Crystal Structure Prediction

Leonid Sarkisyan

Vram Papyan

Supervisor: Aleksandr Hayrapetyan

Submitted for the
degree of

BS Data Science
College of Science and Engineering
American University of Armenia

Yerevan 2025

Abstract

Here, we illustrate an end-to-end data-driven crystal structure analysis from geometric and numerical characteristics obtained from Crystallographic Information Files (CIFs). Using machine learning algorithms, the primary aim was to forecast crystalline materials' space group symmetry class (numerically represented). We extracted a rich dataset from the Crystallography Open Database (COD) (*Gražulis et al. (2009)*), including structural attributes such as unit cell dimensions, angles, volume, and designed aspects such as cell anisotropy and shape factor. Having applied strict preprocessing, we experimented with and compared a pool of standard classifiers like Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbors, AdaBoost, and Support Vector Machine based on cross-validation and various measures of performance. To extend the modeling framework further, we used deep neural networks to address classification and regression tasks such as unit cell volume prediction based on geometric descriptors. We further used TabNet to learn nonlinear relationships with the interpretability of features. We devised a CRYSPNet-inspired design for directly classifying space groups from chemical composition using Matminer-derived descriptors. Our results show that symmetry and volume can be inferred reliably from composition and structure inputs, with ensemble and neural methods being the most accurate. This study emphasizes the value of integrating data science and crystallographic analysis, offering scalable techniques for symmetry classification and structure property prediction. Scripts and Visualisations can be found here: (Sarkisyan (2025))

Keywords: *Crystallography, Machine Learning Applications in Material Science, Deep Learning Applications in Material Science, Crystal Structure Analysis, Data-Driven Materials Science*

Contents

Chapter 1

Introduction

Crystallography is vital in solid-state physics, chemistry, and materials science since it reveals the internal atomic structure in crystalline solids Shaskolskaya (1984). Symmetry is among the simplest properties of a crystal and is typically characterized by its space group. Accurately determining space group symmetry is significant for understanding material properties, synthesis guidance, and computational modeling workflow support.

Traditionally, symmetry determination comes through diffraction experiments and manual or semi-automated interpretation. While effective, they are data-intensive, time-consuming, and sensitive to noise or missing data. With the ongoing growth in crystallographic data facilitated by open-access databases such as the Crystallography Open Database (COD), there is more reason to pursue scalable, automated, and data-driven methods of symmetry classification.

In this work, we propose a machine learning approach for predicting space group symmetry of crystalline compounds from structural descriptors directly extracted from CIFs. We examine how much geometric and numerical descriptors contain symmetry-related information, including unit cell dimensions and angles, and derived quantities such as anisotropy and shape factor. We then apply a conventional classification model suite to assess their ability to learn symmetry classes from these descriptors. Along with classification, we extend our analysis to the regression of physical properties, particularly unit cell volume, through traditional and deep learning methods. We also present an architecture derived from CRYSPNet that predicts symmetry classes directly from chemical composition

based on the Matminer library-derived features. Combined, these features represent a multi-modal, holistic approach to analyzing crystallographic structure and demonstrate the applicability of geometry- and composition-based representation to making accurate and interpretable predictions of crystallographic properties.

Figure 1.1: Photograph of a natural amethyst crystal cluster showcasing well-defined prismatic geometry and inherent structural symmetry. Such visible macroscopic order reflects the underlying periodic atomic arrangement that machine learning models in this study aim to classify through CIF-derived features. (*Iacob (2020)*)

Chapter 2

Literature Review

The field of crystal structure prediction (CSP) has transitioned from purely experimental methods to more sophisticated computational and machine-learning approaches. CSP has traditionally dealt with the sheer configurational space and intricate relationships between atomic structures and symmetry groups. This section provides an overview of some influential developments that have informed our data-driven approach to symmetry classification.

2.1 Evolutionary Algorithms and Global Optimization

Oganov and Glass (2006) suggested a novel approach to CSP by evolutionary algorithms, a radical departure from trial-and-error experimentation towards data-guided exploration. The method mimics evolutionary processes, selection, crossover, and mutation, to guide candidate crystal structures toward lower free-energy states step by step. Underlying their model are symmetry operations and physical constraints that limit the search space and enhance convergence. The result is an extremely efficient system that can identify stable and metastable structures in a wide range of compositions and pressures. This methodology was applied to the popular USPEX code (*Oganov and Glass (2006)*), which has since been used in numerous high-throughput materials discovery projects. Specifically, the authors emphasized the relevance of physically meaningful parameters such as unit cell size (a, b, c, alpha, beta, gamma) and atomic packing efficiency

in guiding the algorithm toward acceptable solutions. Such thought directly impacts our work, where such descriptors are used as input features to supervised learning algorithms for symmetry classification.

