AUA American University of Armenia

CRYSTAL STRUCTURE PREDICTIONS

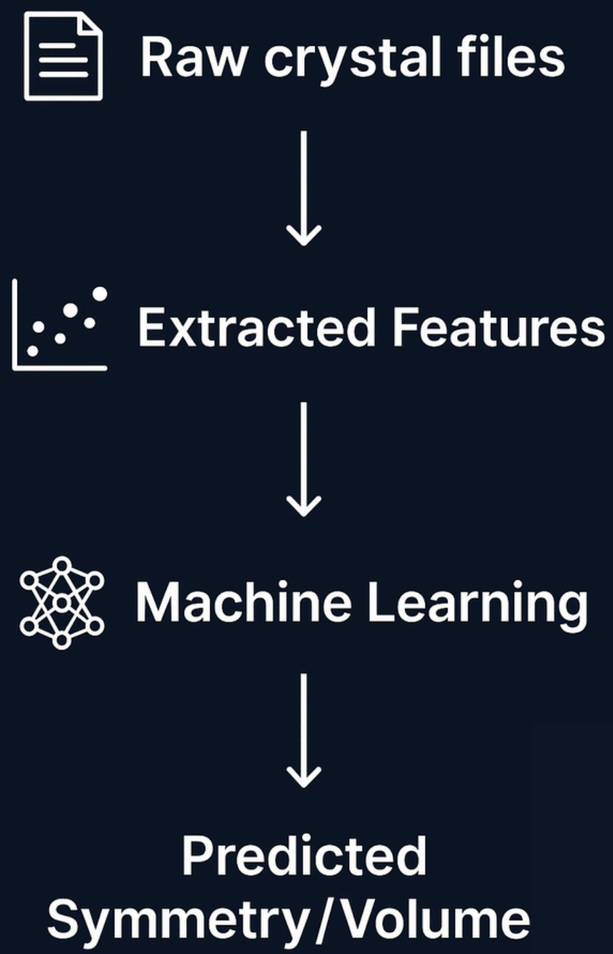
Presented by **Leon Sarkisyan & Vram Papyan**
Supervisor: **Aleksandr Hayrapetyan**

OUTLINE

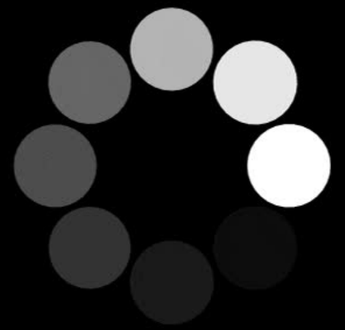
- Introduction
- Data Collection and Processing
- Machine Learning Models
- Deep Learning Models
- Results & Interpretation
- Conclusion

INTRODUCTION

- Crystallography reveals atomic structures in solid materials
- Space group symmetry is a key structural feature
- Traditional methods are slow and complex
- We explore a faster, data-driven approach using machine learning
- Dataset: Crystal structures from the Crystallography Open Database (COD)



DATA COLLECTION & PROCESSING



- Data source: Crystallography Open Database (COD)
- Over 1,000 CIF files parsed for structure info
- Extracted parameters: cell size, angles, volume, symmetry, etc.
- Cleaned & processed data:
  - Removed irrelevant/missing fields
  - Handled missing values (median/mode)
  - Encoded categories, normalized numbers
- Created new features:
  - Anisotropy, axis ratios, shape factor
- Final data: 9 key numerical features + target labels (symmetry, volume)

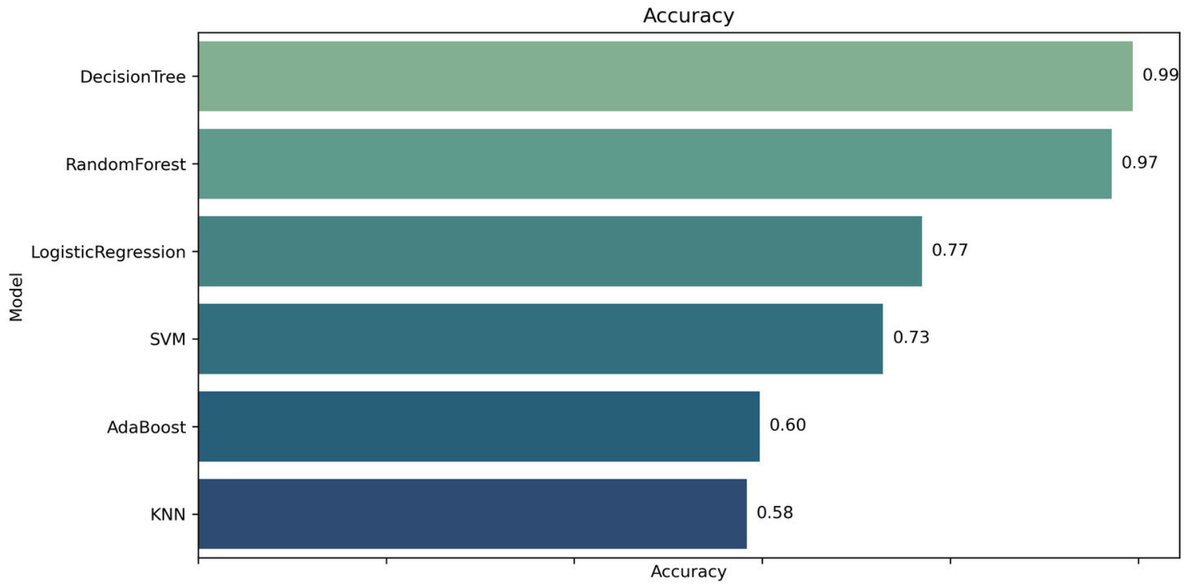
PRE-PROCESSED

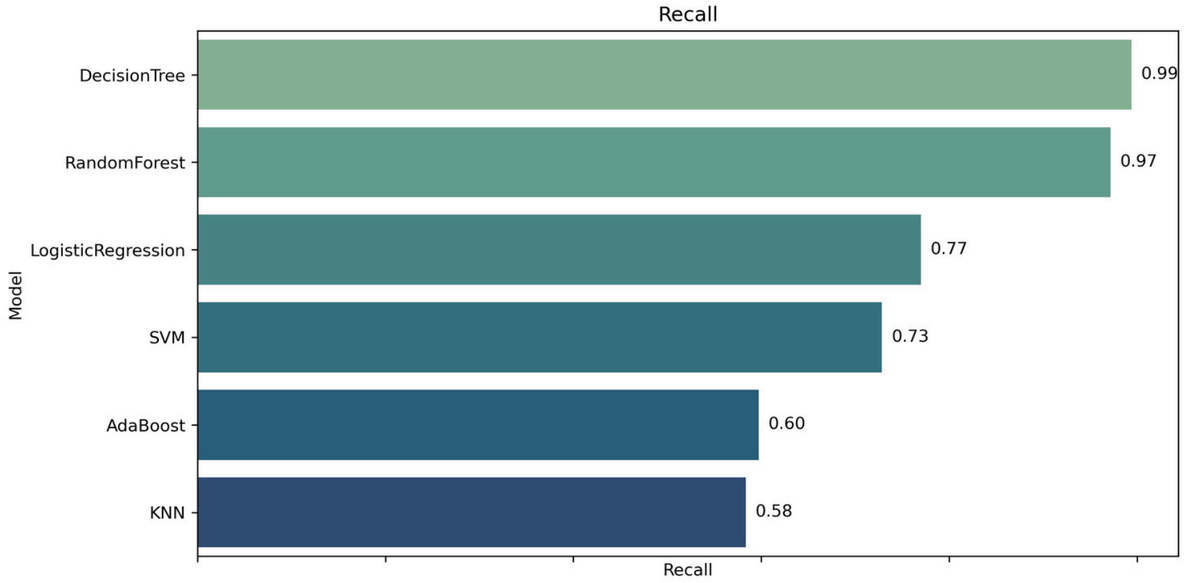
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11328 entries, 0 to 11327
Data columns (total 73 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   file              11328 non-null  int64
 1   a                 11328 non-null  float64
 2   siga              11282 non-null  float64
 3   b                 11328 non-null  float64
 4   sigb              11251 non-null  float64
 5   c                 11328 non-null  float64
 6   sigc              11282 non-null  float64
 7   alpha             11328 non-null  float64
 8   sigalpha          3250 non-null   float64
 9   beta              11328 non-null  float64
 10  sigbeta           8843 non-null   float64
 11  gamma             11328 non-null  float64
 12  siggamma          3253 non-null   float64
 13  vol               11328 non-null  float64
 14  sigvol            11283 non-null  float64
 15  celltemp          11288 non-null  float64
 16  sigcelltemp       8025 non-null   float64
 17  diffrtemp         11294 non-null  float64
 18  sigdiffrtemp      7189 non-null   float64
 19  cellpressure      28 non-null     float64
...
 71  time              11328 non-null  object
 72  onhold            0 non-null      float64
dtypes: float64(42), int64(6), object(25)
memory usage: 6.3+ MB
```

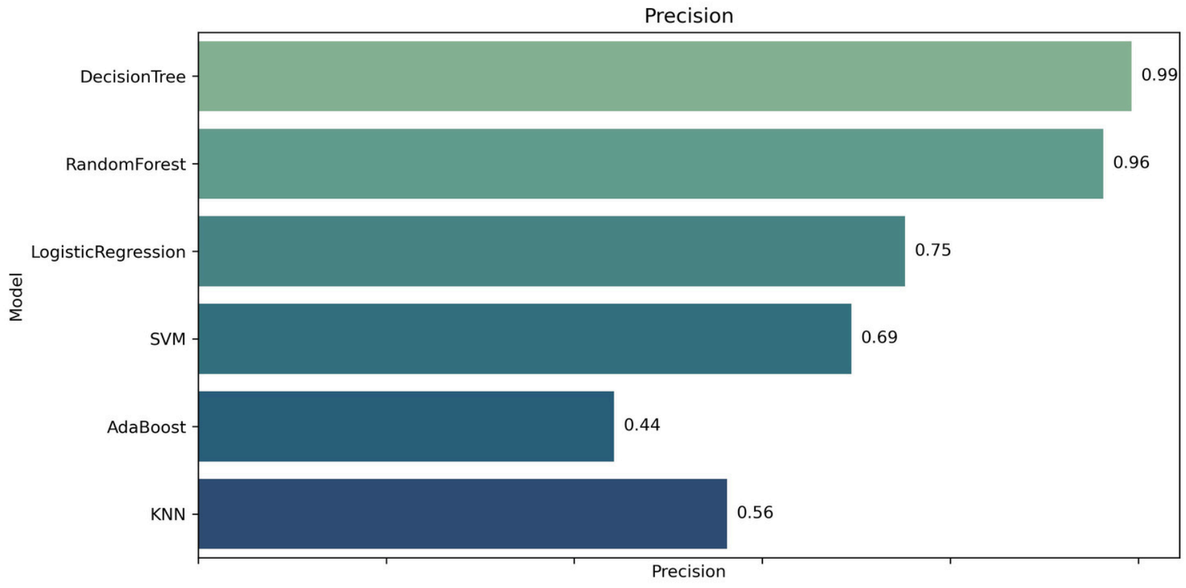
PROCESSED

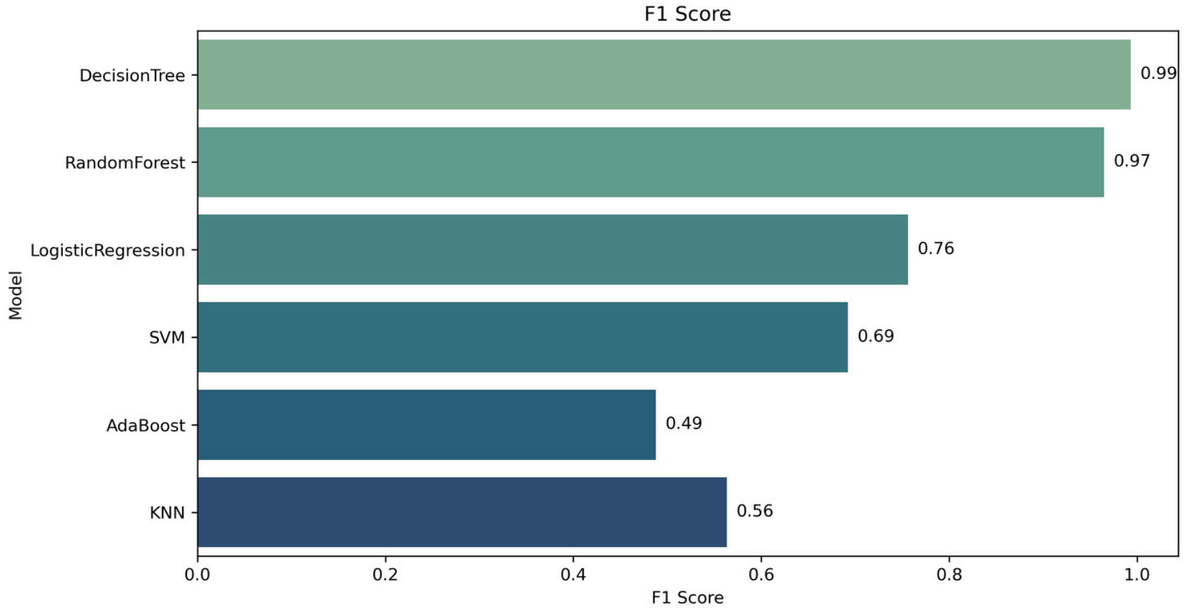
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11328 entries, 0 to 11327
Data columns (total 57 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   file                  11328 non-null  int64
 1   a                     11328 non-null  float64
 2   siga                  11328 non-null  float64
 3   b                     11328 non-null  float64
 4   sigb                  11328 non-null  float64
 5   c                     11328 non-null  float64
 6   sigc                  11328 non-null  float64
 7   alpha                 11328 non-null  float64
 8   beta                  11328 non-null  float64
 9   sigbeta               11328 non-null  float64
 10  gamma                 11328 non-null  float64
 11  vol                   11328 non-null  float64
 12  sigvol                11328 non-null  float64
 13  celltemp              11328 non-null  float64
 14  sigcelltemp           11328 non-null  float64
 15  diffrtemp             11328 non-null  float64
 16  sigdiffrtemp          11328 non-null  float64
 17  nel                   11328 non-null  int64
 18  sg                    11328 non-null  object
 19  sgHall                11328 non-null  object
...
 55  radType_encoded       11328 non-null  int32
 56  flags_encoded         11328 non-null  int32
dtypes: float64(25), int32(14), int64(4), object(14)
memory usage: 4.4+ MB
```

