(h + h2)
    h2 = self.ff(h)
    h = self.ln2(h + h2)
    # extract encoding at target positions
    batch_idx = torch.arange(x.size(0), device=x.device, dtype=int)
    target_embedding = h[batch_idx, positions]
    encoded = self.cls(target_embedding)
    return encoded, target_embedding, attn
```

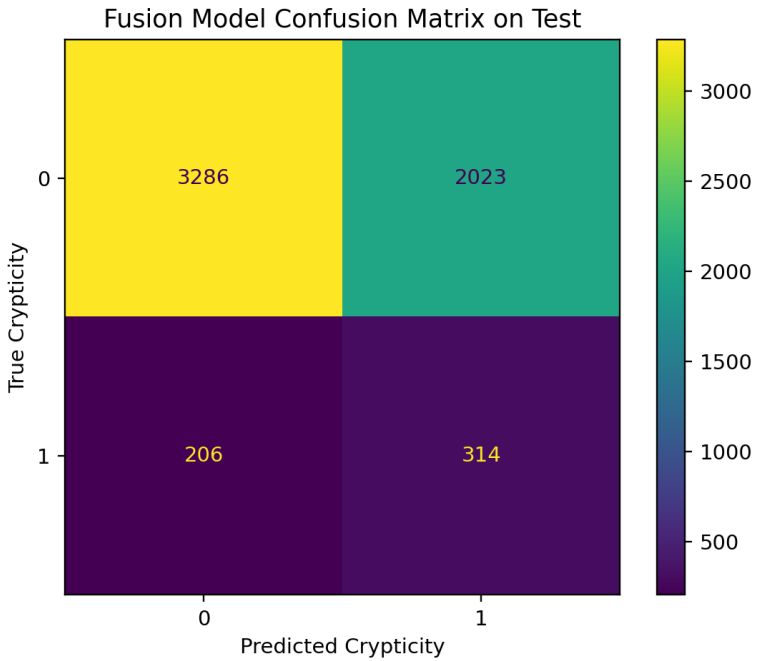
b. Fusion Model With Pre-trained BiomedBERT [9]

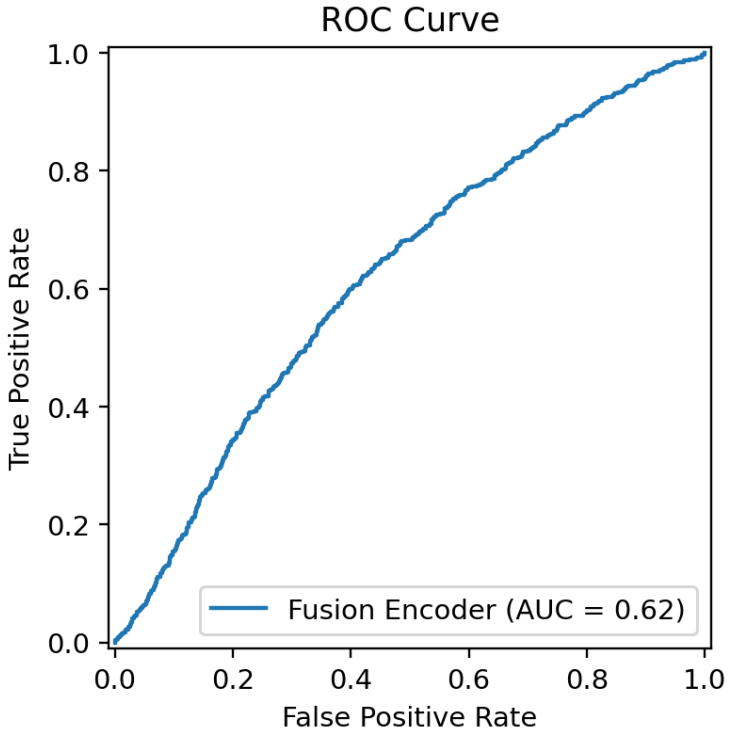
```
def forward(self, protein_encoding, site_encoding):
  # Modality Dropout (Training Only)
  if self.training:
    if torch.rand(1).item() < 0.3:
      protein_encoding = torch.zeros_like(protein_encoding)
    elif torch.rand(1).item() < 0.1:
      site_encoding = torch.zeros_like(site_encoding)
  mat = self.bn(protein_encoding)
  mat = self.ff1(mat)
  mat = self.ln1(mat)
  mat = self.ff2(mat)
  mat = site_encoding + 0.1*mat
  mat = self.ln2(mat)
  logit = self.cls(mat)
  return logit, mat
```

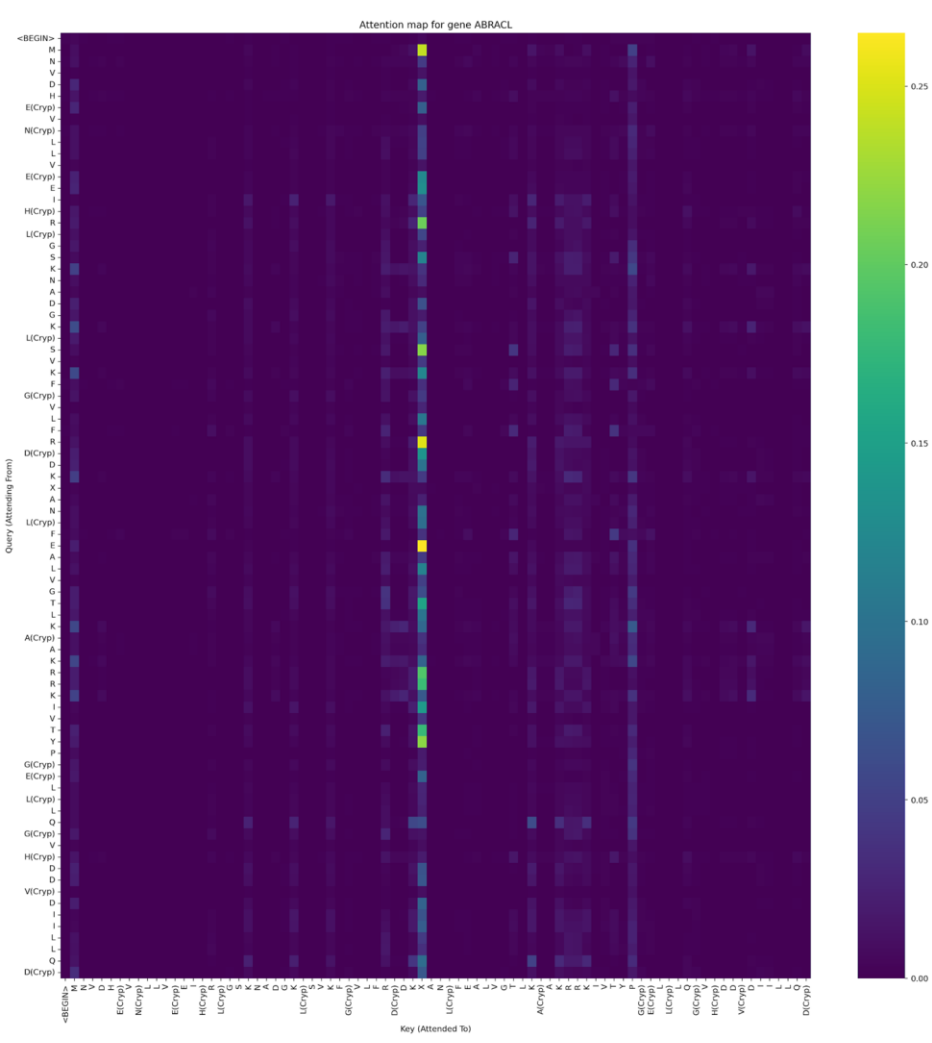
(9) Gu et al., 2022

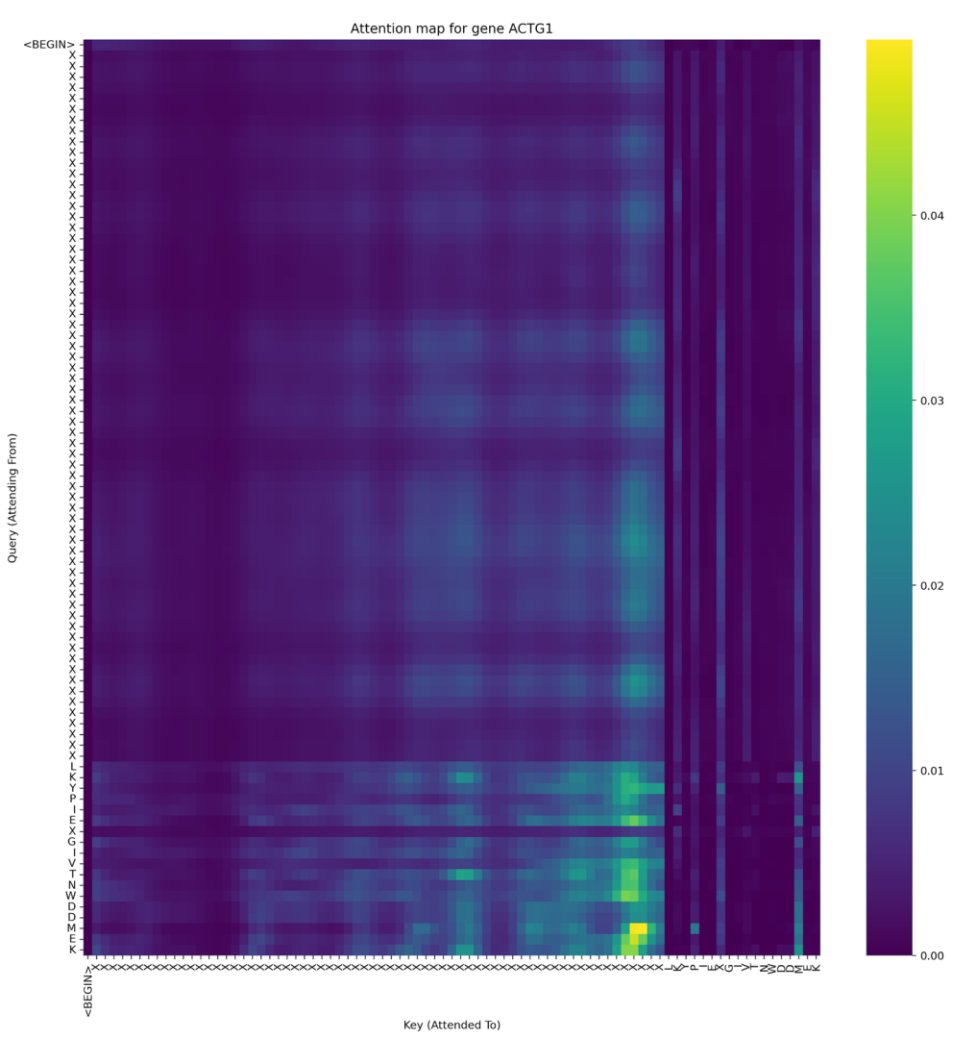
- Accuracy: 0.62
- F1 Score: 0.22
- ROC AUC: 0.621

Decent accuracy but poor F1 score






Attention map for gene ABRACL
Query (Attending From)
Key (Attended To)
0.25
0.20
0.15
0.10