2.2 Chemical and Physical Intuition in Structure Classification

Wang and Ma (2014) provided a broad and critical survey of CSP strategies, emphasizing the role of chemical intuition as well as empirical statistical methods to analyze symmetry relationships. According to their review, several heuristic models are defined based on previous tendencies, e.g., the tendency of certain molecular shapes or functional groups to pack in certain space groups (e.g., non-centrosymmetric vs. centrosymmetric systems). They also mapped out the limitations of such rule-based approaches to molecular diversity and crystallographic complexity. One valuable point they made is that while specific patterns (e.g., networks of hydrogen bonds or molecular shape) accompany particular types of symmetry, the absence of a quantitative foundation limits generalizability. This gap allowed for the creation of data-driven approaches like ours to try to extract statistical patterns from big databases using formal machine learning tools. In addition, they underscored the increasing significance of open crystallographic databases such as COD in enabling reproducible and large-scale CSP research datasets, which we also use in this study.

2.3 Machine Learning Models for Space Group Prediction

The CRYSPNet framework (*Liang et al. (2020)*) represents a significant milestone in machine learning for CSP applications. Unlike the majority of conventional CSP methods that are highly reliant on quantum mechanical computation or structural heuristics, CRYSPNet formulates space group prediction as a supervised learning task. Using chemical compositions (elemental ratios) as inputs, the model applies a deep neural network to simultaneously predict three struc-

tural descriptors: Bravais lattice type, space group, and lattice parameters. The work demonstrated that CRYSPNet can achieve over 90% accuracy on a variety of datasets, highlighting the promise of symmetry classification without complete structural information. However, because they only use compositional features, the model does not involve physical descriptors in terms of unit cell geometry, so how much independently contributing structural information can enhance predictive accuracy is a question left open. Our study addresses this question by using only geometric and calculated features (e.g., anisotropy, shape factor) to establish how much symmetry can be deduced from crystallographic structure alone.

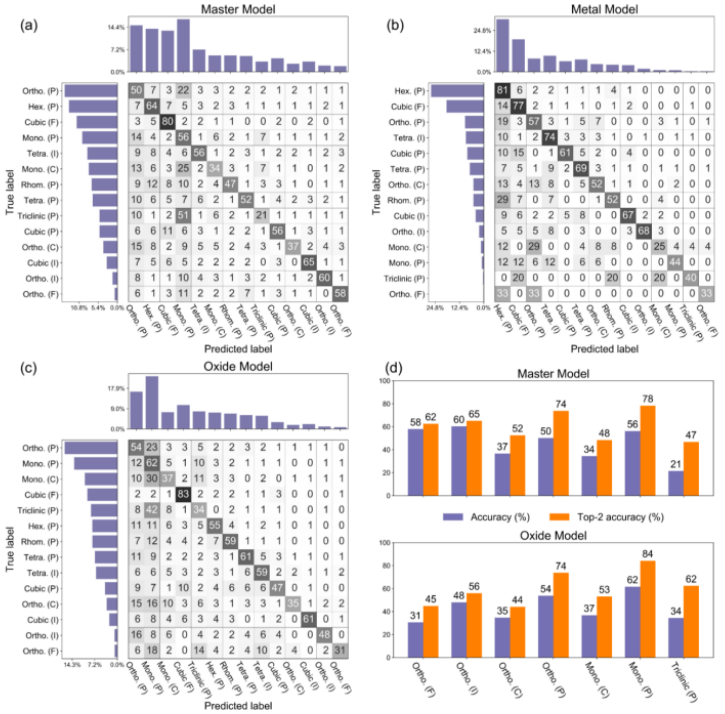


Figure 2.1: Performance of the model predicting the Bravais lattices: (a) Confusion matrix of the Master Model; (b) Confusion matrix of the Metal Model; (c) Confusion matrix of the Oxide Model. The grey shade corresponds to the relative density of the true labels (also provided as percentages). The numbers on the diagonal shows the prediction accuracy of each Bravais lattice. Histograms on the left and on the top represent the distribution of the true and predicted labels, respectively. (d) The (top-1) accuracy and top-2 accuracy of the Master Model (top) and Oxide Model (bottom). The abbreviations are "Hex." for hexagonal, "Rhom." for rhombohedral, "Tetra." for tetragonal, "Ortho." for orthogonal, "Mono." for monoclinic. "P", "I", "C", and "F" denote primitive, body-centered, base-centered with unique c-axis, and face-centered system, respectively. (*Liang et al. (2020)*)