MACHINE LEARNING MODELS


Accuracy
DecisionTree
RandomForest
LogisticRegression
SVM
AdaBoost
KNN
0.99
0.97
0.77
0.73
0.60
0.58
Model
Accuracy

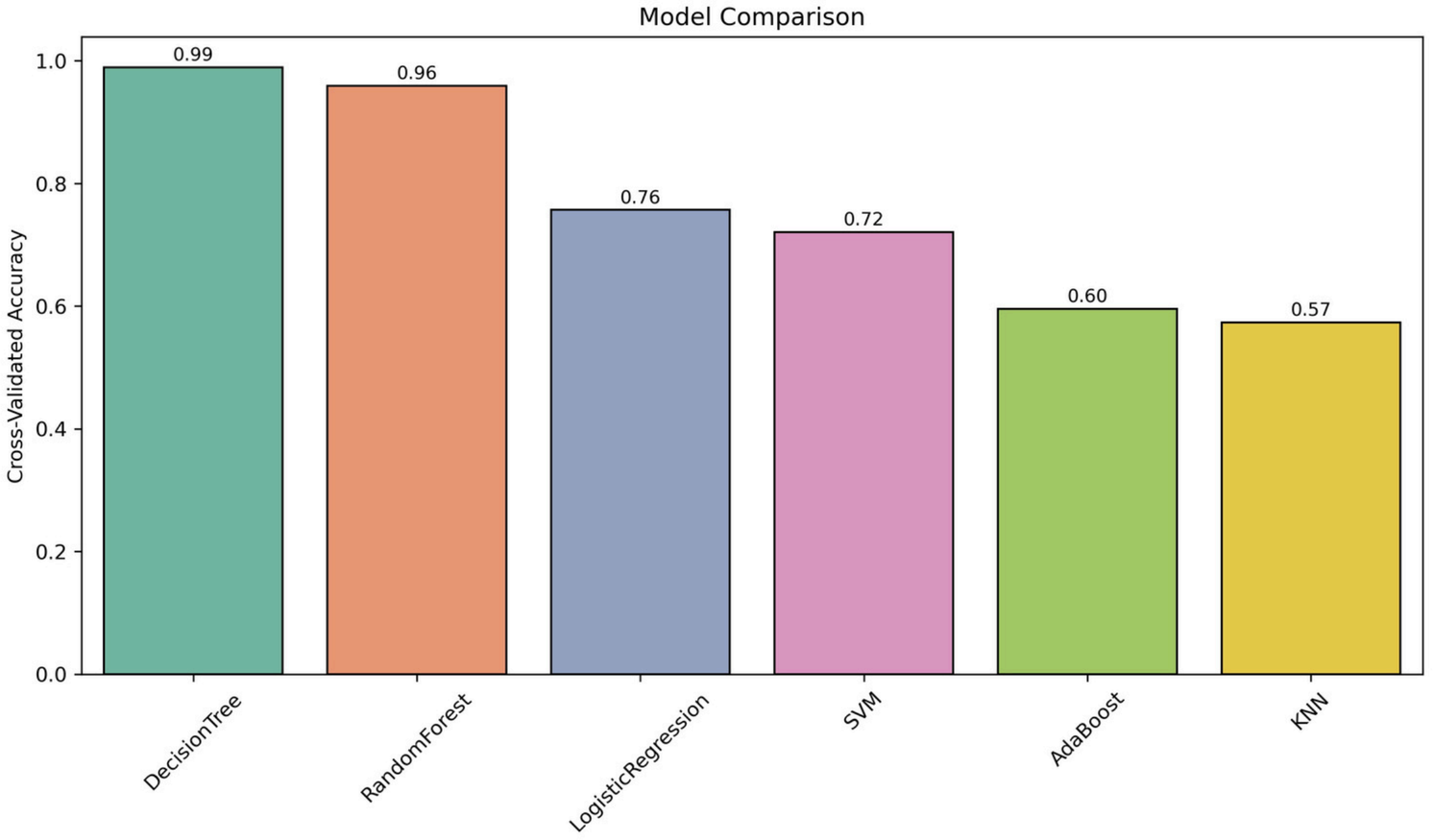

Recall
DecisionTree
RandomForest
LogisticRegression
SVM
AdaBoost
KNN
0.99
0.97
0.77
0.73
0.60
0.58
Model
Recall


Precision
DecisionTree
RandomForest
LogisticRegression
SVM
AdaBoost
KNN
0.99
0.96
0.75
0.69
0.44
0.56
Model
Precision


F1 Score
DecisionTree
RandomForest
LogisticRegression
SVM
AdaBoost
KNN
0.99
0.97
0.76
0.69
0.49
0.56
Model
0.0
0.2
0.4
0.6
0.8
1.0
F1 Score