0.05
0.00


Attention map for gene ACTG1
Query (Attending From)
Key (Attended To)
0.04
0.03
0.02
0.01
0.00

c. Single Model Analysis

- Utilized 5 Single Models
  - Random Forest
  - Logistic Regression
  - Decision Tree
  - KNN Classifier
  - Support Vector Machine
- Rank by F1 score

```
class_weight={0: 1, 1: 6.5} # it's 6.5 times more important to correctly classify a cryptic site
```

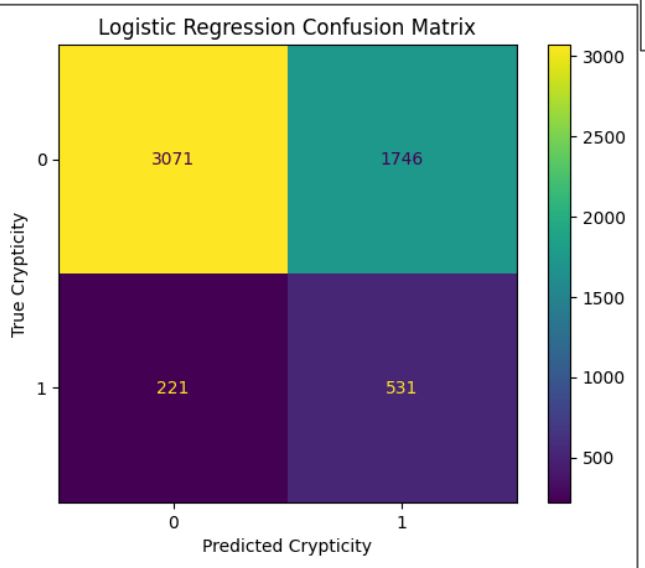
```
rf_model = RandomForestClassifier(
  n_estimators=300,
  min_samples_split=20,
  min_samples_leaf=10,
  class_weight=class_weight,
  random_state=114514,
  n_jobs=-1,
  verbose=True
)
```

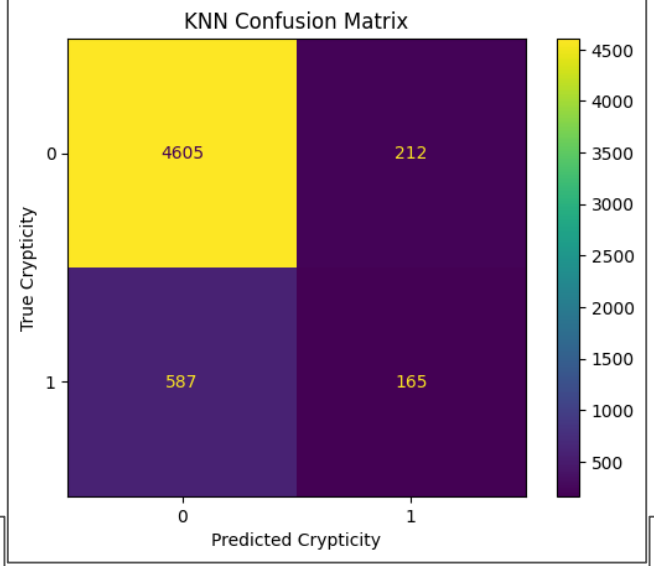
```
rf_model.fit(X_train, y_train)
y_pred = rf_model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("F1 Score:", f1_score(y_test, y_pred))
```

```
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 2 concurrent workers.
[Parallel(n_jobs=-1)]: Done  46 tasks      | elapsed:    1.7s
[Parallel(n_jobs=-1)]: Done 196 tasks      | elapsed:   11.3s
[Parallel(n_jobs=-1)]: Done 300 out of 300 | elapsed:   15.7s finished
[Parallel(n_jobs=2)]: Using backend ThreadingBackend with 2 concurrent workers.
[Parallel(n_jobs=2)]: Done  46 tasks      | elapsed:    0.1s