2.4 Evolutionary Algorithms and USPEX Method

Oganov et al. (2011) gave a wide-ranging theoretical and practical overview of CSP using evolutionary algorithms, i.e., the USPEX method. The algorithm addresses two key CSP issues: crossing the vast configuration space (search) and accurately determining relative structure energies (ranking). The evolutionary algorithm converges into low-energy, high-symmetry arrangements by modeling natural selection processes and implementing processes like heredity, mutation, and permutation of atomic arrangements. Particularly, the authors refer to reducing the dimensionality of the energy landscape upon relaxation of the structure and enabling faster convergence. The work demonstrated that even very dimensional problems with dozens of atoms in the unit cell could be sampled reasonably with such biologically motivated methods. The paper also introduced the "quasientropy" diversity measure and how initialization based on symmetry and pseudosymmetry can prevent premature convergence to local minima. Interestingly, Oganov's method proved to be quite successful in discovering new, unexpected phases in high-pressure chemistry, confirming the general idea that physically meaningful geometric descriptors (like lattice parameters and anisotropy) are at the core of symmetry prediction tasks. This idea directly informs and justifies our research's focus on learning symmetry from numerical structure-derived features alone. Our contribution is rooted in this by formalizing the prediction task and evaluating the discriminative power of such features using modern machine learning methods.

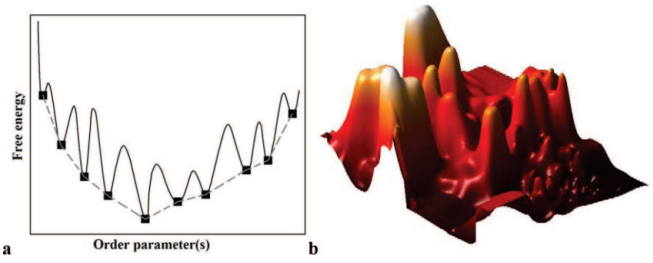


Figure 2.2: Energy landscape: (a) 1D scheme showing the full landscape (solid line) and reduced landscape (dashed line joining local minima); (b) 2D Projection of the reduced landscape of Au8Pd4, showing clustering of low-energy structures in one region. (*Oganov et al. (2011)*)

2.5 TabNet

TabNet (*Arik and Pfister (2021)*) is an end-to-end state-of-the-art neural architecture for tabular data that synergistically integrates instance-wise sparse attention on features and deep feature transformers. At a series of different decision steps, TabNet learns a sparse choice mask (by sparsemax) that picks just the most prominent inputs, processes them with shared and step-dependent gated linear units, and mixes their results into the final prediction. This design supports both high prediction accuracy, generally comparable or better than gradient-boosted trees, as well as inbuilt interpretability since the learned masks themselves yield insight into what features drove each prediction. TabNet also enables self-supervised pre-training using masked feature reconstruction, with further performance gain when little labeled data is present.

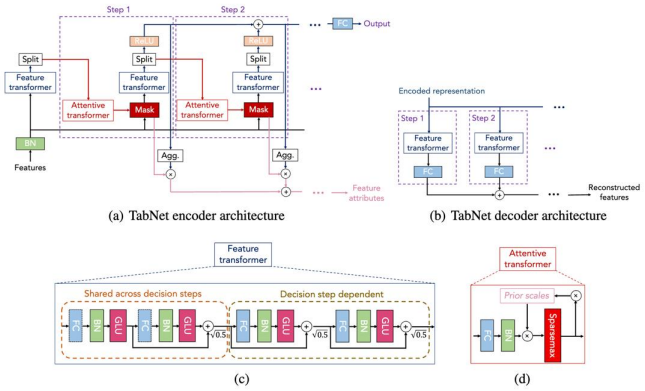


Figure 2.3: (a) TabNet encoder comprises a feature transformer, an attentive transformer, and feature masking. A split block divides the processed representation to be used by the attentive transformer of the subsequent step, as well as for the overall output. For each step, the feature selection mask provides interpretable information about the model's functionality, and the masks can be aggregated to obtain global feature important attribution. (b) TabNet decoder, composed of a feature transformer block at each step. (c) A feature transformer block example – 4-layer network is shown, where 2 are shared across all decision steps and 2 are decision step-dependent. Each layer is composed of a fully-connected (FC) layer, BN and GLU nonlinearity. (d) An attentive transformer block example – a single layer mapping is modulated with a prior scale information which aggregates how much each feature has been used before the current decision step. sparsemax (Martins and Astudillo 2016) is used for normalization of the coefficients, resulting in sparse selection of the salient features.(*Arik and Pfister (2021)*)