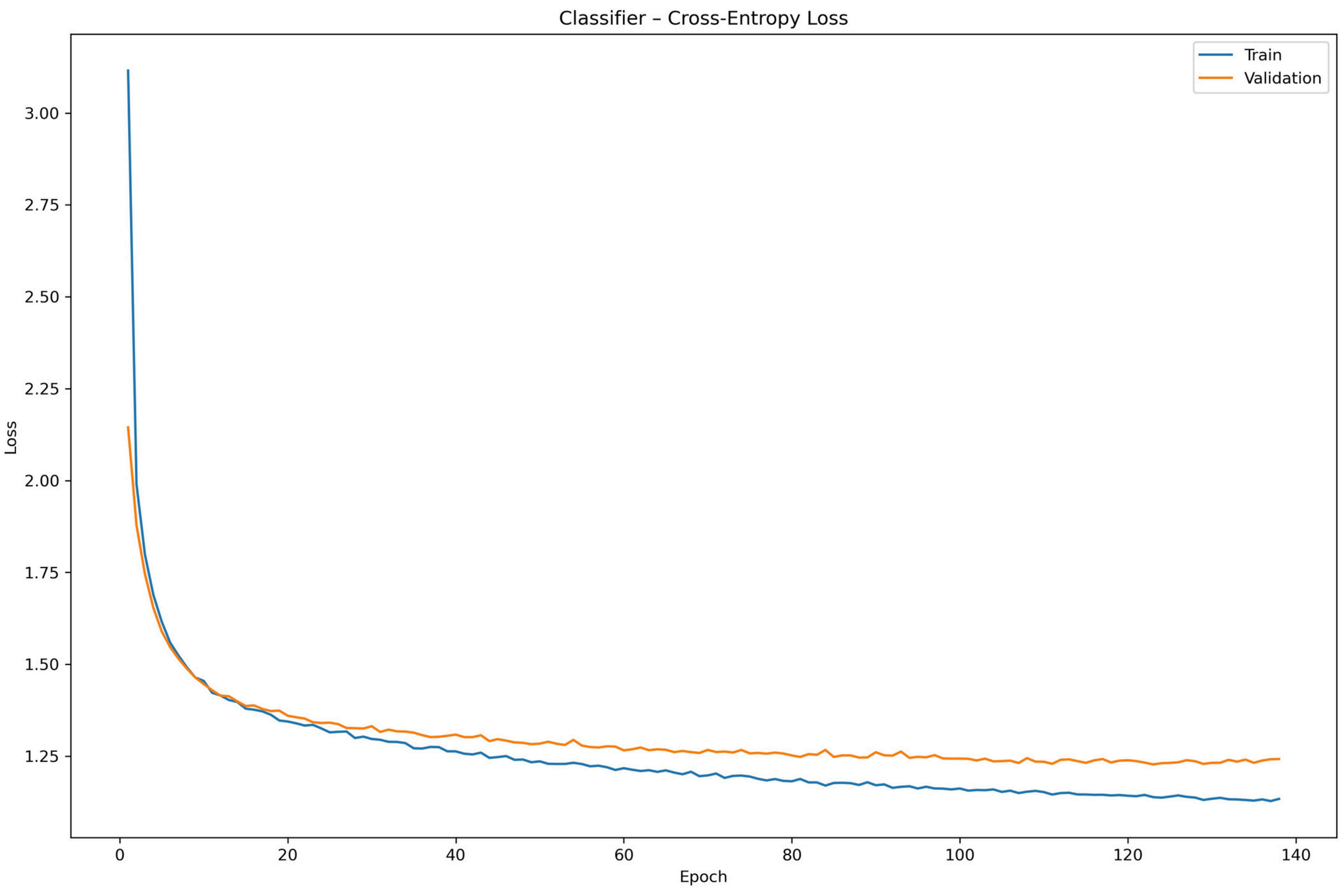
RANKINGS

After 5-fold cross-validation, the Decision Tree and Random Forest Models performed best in terms of mean accuracy. The graph indicates the cross-validated average accuracy for each model



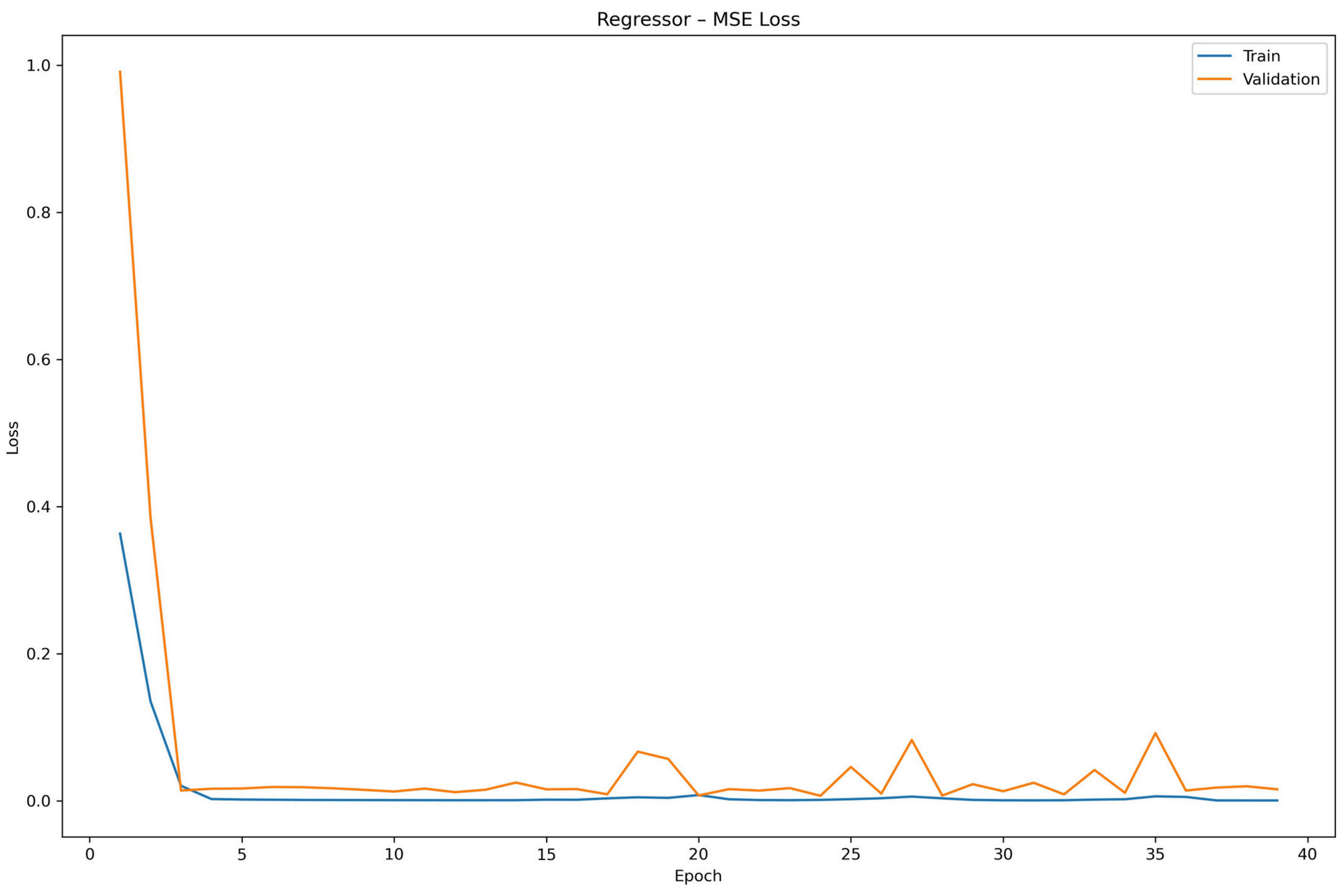
THREE-LAYER MULTILAYER PERCEPTRON IN PYTORCH

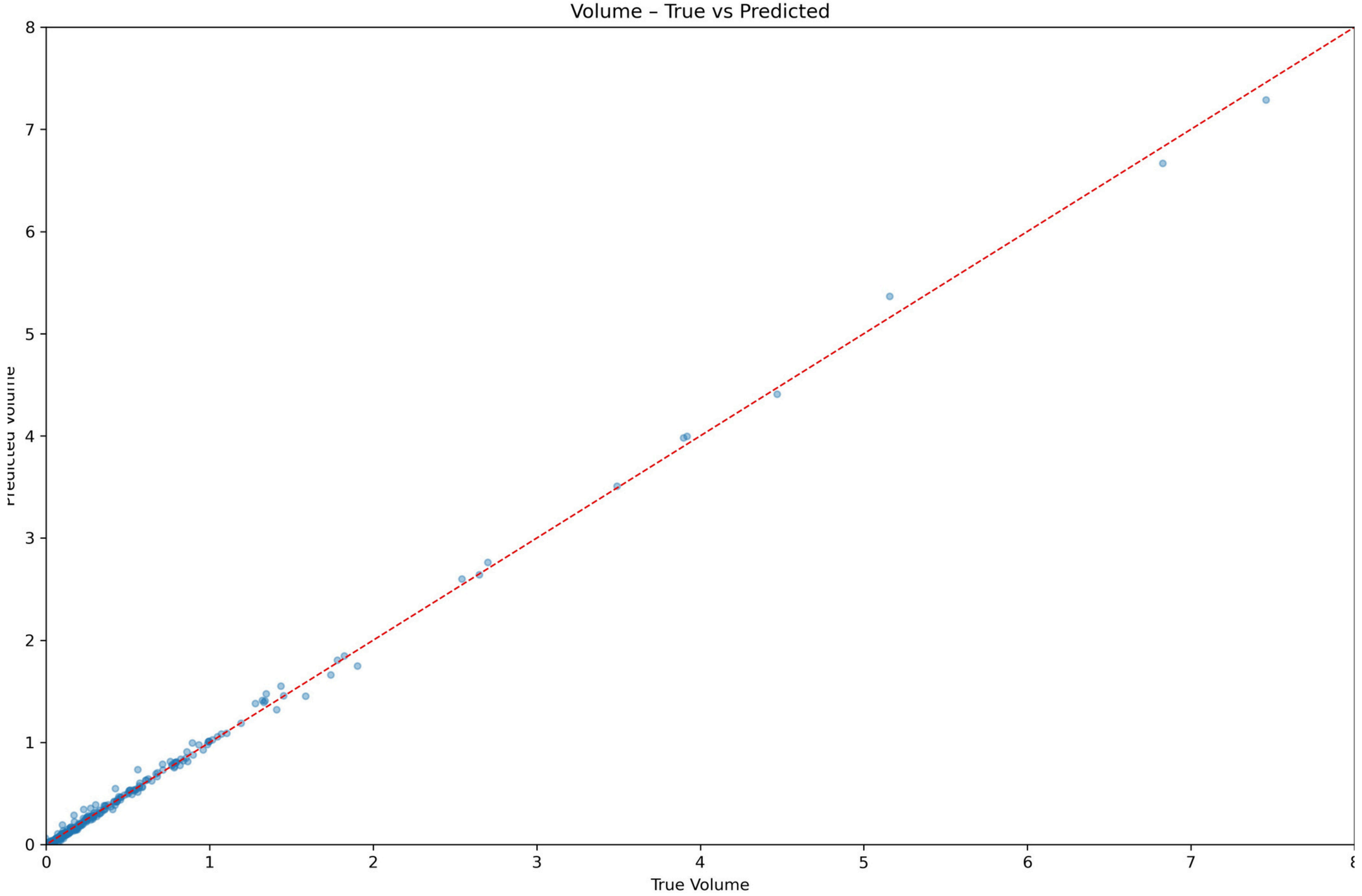
- Input features passed to first hidden layer with 64 neurons
- Activation function: ReLU
- Dropout rate: 30% (to reduce overfitting)
- Second hidden layer: 32 ReLU-activated neurons
- Final output layer matches number of symmetry classes (space groups)
- Performance
  - Accuracy: 59.9%
  - Weighted F1: 0.557
  - Macro-averaged F1: 0.210