Accuracy: 0.8125336685221763
F1 Score: 0.4846989141164857
[Parallel(n_jobs=2)]: Done 196 tasks      | elapsed:    0.2s
[Parallel(n_jobs=2)]: Done 300 out of 300 | elapsed:    0.4s finished
```

Logistic Regression

- Accuracy: 0.65
- F1 Score: 0.35



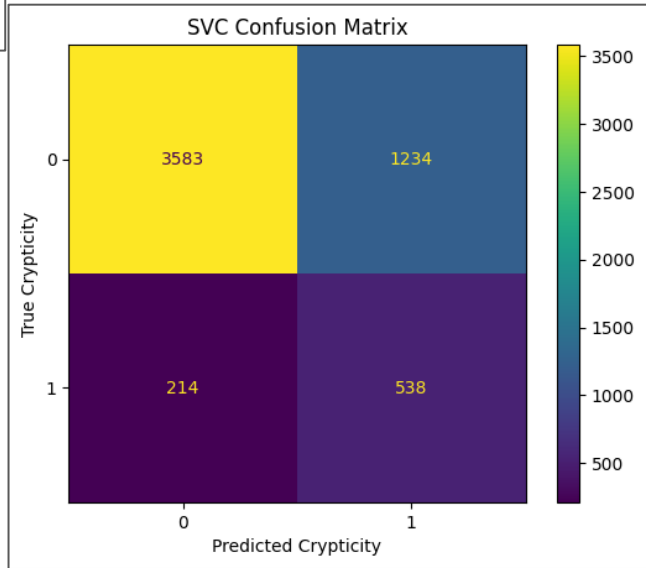


KNN

- Accuracy: 0.86
- F1 Score: 0.29

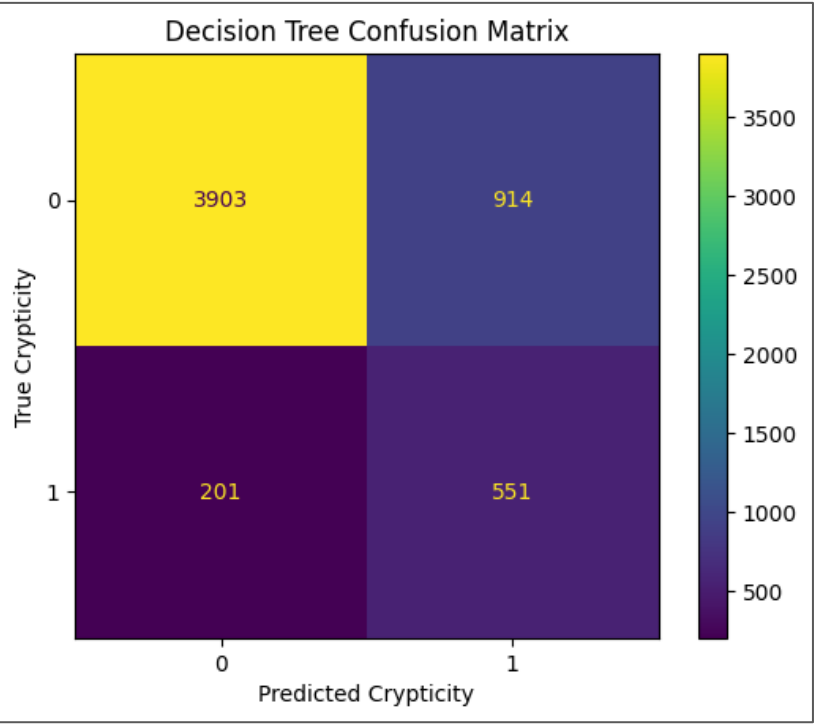
SVM

- Accuracy: 0.74
- F1 Score: 0.43



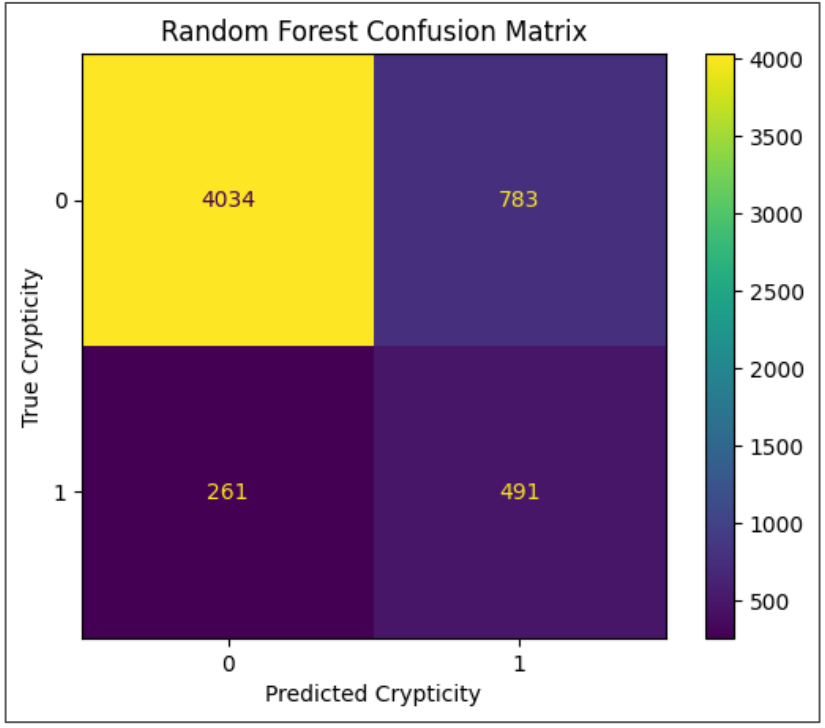
Decision Tree

- Accuracy: 0.80
- F1 Score: 0.50



Random Forest

- Accuracy: 0.81
- F1 Score: 0.48

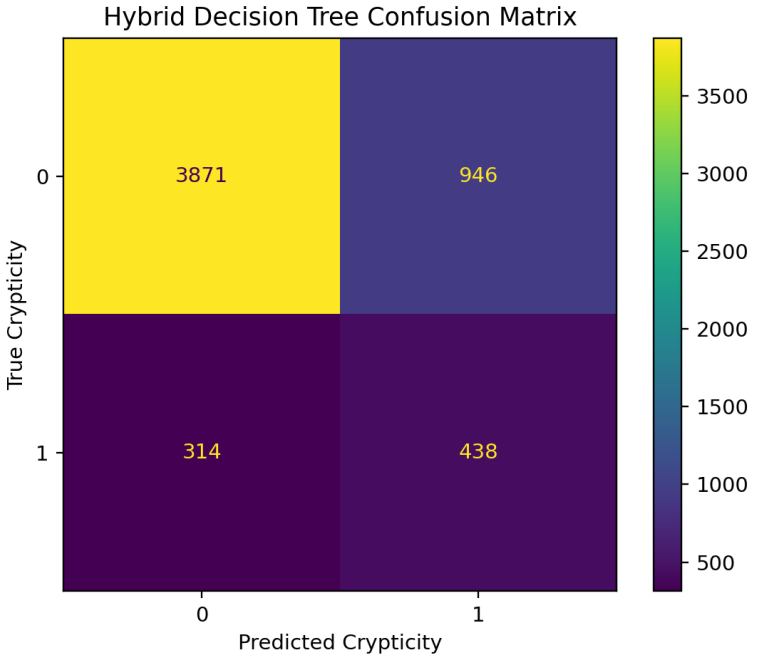


d. Hybrid Models

- Decision Tree, Random Forest
- Combined Features
  - Transformer encoding (64D vectors)
  - Three RSA random samples
  - Evolutionary conservation (RSS)
  - Global protein descriptors
  - Residue chemical properties
  - Residue post-translational modifications

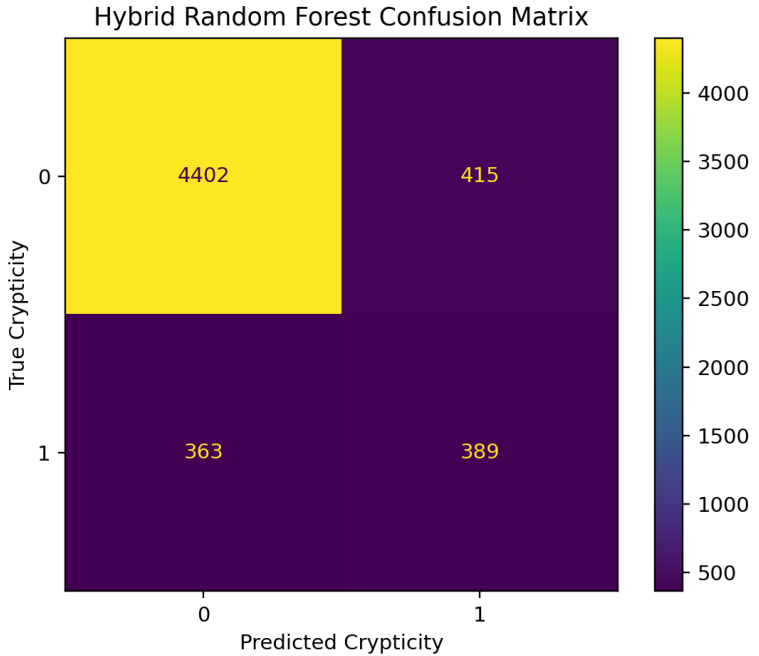
Hybrid Decision Tree

- Accuracy: 0.77
- F1 Score: 0.41
- Performance dropped!



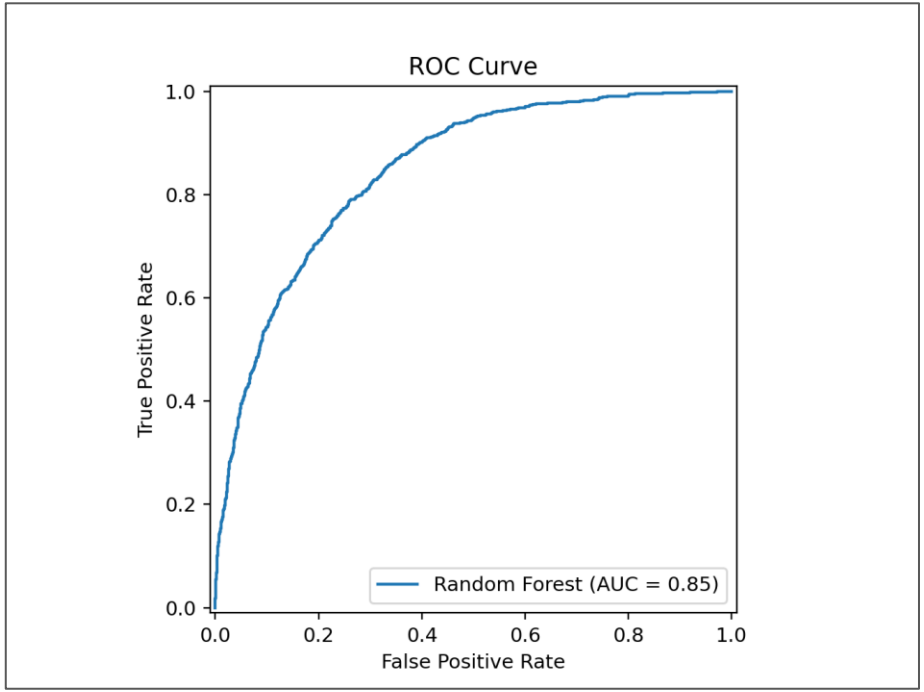
Hybrid Random Forest

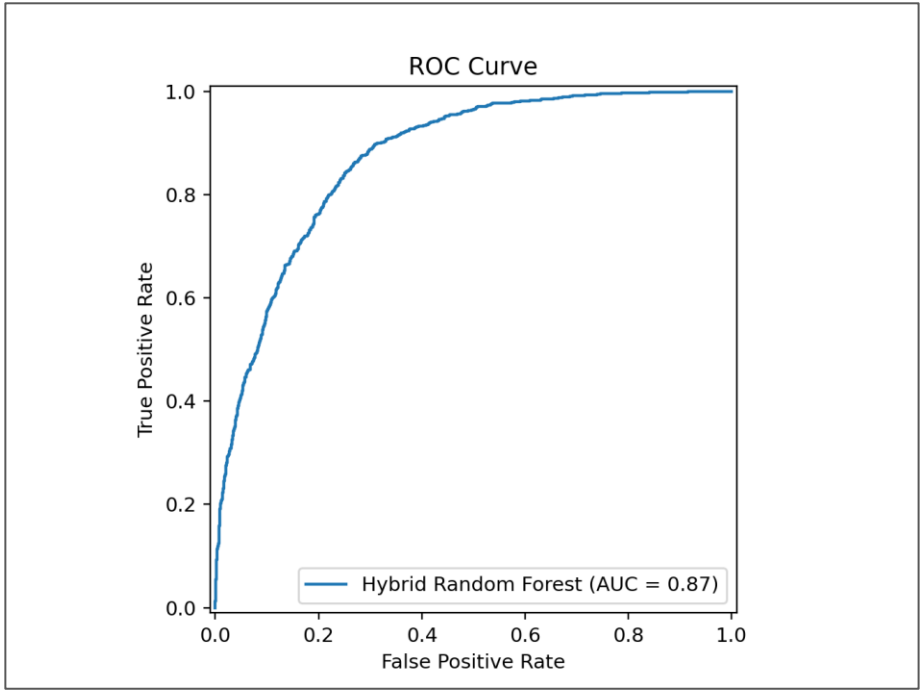
- Accuracy: 0.86
- F1 Score: 0.50
- More stable!



ROC AUC before and after

Slight improvement!