Chapter 3

Methodology

3.1 Data Source and Acquisition

We acquired our dataset from the COD public database of over a thousand CIFs. A CSV metadata file in structured format was utilized to identify CIF file names, which were then downloaded in parallel with the help of a homegrown script. Each .cif file contains important structural parameters such as unit cell dimensions, angles, volume, symmetry group, and chemical metadata. Below are depictions of Crystal Structures from various perspectives, such as a 2x2x2 Supercell and a top-down view.

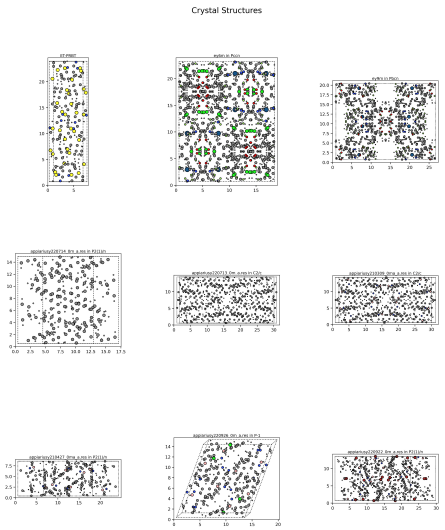


Figure 3.1: Visualization of unit cell crystal structures. Each subplot represents a CIF-based atomic arrangement using different compounds and symmetries. Atom colors represent different element types or roles in the structure.

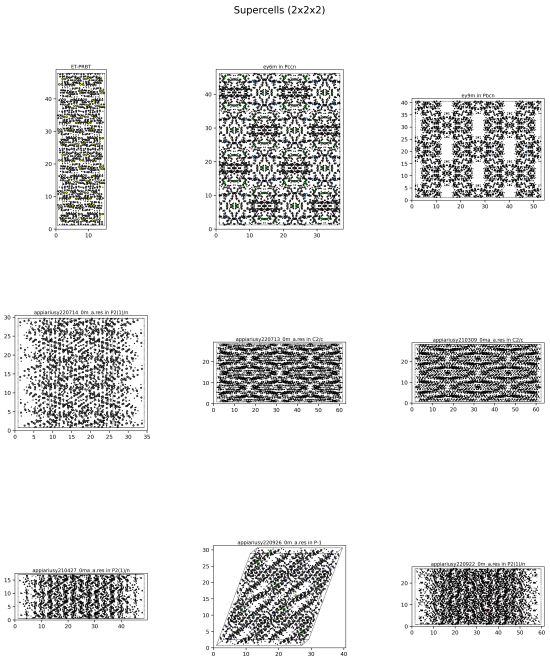


Figure 3.2: 2x2x2 supercells generated from the original unit cells. These expanded views reveal periodic atomic packing and provide insights into crystalline regularity and defect propagation.

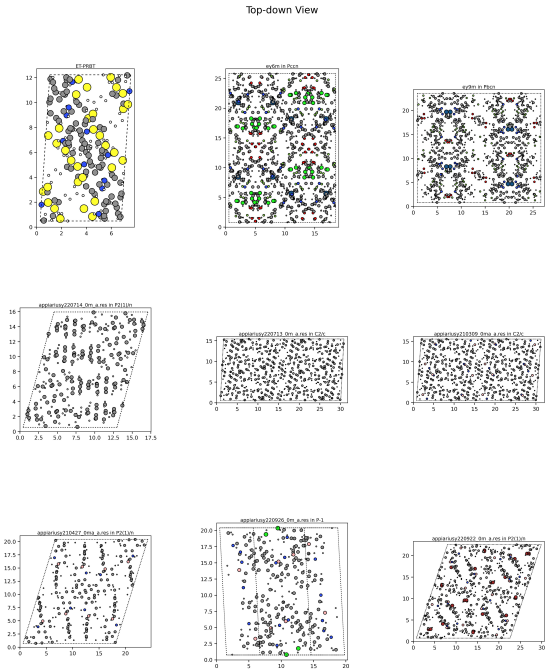


Figure 3.3: Top-down view of each structure along the c-axis. This perspective highlights symmetry within planes, spatial repetition, and possible layering effects that aren't visible in side views.