Classifier – Cross-Entropy Loss
Train
Validation
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
Loss
0
20
40
60
80
100
120
140
Epoch

VOLUME PREDICTION MODEL (UNIT CELL VOLUME REGRESSION)

- Same architecture as classifier, but:
  - No dropout
  - Single output neuron for continuous prediction
- Structure: Input (9) → 64 → 32 → 1
- ReLU activations on hidden layers
- Trained with:
  - Adam optimizer (lr = 1e-3)
  - Batch size = 64
  - Early stopping (~epoch 35)
- Performance:
  - RMSE: 0.062
  - MAE: 0.018
  - R^2: 0.9972


Regressor – MSE Loss
Train
Validation
Loss
Epoch
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40

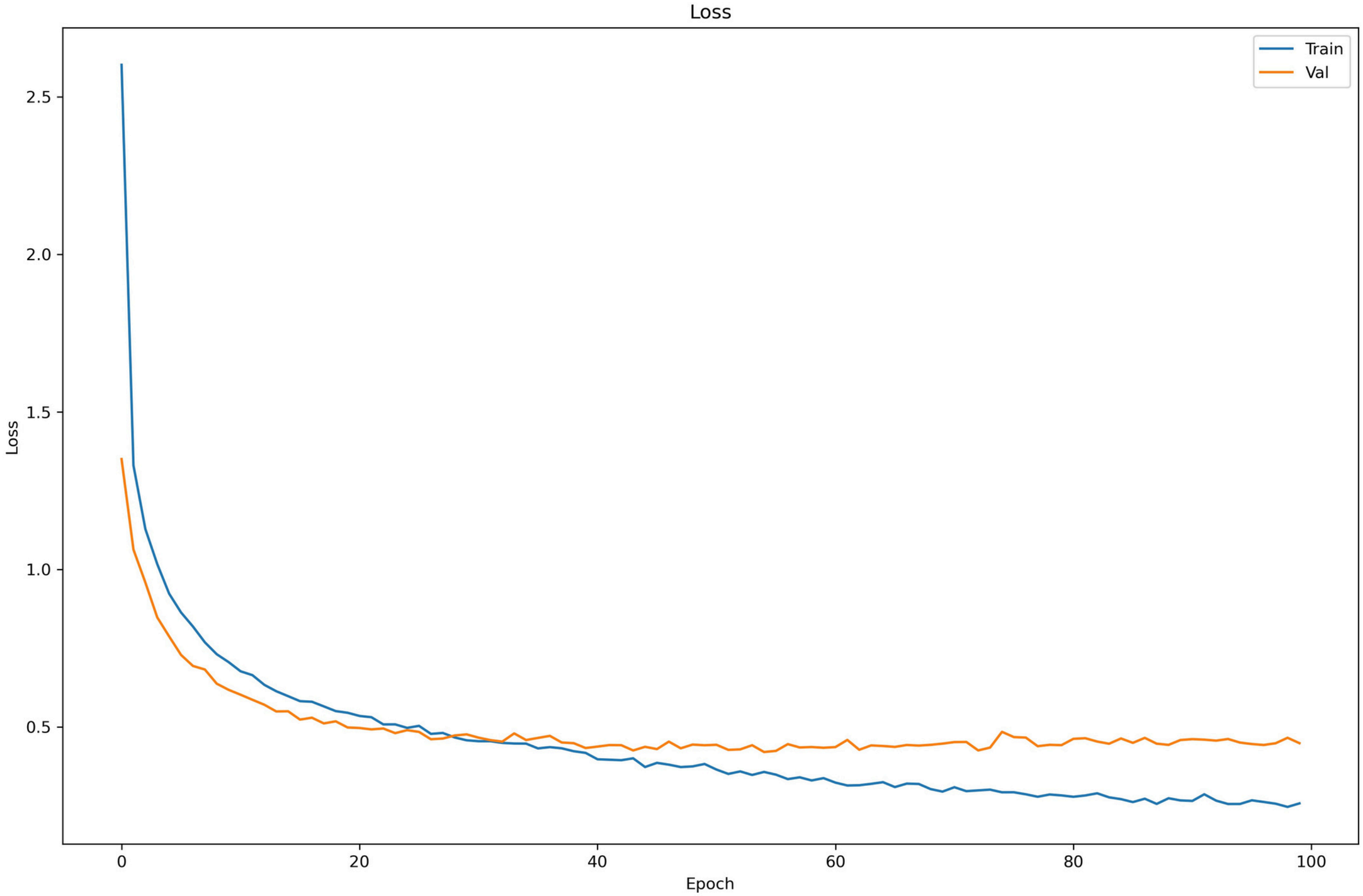

Volume – True vs Predicted
Predicted Volume
True Volume
0
1
2
3
4
5
6
7
8

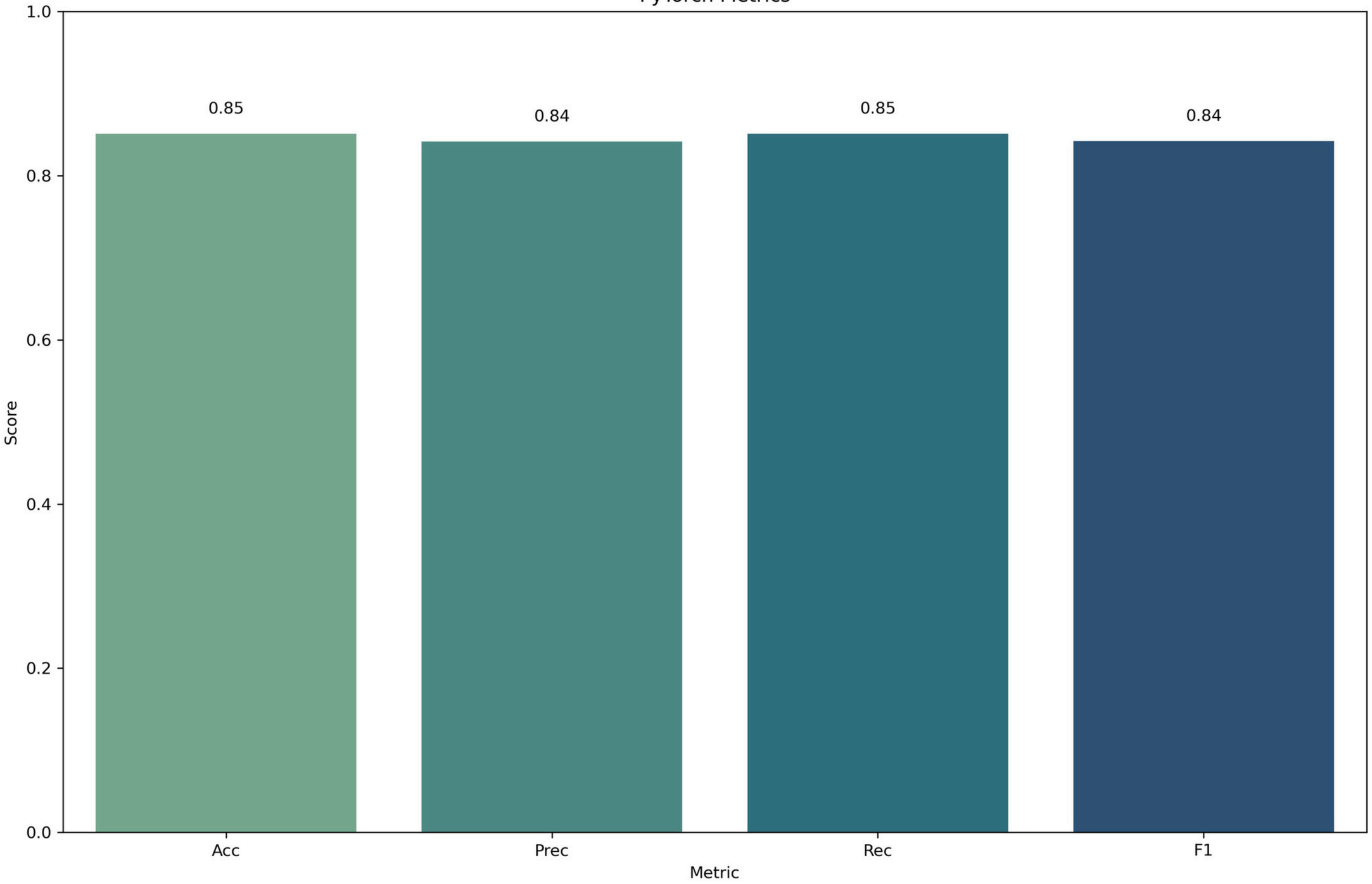
COMPOSITION-BASED SYMMETRY CLASSIFICATION

- Input: Chemical formula converted to vector via Matminer descriptors
  - Includes stoichiometry & elemental statistics

- Architecture (CRYSPNet-inspired):
  - Two-stage MLP:
    - Stage 1: Predicts Bravais lattice
    - Stage 2: Predicts space group, using lattice embedding + original features
  - Each stage: 128 → 64 units
    - ReLU + BatchNorm + Dropout (20%)

- Training Setup:
  - Cross-entropy loss, Adam optimizer (lr = 1e-3)
  - Batch size: 128
  - Early stopping after 10 epochs w/o improvement
- Results:
  - Precision: 0.85
  - Recall: 0.86
  - F1 Score: 0.85