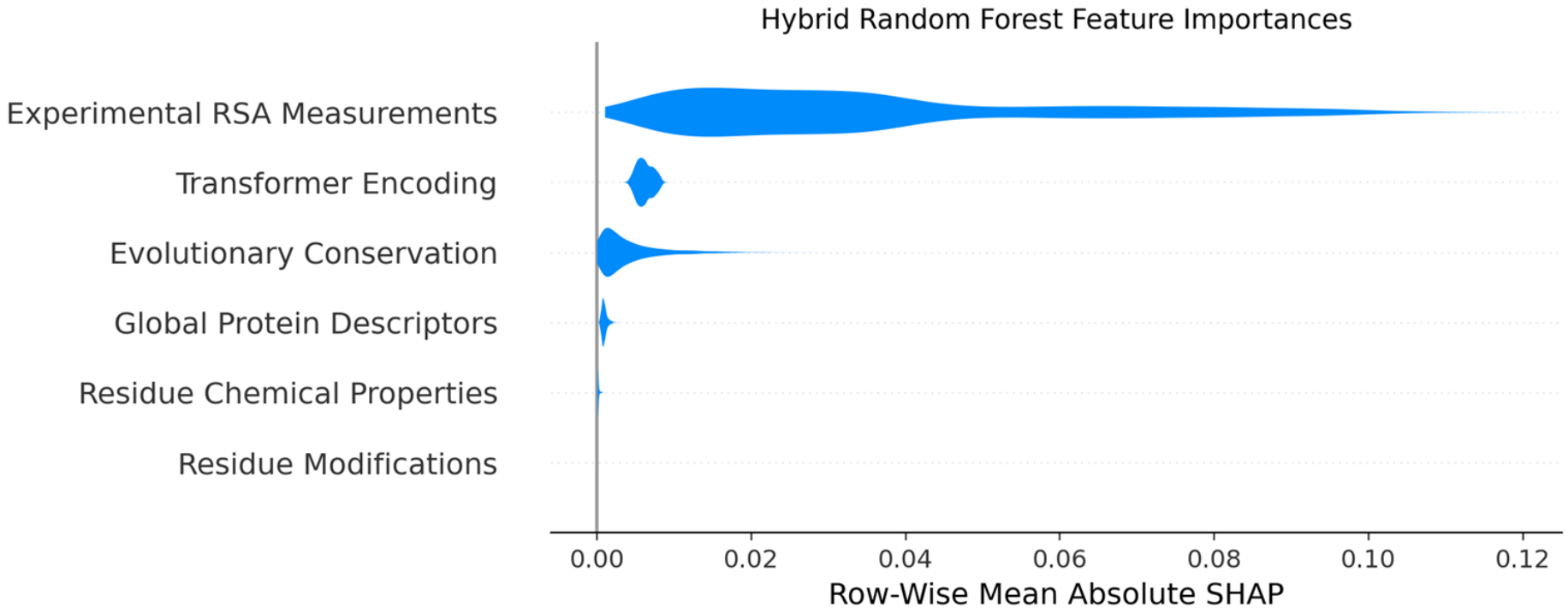


Contents

SHAP: SHapley Additive exPlanations

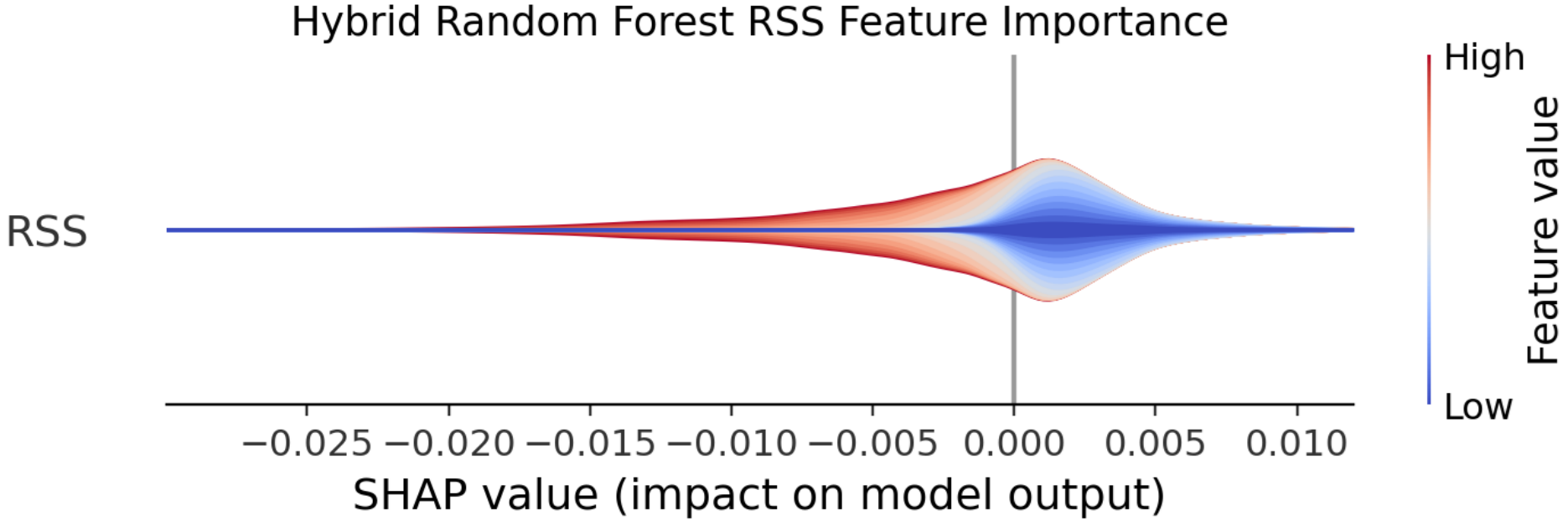
- **Contribution of each feature to the final prediction, averaged across all combinations of other features**
- **Great properties:**
  - Same units as output (probabilities in our case)
  - Additive

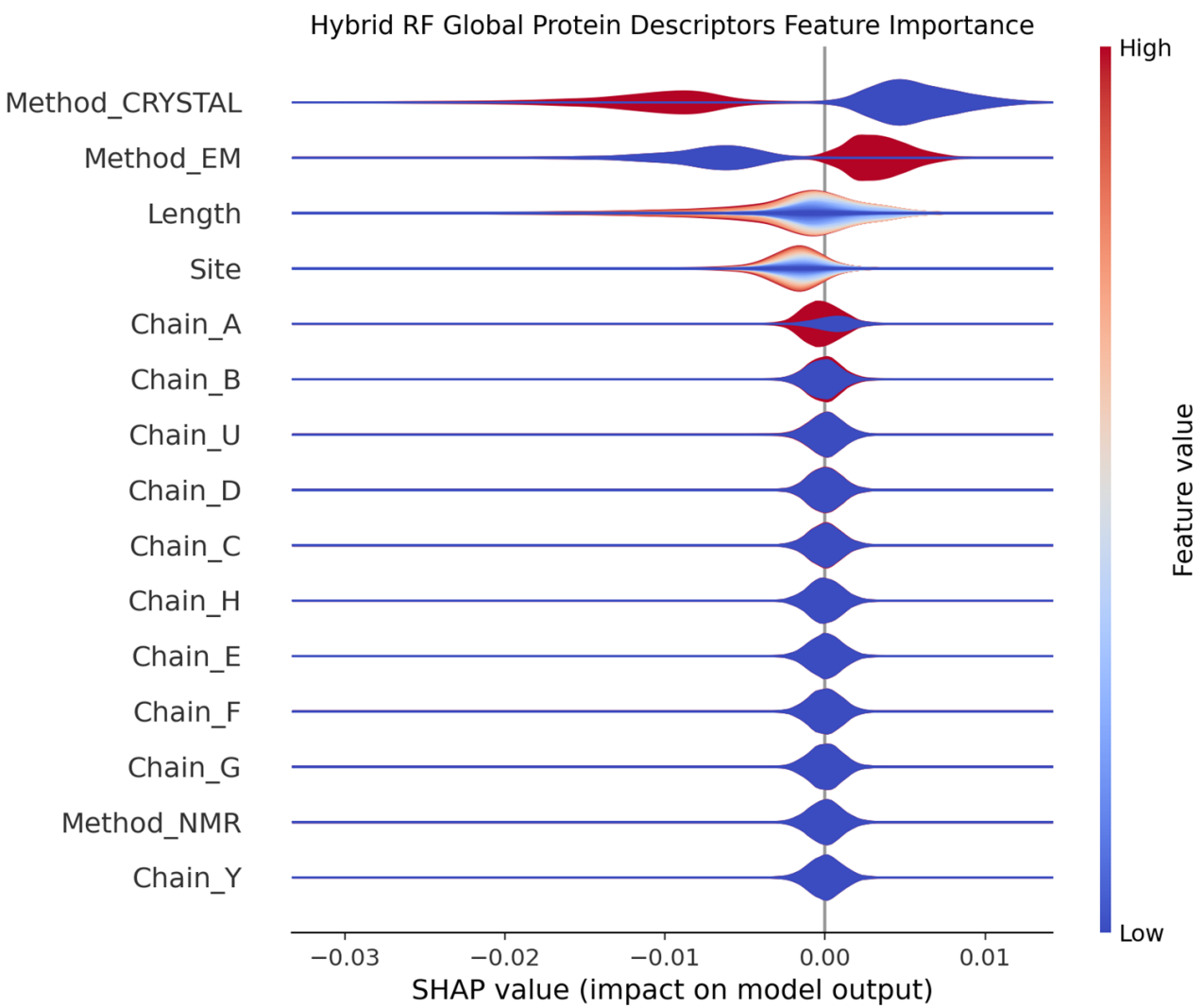
Row-Wise Mean Absolute SHAP By Categories



SHAP Layered Violin Plot

- **High feature value + positive SHAP = positive contribution**
- **High feature value + negative SHAP = negative contribution**




Hybrid RF Global Protein Descriptors Feature Importance
Method_CRYSTAL
Method_EM
Length
Site
Chain_A
Chain_B
Chain_U
Chain_D
Chain_C
Chain_H
Chain_E
Chain_F
Chain_G
Method_NMR
Chain_Y
−0.03
−0.02
−0.01
0.00
0.01
SHAP value (impact on model output)
High
Feature value
Low

Feature Importance Summary

- Functionally important sites tend to be more cryptic
- Method influences crypticity observation
- Residue-level properties are poor predictors

Contents

Limitations and Risks

- Class imbalance between crypticity target variable
- Use of experimental RSA samples
- Poor PTM coverage
- Lack of relevant training features
  - Ligand effects, 3D structure, structural dynamics
- Transformer largely failed to recognize global structural features (monotonous attention maps)
- Chose 25 and 20 for RSA somewhat arbitrarily

Outlook

- Integrate with pretrained structural predictors (e.g. AlphaFold)
- Integrate with molecular dynamics simulations
- Integrate with ligand docking predictors
- Include more comprehensive PTM data
- Include protein functional annotations

References

1. Oleinikovas, Vladimiras; Saladino, G.; Cossins, B. P.; Francesco Luigi Gervasio. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *JACS* **2016**, *138* (43), 14257–14263. https://doi.org/10.1021/jacs.6b05425.
2. Bemelmans, M. P.; Cournia, Z.; Damm-Ganamet, K. L.; Gervasio, F. L.; Pande, V. Computational Advances in Discovering Cryptic Pockets for Drug Discovery. *Curr. Opin. Struct. Biol.* **2025**, *90*, 102975. https://doi.org/10.1016/j.sbi.2024.102975.
3. Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Bio.* **2016**, *428* (4), 709–719. https://doi.org/10.1016/j.jmb.2016.01.029.
4. Meller, A.; Ward, M.; Borowsky, J.; Kshirsagar, M.; Lotthammer, J. M.; Oviedo, F.; Ferres, J. L.; Bowman, G. R. Predicting Locations of Cryptic Pockets from Single Protein Structures Using the PocketMiner Graph Neural Network. *Nat. Commun.* **2023**, *14* (1). https://doi.org/10.1038/s41467-023-36699-3.
5. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
6. Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Adesina, A.; Ahmad, S.; Bowler-Barnett, E. H.; Hema Bye-A-Jee; Carpentier, D.; Denny, P.; Fan, J.; Garmiri, P.; Jose, L.; Hussein, A.; Alexandr Ignatchenko; Insana, G.; Rizwan Ishtiaq; Joshi, V.; Dushyanth Jyothi; Swaathi Kandasaamy. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2024**, *53* (D1). https://doi.org/10.1093/nar/gkae1010.
7. Chang, K. T.; Guo, J.; di Ronza, A.; Sardiello, M. Aminode: Identification of Evolutionary Constraints in the Human Proteome. *Sci. Rep.* **2018**, *8* (1). https://doi.org/10.1038/s41598-018-19744-w.
8. Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J. M.; Latham, V.; Skrzypek, E. PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations. *Nucleic Acids Res.* **2014**, *43* (D1), D512–D520. https://doi.org/10.1093/nar/gku1267.
9. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* **2022**, *3* (1), 1–23. https://doi.org/10.1145/3458754.

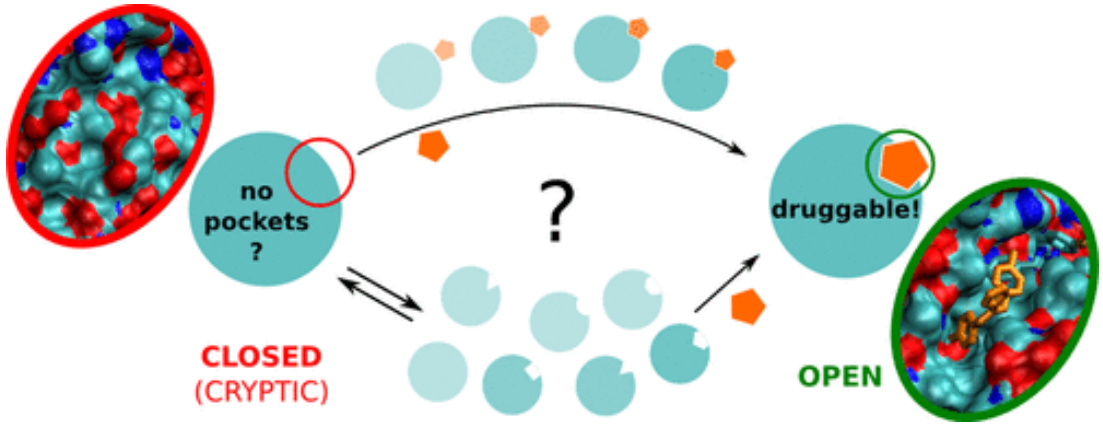
Acknowledgement

Nano Banana (https://gemini.google/overview/image-generation) was used for the generation of certain images as noted on the relevant slides.

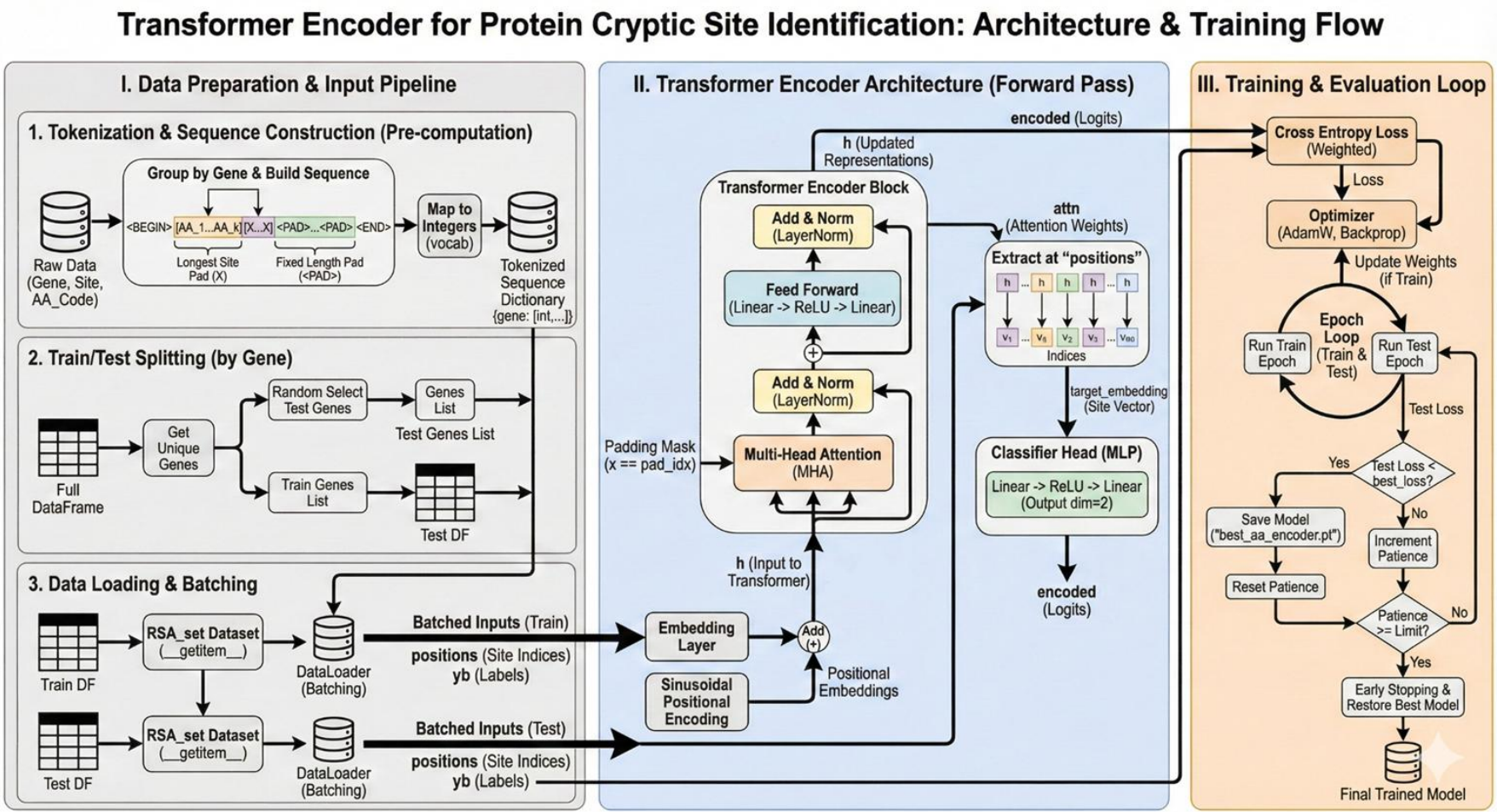
Supplementary Information

Broad Mechanisms for Cryptic Pocket Opening

- **Induced fit**
- **Conformation selection**

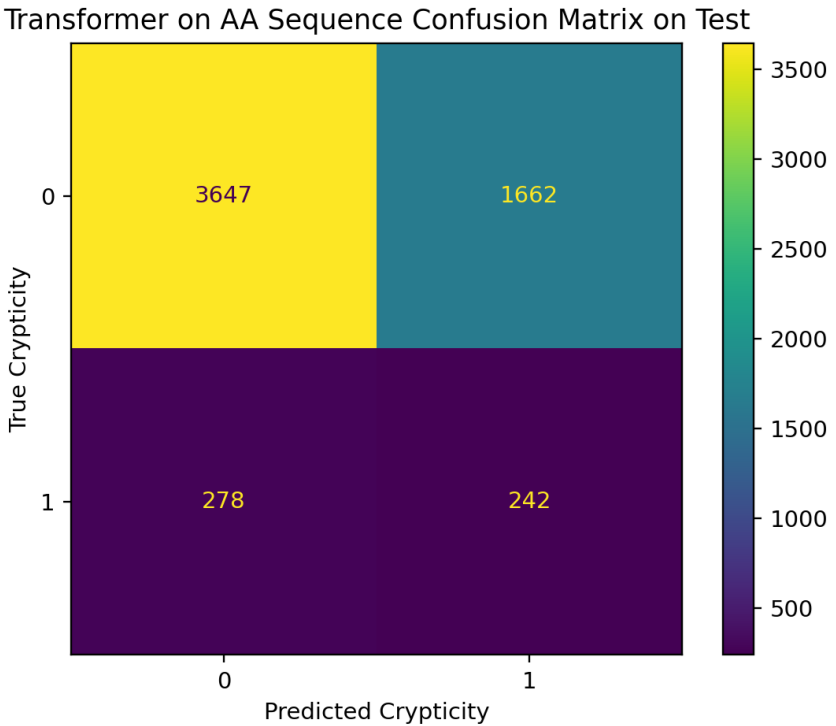


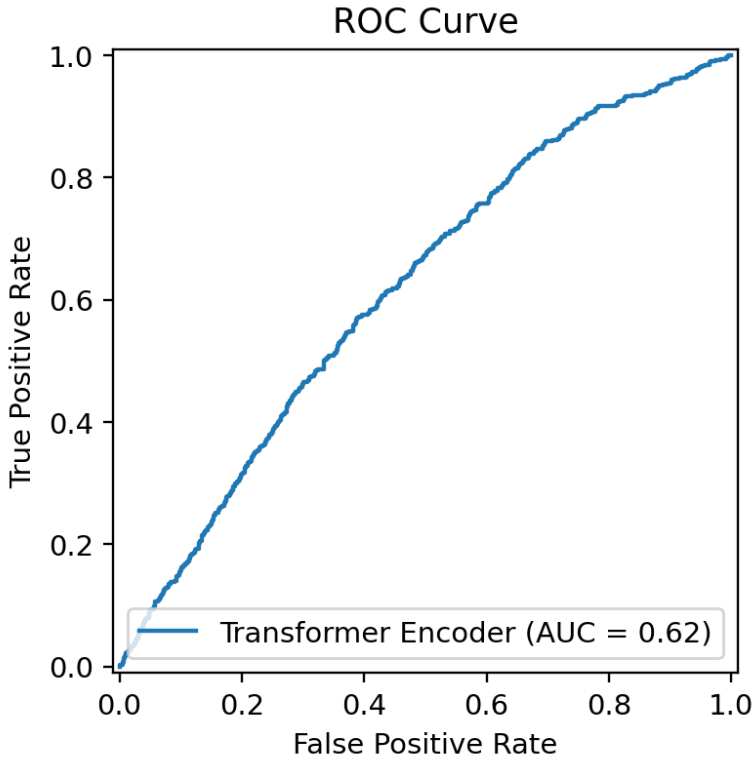
Adapted from Oleinikovas et al.