Loss
Train
Val
2.5
2.0
1.5
1.0
0.5
Loss
0
20
40
60
80
100
Epoch

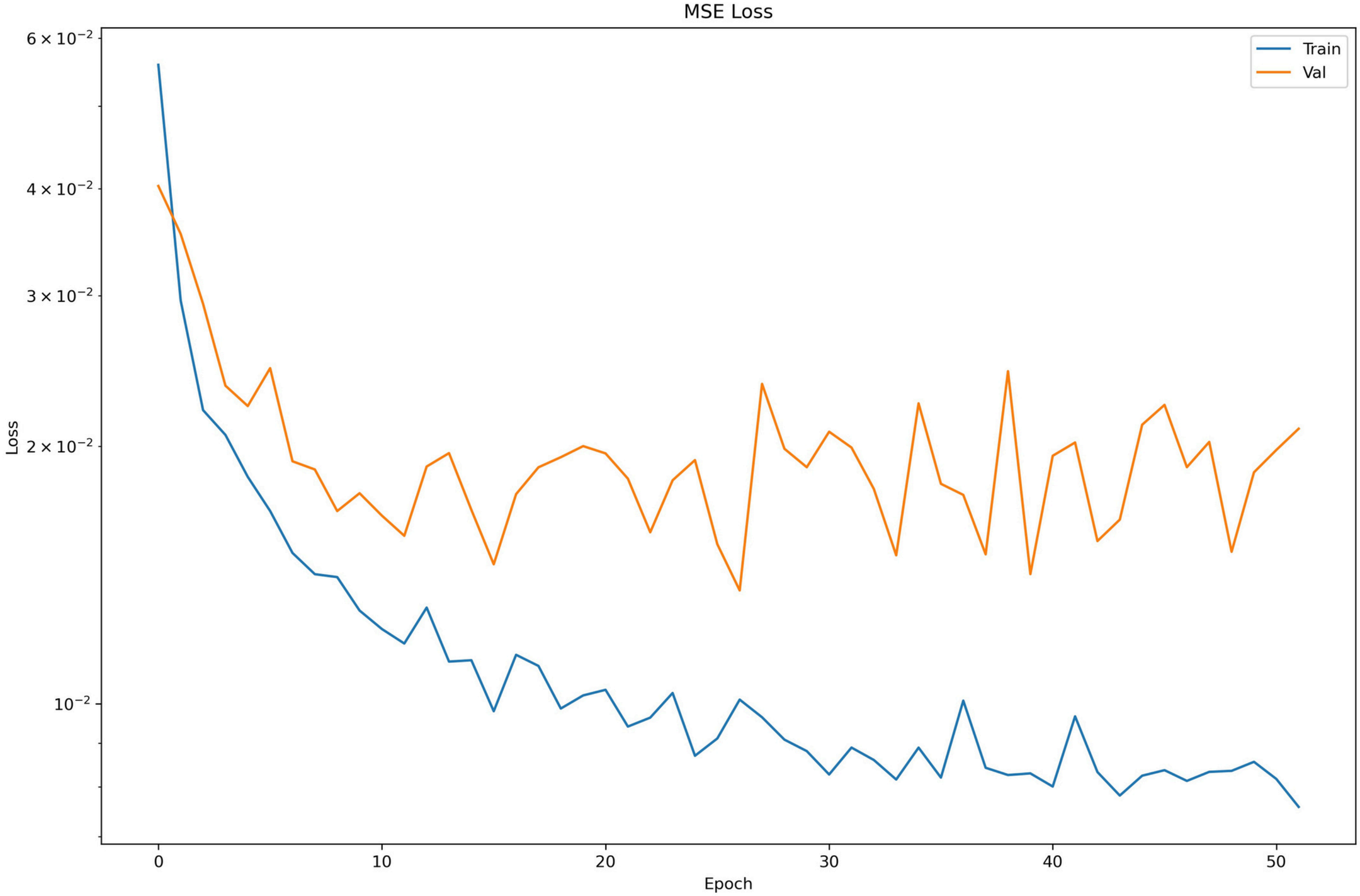

PyTorch Metrics
1.0
0.8
0.6
0.4
0.2
0.0
Score
0.85
0.84
0.85
0.84
Acc
Prec
Rec
F1
Metric

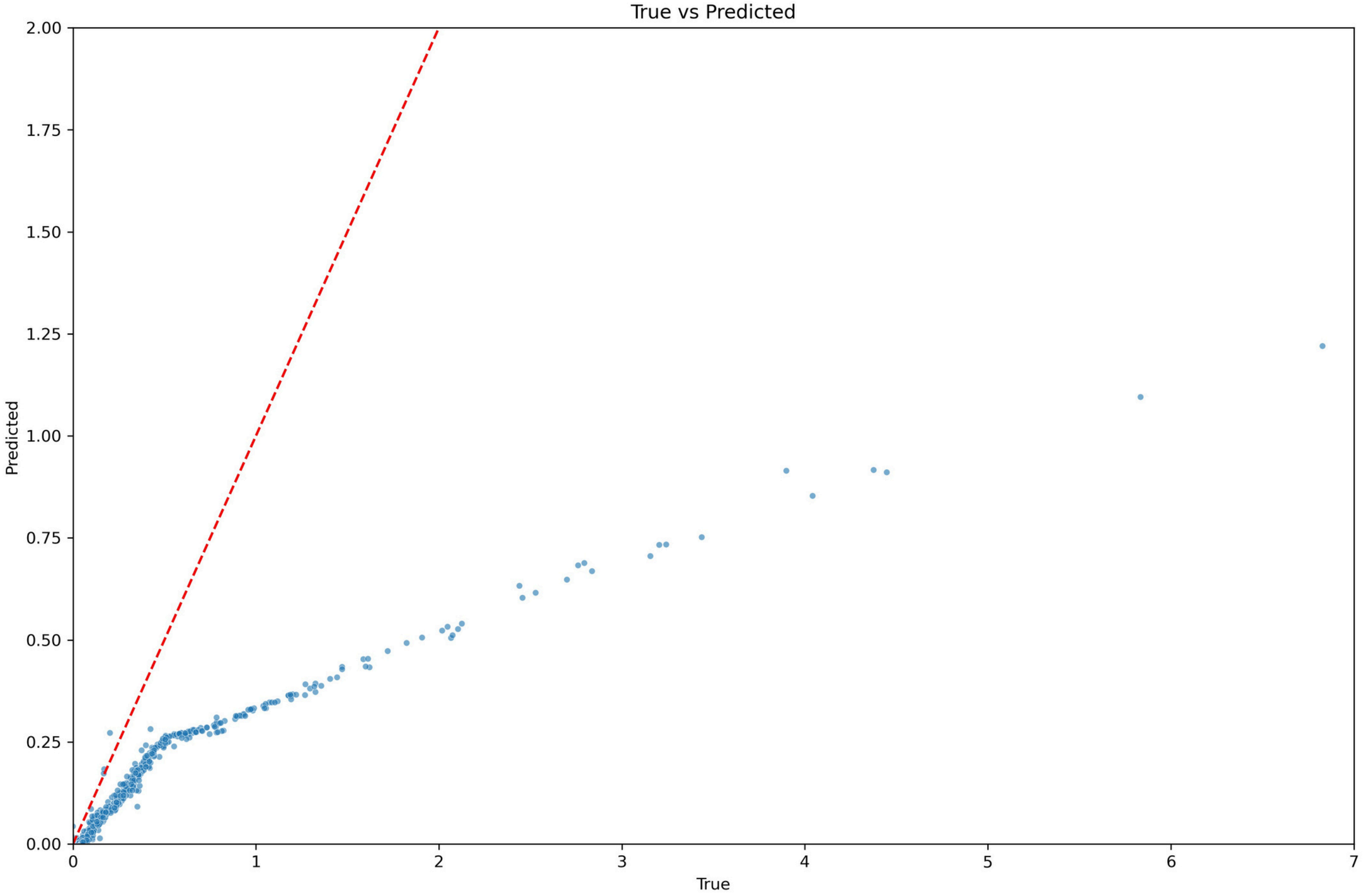
DEEP FEEDFORWARD VOLUME REGRESSION

- Goal: Predict volume using 6 primitive parameters + analytic volume

- Model Design:
  - Deep branch: 10-layer MLP
    - Layer sizes: 512 → … → 16
    - Each layer: Linear → BatchNorm → ReLU → Dropout (p = 0.2)
    - Includes SE-blocks for better feature recalibration
  - Final 16-D embedding + analytic volume → fusion layers (64 → 1)
  - Output: log(volume)

- Performance:
  - Train loss: < 0.015
  - Test MSE: 0.915
  - MAE: 0.121
  - R^2: 0.44


MSE Loss
Train
Val
Loss
Epoch
6×10^{-2}
4×10^{-2}
3×10^{-2}
2×10^{-2}
10^{-2}
0
10
20
30
40
50


True vs Predicted
Predicted
True
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
6
7

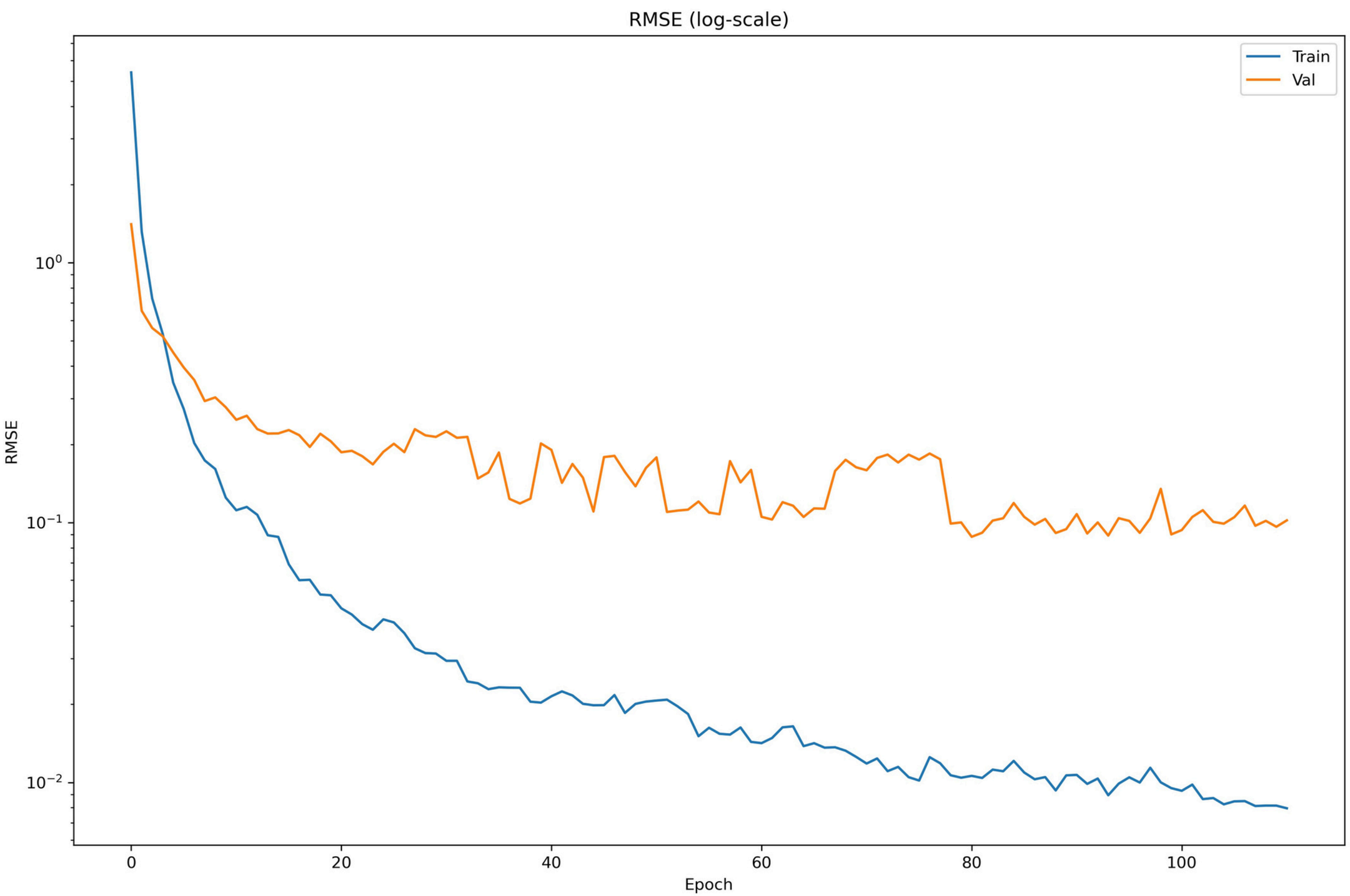
TABNET VOLUME REGRESSION

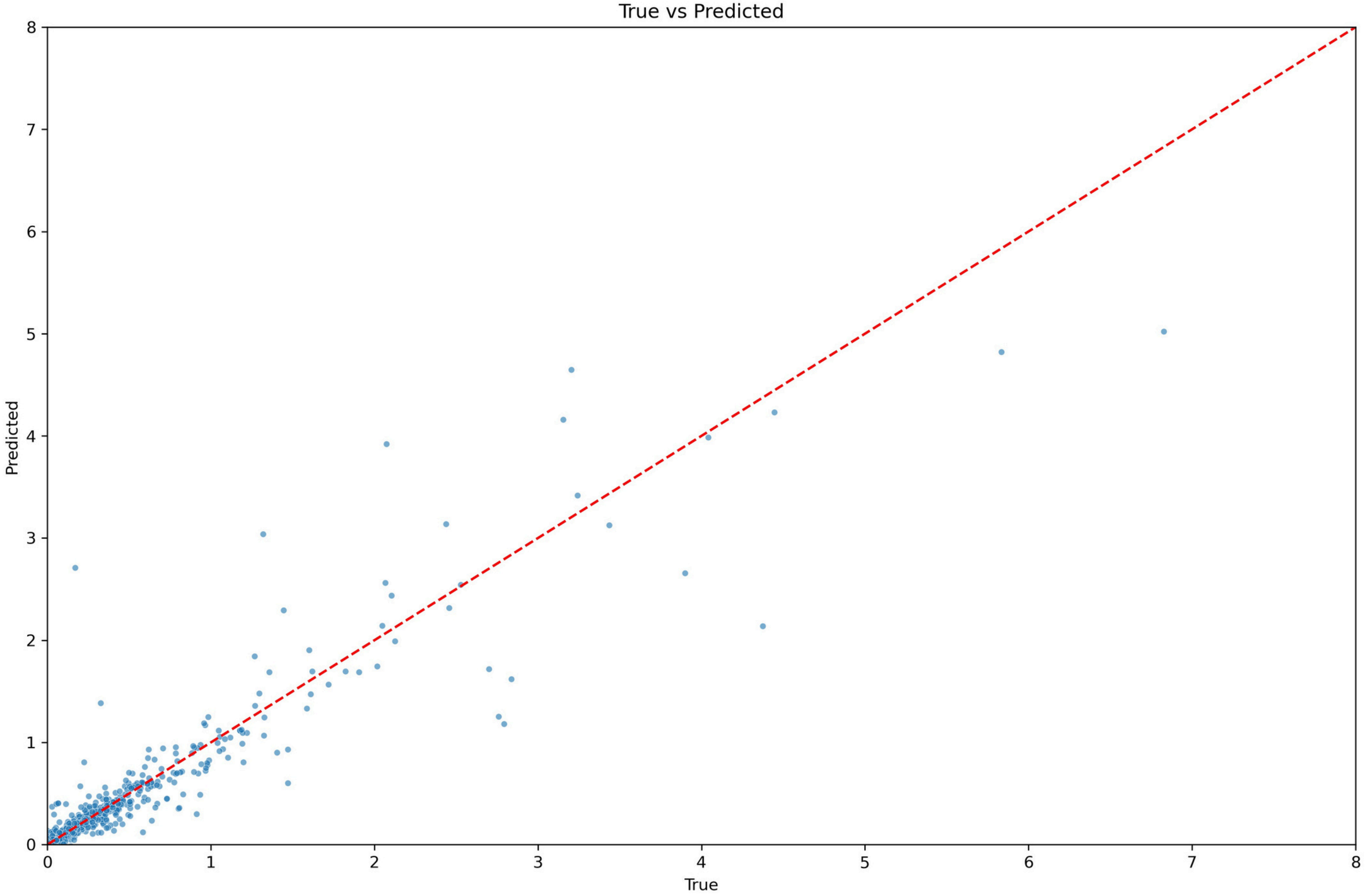
- Architecture:
  - Decision-step model with:
    - $n_d = n_a = 64$, $n_{steps} = 5$, $\gamma = 1.5$, $\lambda_{sparse} = 1e{-}4$

- Training Setup:
  - Adam optimizer, lr = 1e-3
  - StepLR schedule: step size = 50, $\gamma = 0.9$
  - Batch size: 1024 (virtual: 128)
  - Loss on log(1 + volume)
  - Early stopping: 30 epochs without RMSE improvement

TABNET VOLUME REGRESSION

- **Training Results:**
  - **Training RMSE: < 0.007**
  - **Validation RMSE: ~0.056 (at ~170 epochs)**
- **Test Performance:**
  - **MSE: 0.122**
  - **MAE: 0.051**
  - **R^2: 0.9255**
- **Insights:**
  - **Major contributing features: b, c, a, α**
  - **v formula added consistent but smaller value**
  - **Predictions closely matched true values; residuals were symmetric and stable**


RMSE (log-scale)
Train
Val
RMSE
10^{0}
10^{-1}
10^{-2}
0
20
40
60
80
100
Epoch


True vs Predicted
Predicted
True
0
1
2
3
4
5
6
7
8

REFERENCES

Yanchao Wang and Yanming Ma. Perspective: Crystal structure prediction at high pressures. The Journal of chemical physics, 140(4), 2014.

MP Shaskolskaya. Crystallography: Manual for institutes of higher education. Higher School, pages 10–14, 1984.

Artem R Oganov, Andriy O Lyakhov, and Mario Valle. How evolutionary crystal structure prediction works-and why. Accounts of chemical research, 44(3): 227–237, 2011.

Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. The Journal of chemical physics, 124(24), 2006.

Haotong Liang, Valentin Stanev, A Gilad Kusne, and Ichiro Takeuchi. Cryspnet: Crystal structure predictions via neural networks. Physical Review Materials, 4(12):123802, 2020.

Saulius Gra˘zulis, Daniel Chateigner, Robert T Downs, AFT Yokochi, Miguel Quir´os, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography open database–an open-access collection of crystal structures. Applied Crystallography, 42(4):726–729, 2009.

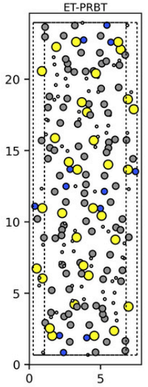
Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learn- ¨ ing. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 6679–6687, 2021.

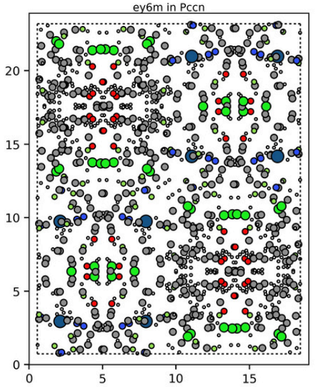
THANK YOU

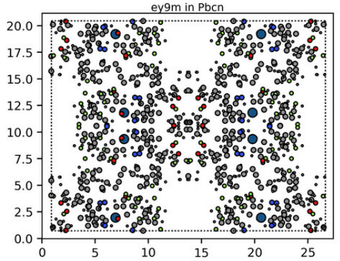
ANY QUESTIONS?