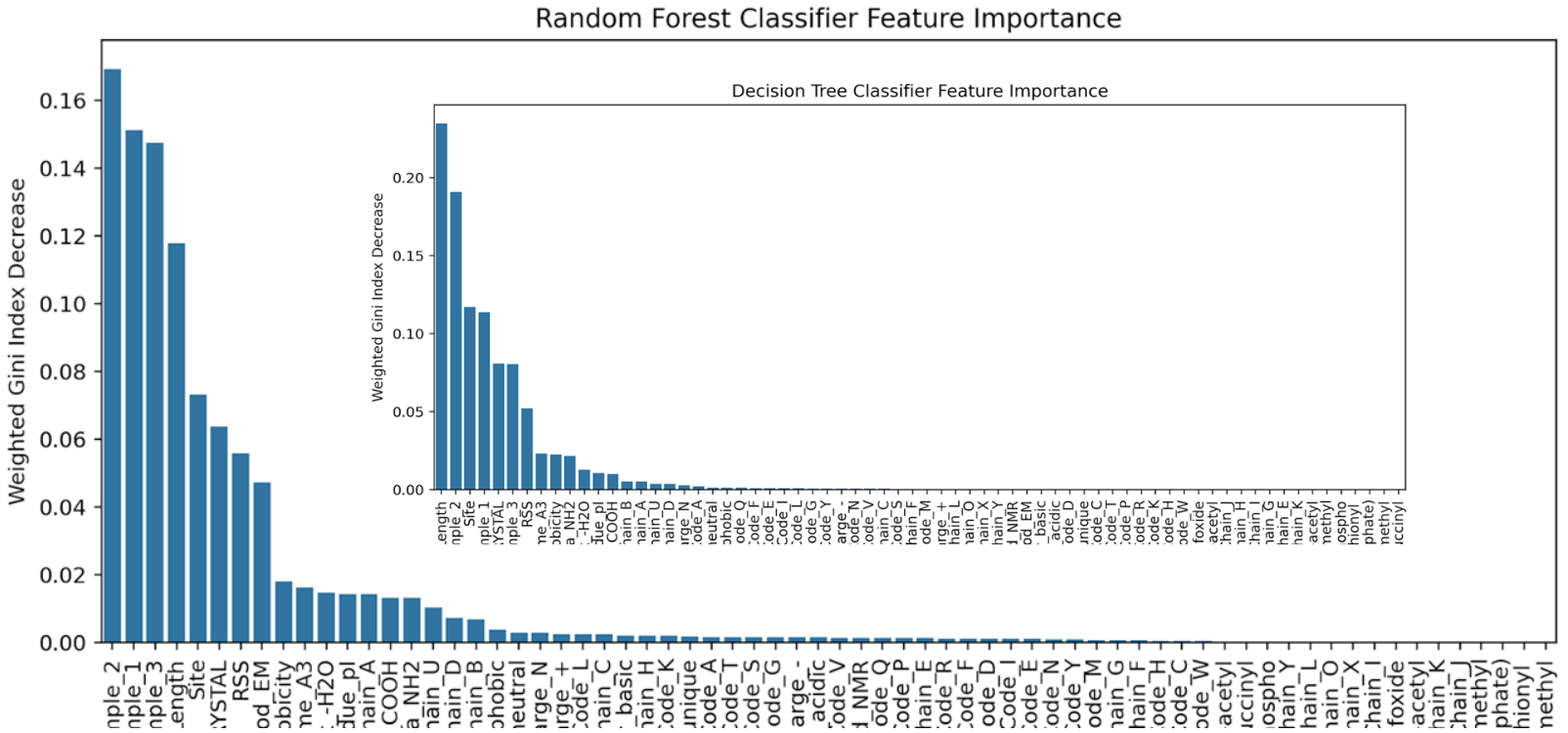
Transformer Encoder Performance

- Accuracy: 0.67
- F1 Score: 0.20
- ROC AUC: 0.618

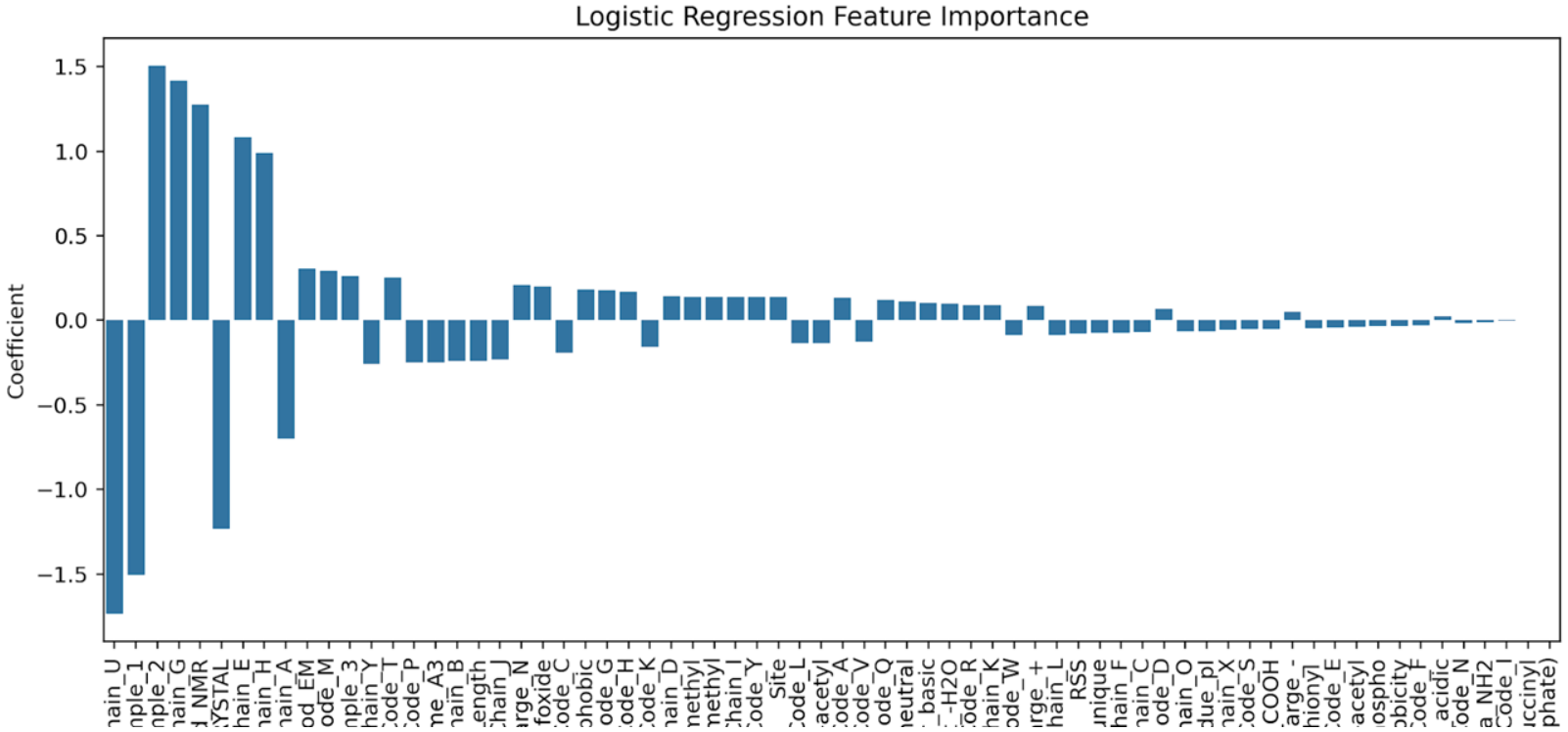


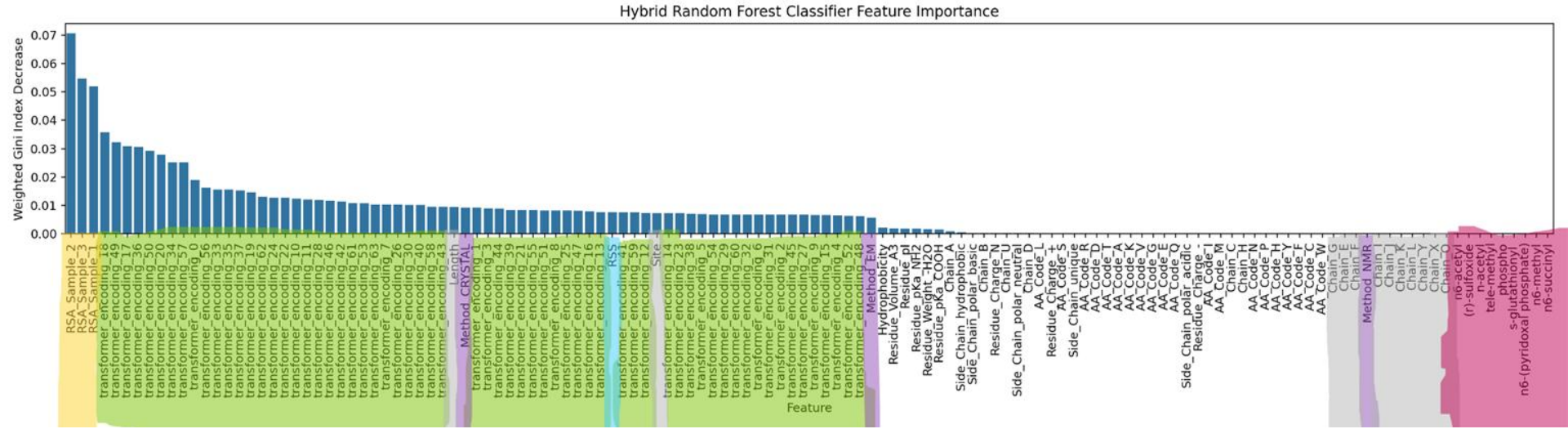


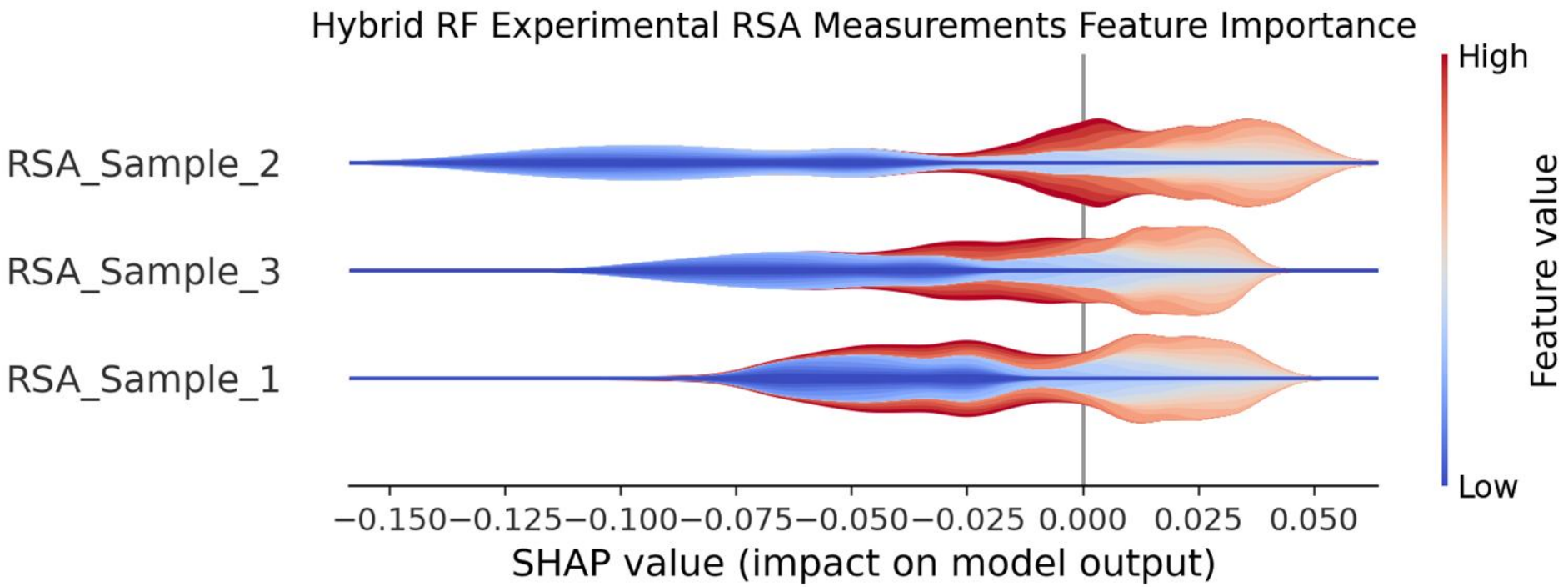
Model dependent: Gini importance

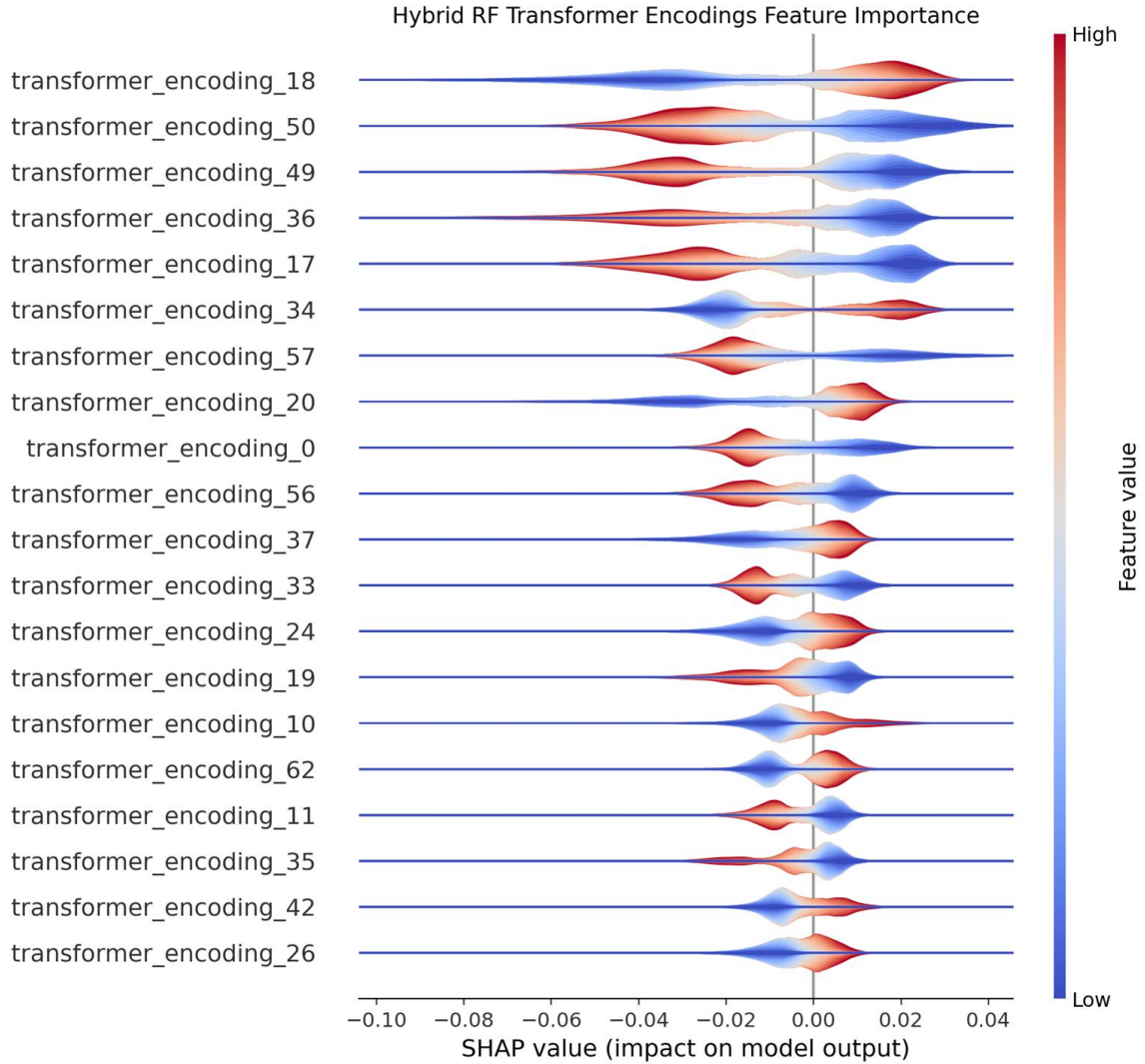


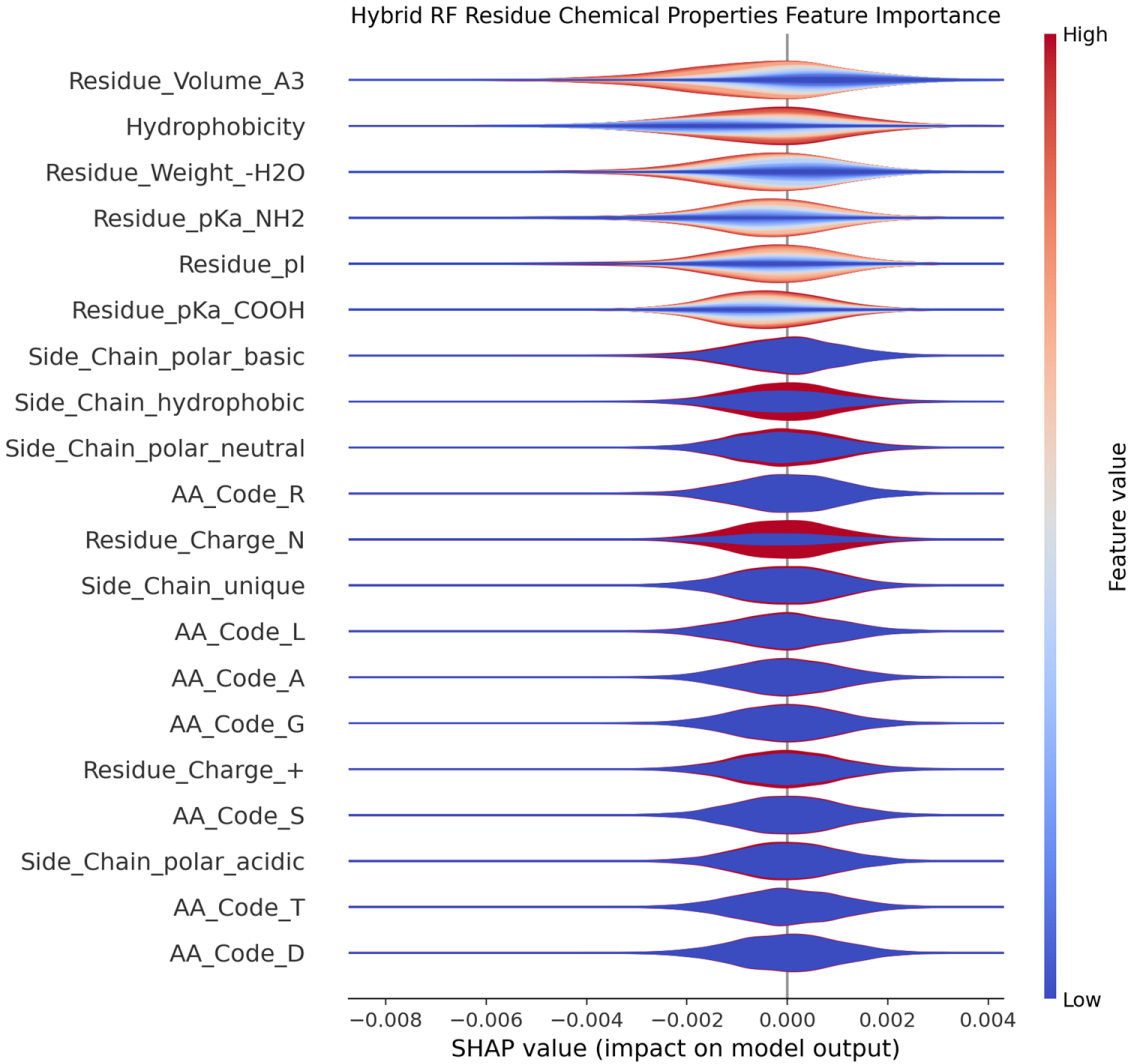
Model dependent: Coefficients

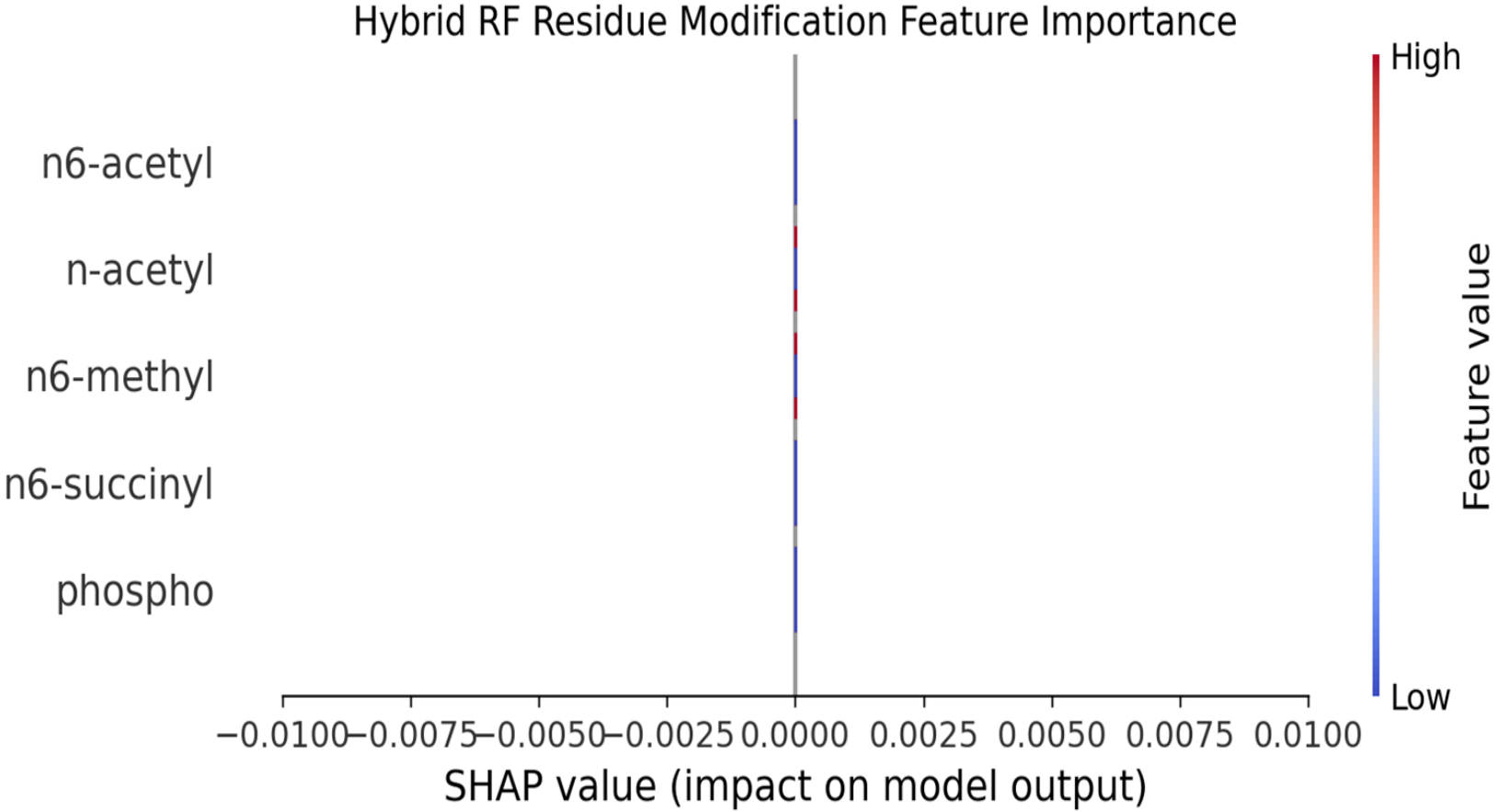



Hybrid Random Forest Classifier Feature Importance
Weighted Gini Index Decrease
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
RSA_Sample_2
RSA_Sample_3
RSA_Sample_1
transformer_encoding_18
transformer_encoding_49
transformer_encoding_17
transformer_encoding_36
transformer_encoding_50
transformer_encoding_20
transformer_encoding_34
transformer_encoding_57
transformer_encoding_0
transformer_encoding_35
transformer_encoding_37
transformer_encoding_19
transformer_encoding_62
transformer_encoding_24
transformer_encoding_22
transformer_encoding_10
transformer_encoding_11
transformer_encoding_28
transformer_encoding_46
transformer_encoding_42
transformer_encoding_61
transformer_encoding_53
transformer_encoding_63
transformer_encoding_7
transformer_encoding_26
transformer_encoding_30
transformer_encoding_40
transformer_encoding_58
transformer_encoding_43
Length
Method_CRYSTAL
transformer_encoding_1
transformer_encoding_3
transformer_encoding_44
transformer_encoding_39
transformer_encoding_21
transformer_encoding_55