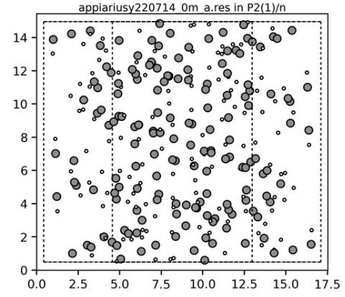
APENDIX

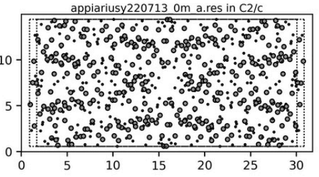
Crystal Structures

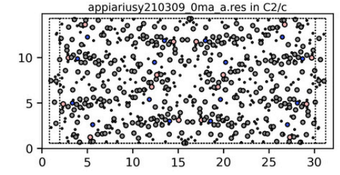


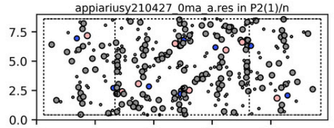


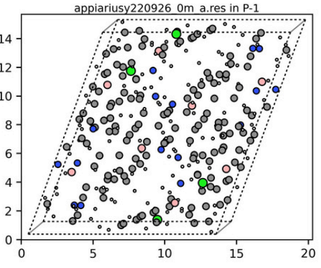


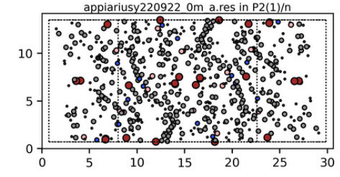




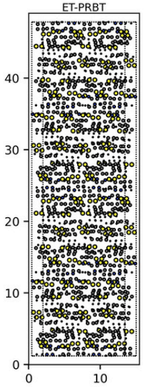


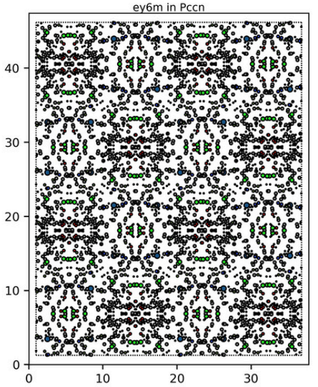


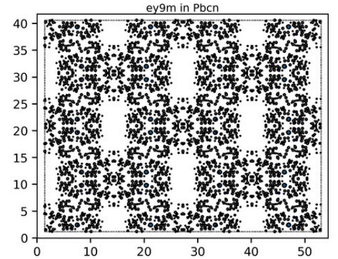


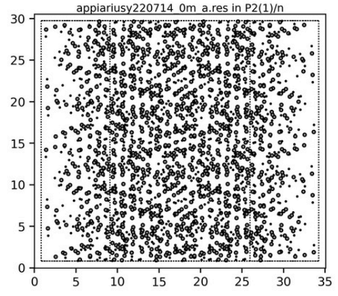


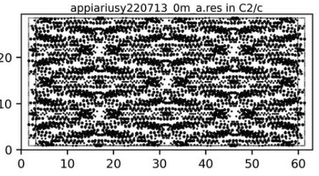
Supercells (2x2x2)