transformer_encoding_51
transformer_encoding_8
transformer_encoding_25
transformer_encoding_47
transformer_encoding_16
transformer_encoding_13
RSS
transformer_encoding_41
transformer_encoding_59
transformer_encoding_12
Site_1
transformer_encoding_14
transformer_encoding_23
transformer_encoding_38
transformer_encoding_5
transformer_encoding_54
transformer_encoding_29
transformer_encoding_60
transformer_encoding_32
transformer_encoding_6
transformer_encoding_31
transformer_encoding_2
transformer_encoding_45
transformer_encoding_27
transformer_encoding_9
transformer_encoding_15
transformer_encoding_4
transformer_encoding_52
transformer_encoding_48
Method_EM
Hydrophobicity
Residue_Volume_A3
Residue_pI
Residue_pKa_NH2
Residue_Weight_H2O
Residue_pKa_COOH
Chain_A
Side_Chain_hydrophobic
Side_Chain_polar_basic
Chain_B
Residue_Charge_N
Chain_U
Side_Chain_polar_neutral
Chain_D
AA_Code_L
Residue_Charge_+
AA_Code_S
Side_Chain_unique
AA_Code_R
AA_Code_T
AA_Code_A
AA_Code_K
AA_Code_Y
AA_Code_G
AA_Code_E
AA_Code_Q
Side_Chain_polar_acidic
Residue_Charge_-
AA_Code_I
AA_Code_M
Chain_C
Chain_H
AA_Code_N
AA_Code_P
AA_Code_H
AA_Code_Y
AA_Code_F
AA_Code_C
AA_Code_W
Chain_G
Chain_E
Chain_F
Method_NMR
Chain_I
Chain_J
Chain_K
Chain_L
Chain_Y
Chain_X
Chain_O
n6-acetyl
(r)-sulfoxide
n-acetyl
tele-methyl
phospho
s-glutathionyl
n6-(pyridoxal phosphate)
n6-methyl
n6-succinyl
Feature


Hybrid RF Experimental RSA Measurements Feature Importance
RSA_Sample_2
RSA_Sample_3
RSA_Sample_1