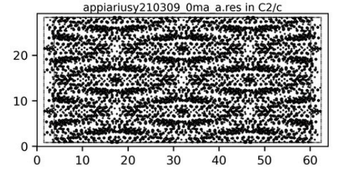


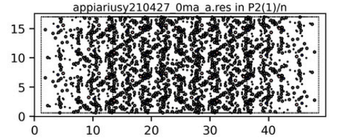


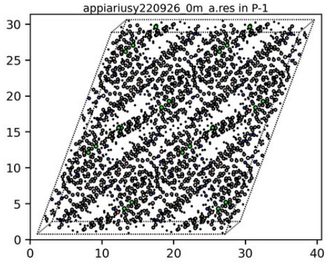


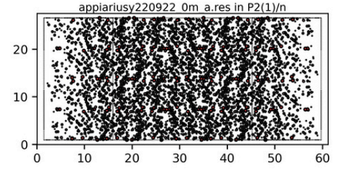


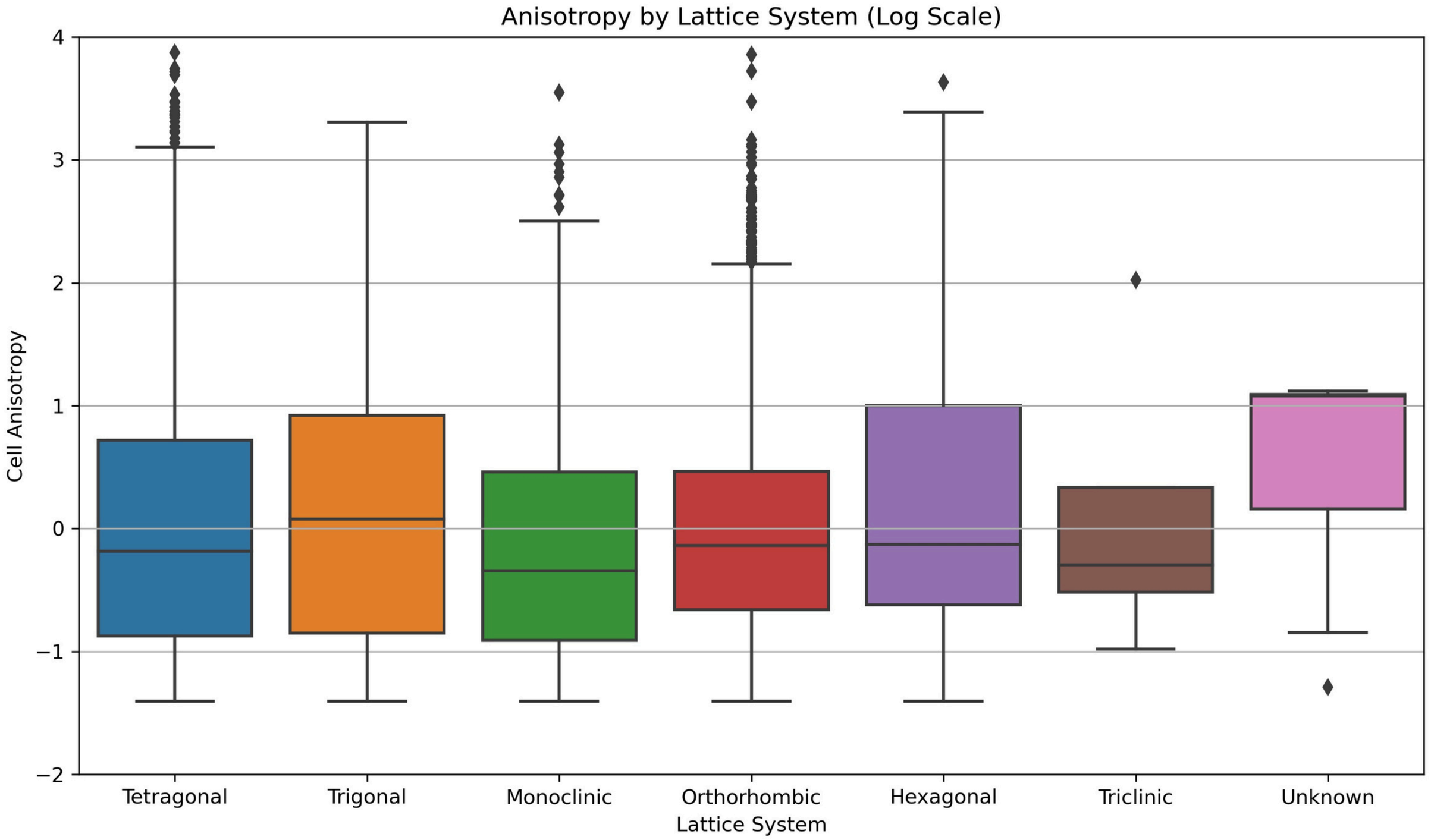


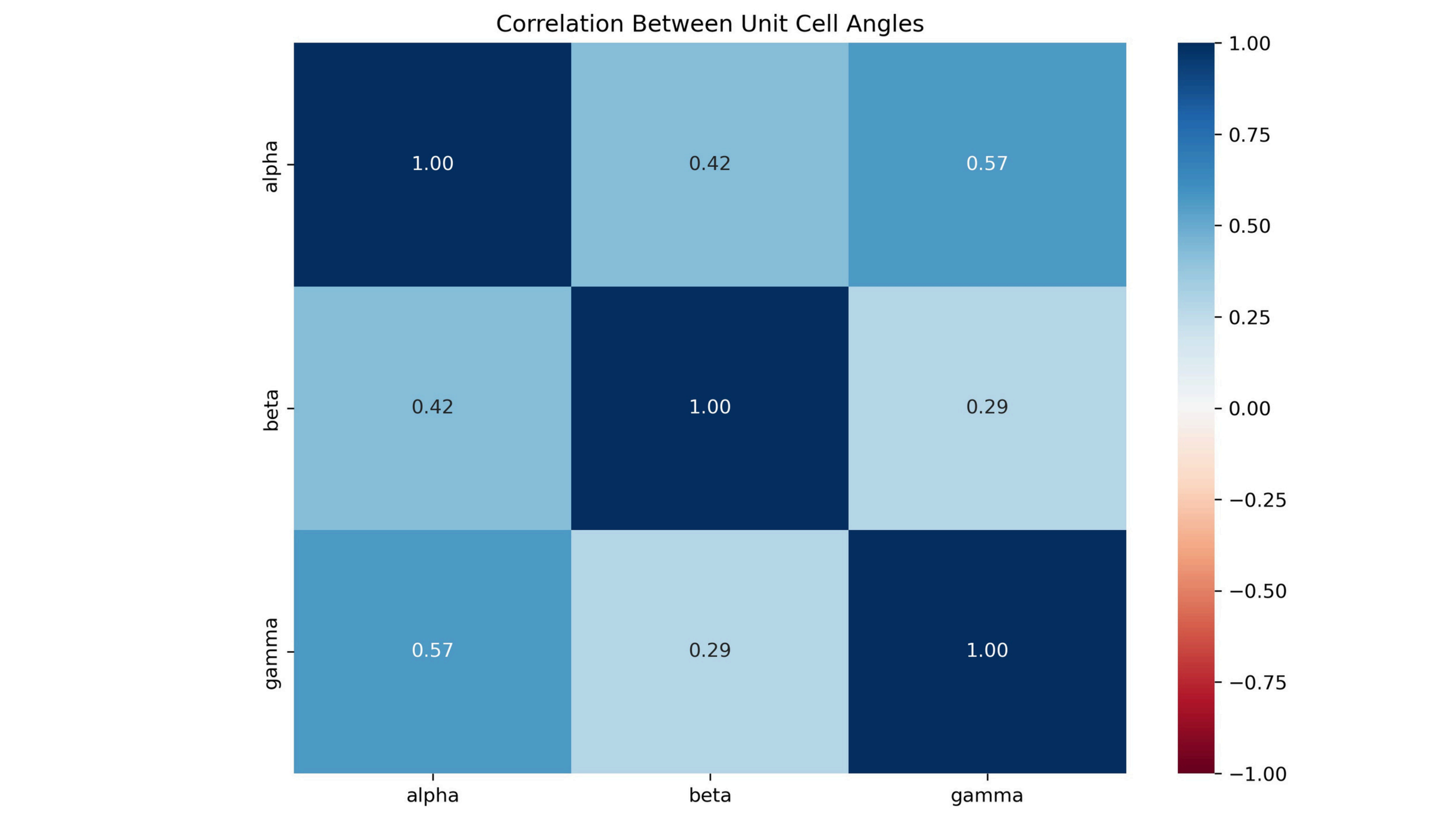


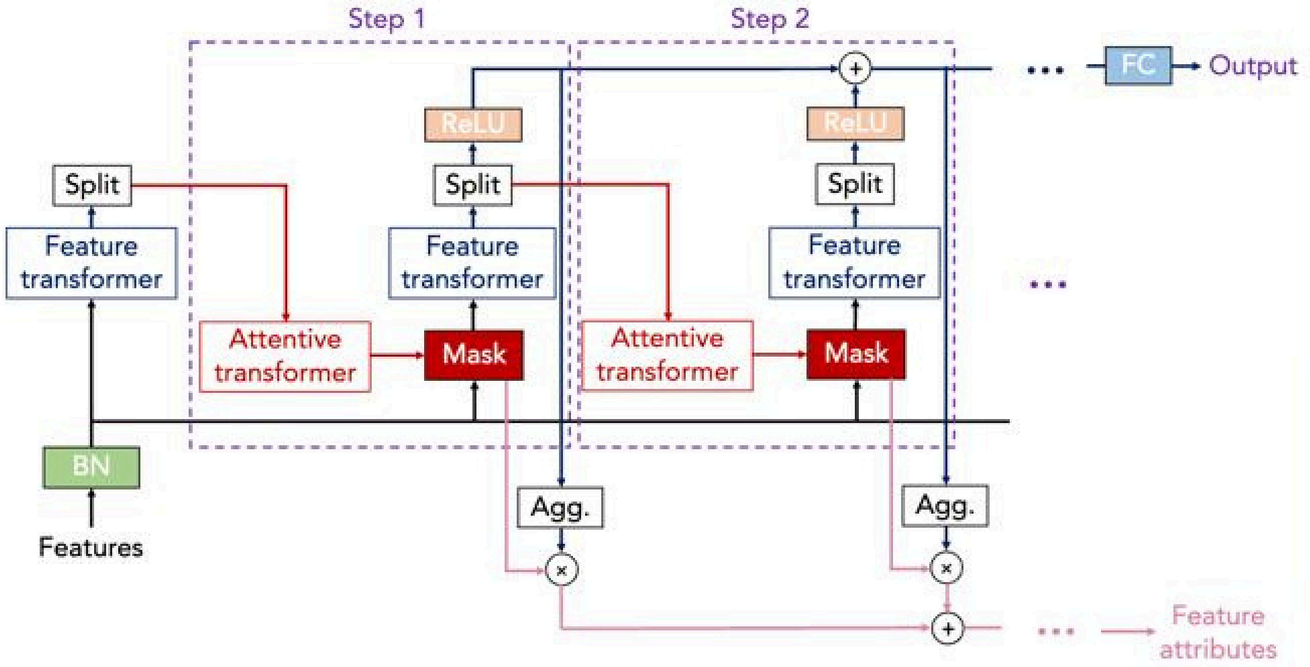




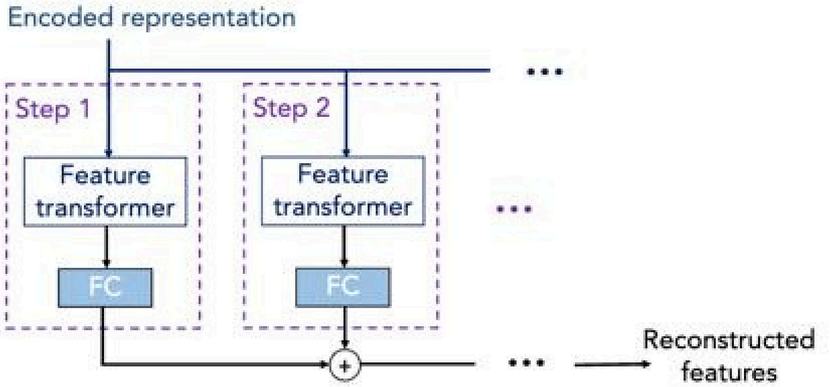



Anisotropy by Lattice System (Log Scale)
Cell Anisotropy
4
3
2
1
0
−1
−2
Tetragonal
Trigonal
Monoclinic
Orthorhombic
Hexagonal
Triclinic
Unknown
Lattice System

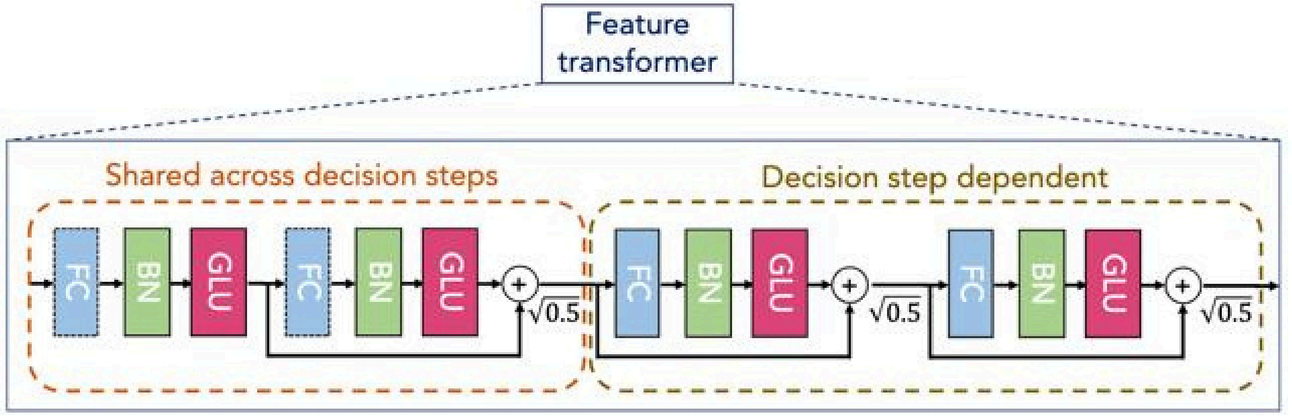

Correlation Between Unit Cell Angles
alpha
beta
gamma
1.00
0.42
0.57
0.42
1.00
0.29
0.57
0.29
1.00
alpha
beta
gamma
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



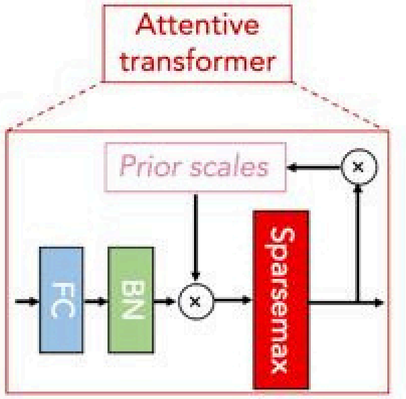
(a) TabNet encoder architecture



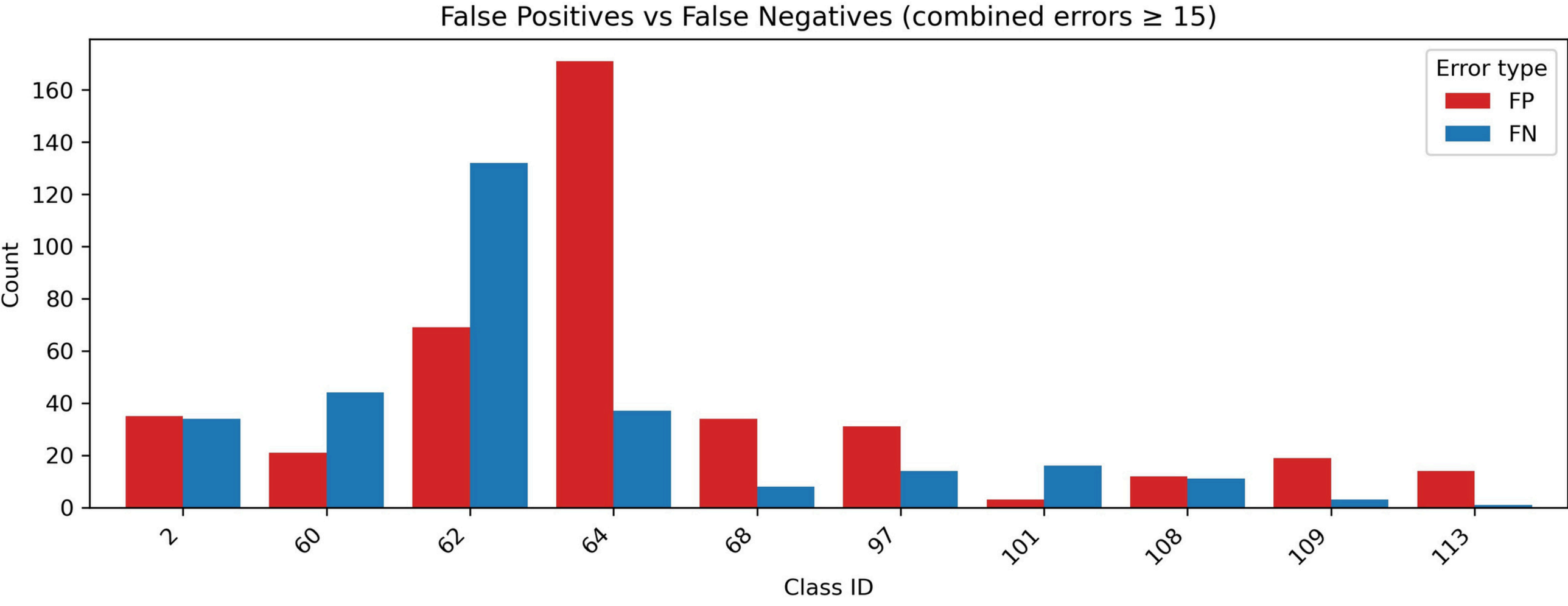
(b) TabNet decoder architecture



(c)



(d)


False Positives vs False Negatives (combined errors ≥ 15)
Error type
FP
FN
Count
0
20
40
60
80
100
120
140
160
2
60
62
64
68
97
101
108
109
113
Class ID