−0.150
−0.125
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
SHAP value (impact on model output)
High
Feature value
Low


Hybrid RF Transformer Encodings Feature Importance
transformer_encoding_18
transformer_encoding_50
transformer_encoding_49
transformer_encoding_36
transformer_encoding_17
transformer_encoding_34
transformer_encoding_57
transformer_encoding_20
transformer_encoding_0
transformer_encoding_56
transformer_encoding_37
transformer_encoding_33
transformer_encoding_24
transformer_encoding_19
transformer_encoding_10
transformer_encoding_62
transformer_encoding_11
transformer_encoding_35
transformer_encoding_42
transformer_encoding_26
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
SHAP value (impact on model output)
High
Feature value
Low


Hybrid RF Residue Chemical Properties Feature Importance
Residue_Volume_A3
Hydrophobicity
Residue_Weight_-H2O
Residue_pKa_NH2
Residue_pI
Residue_pKa_COOH
Side_Chain_polar_basic
Side_Chain_hydrophobic
Side_Chain_polar_neutral
AA_Code_R
Residue_Charge_N
Side_Chain_unique
AA_Code_L
AA_Code_A
AA_Code_G
Residue_Charge_+
AA_Code_S
Side_Chain_polar_acidic
AA_Code_T
AA_Code_D
−0.008
−0.006
−0.004
−0.002
0.000
0.002
0.004
SHAP value (impact on model output)
High
Feature value
Low


Hybrid RF Residue Modification Feature Importance
n6-acetyl
n-acetyl
n6-methyl
n6-succinyl
phospho
−0.0100
−0.0075
−0.0050
−0.0025
0.0000
0.0025
0.0050
0.0075
0.0100
SHAP value (impact on model output)
High
Feature value
Low