The plots below show the bond length distribution in different crystal structures. Each of the histograms gives a representation of internal atomic connectivity and variability of the bond environment. The distribution indicates structural variety and assists in the numerical features applied to predict symmetry.

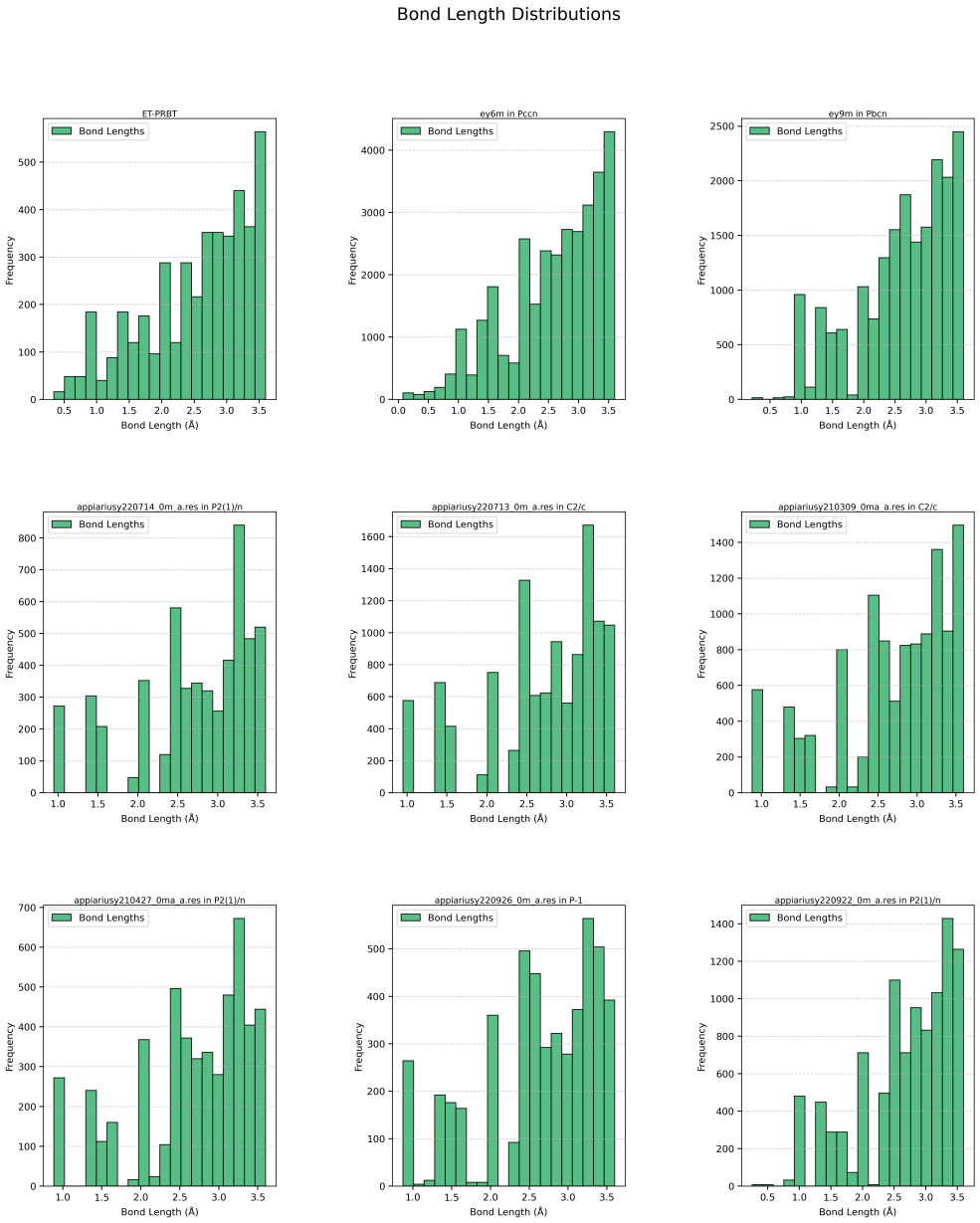


Figure 3.4: Histogram grid shows bond length variation across CIF structures.

The plots below illustrate the elemental composition of selected crystal structures. The bar chart in each expresses the number of atoms per element and captures chemical diversity and stoichiometry. These compositions serve as input to compositional symmetry prediction.

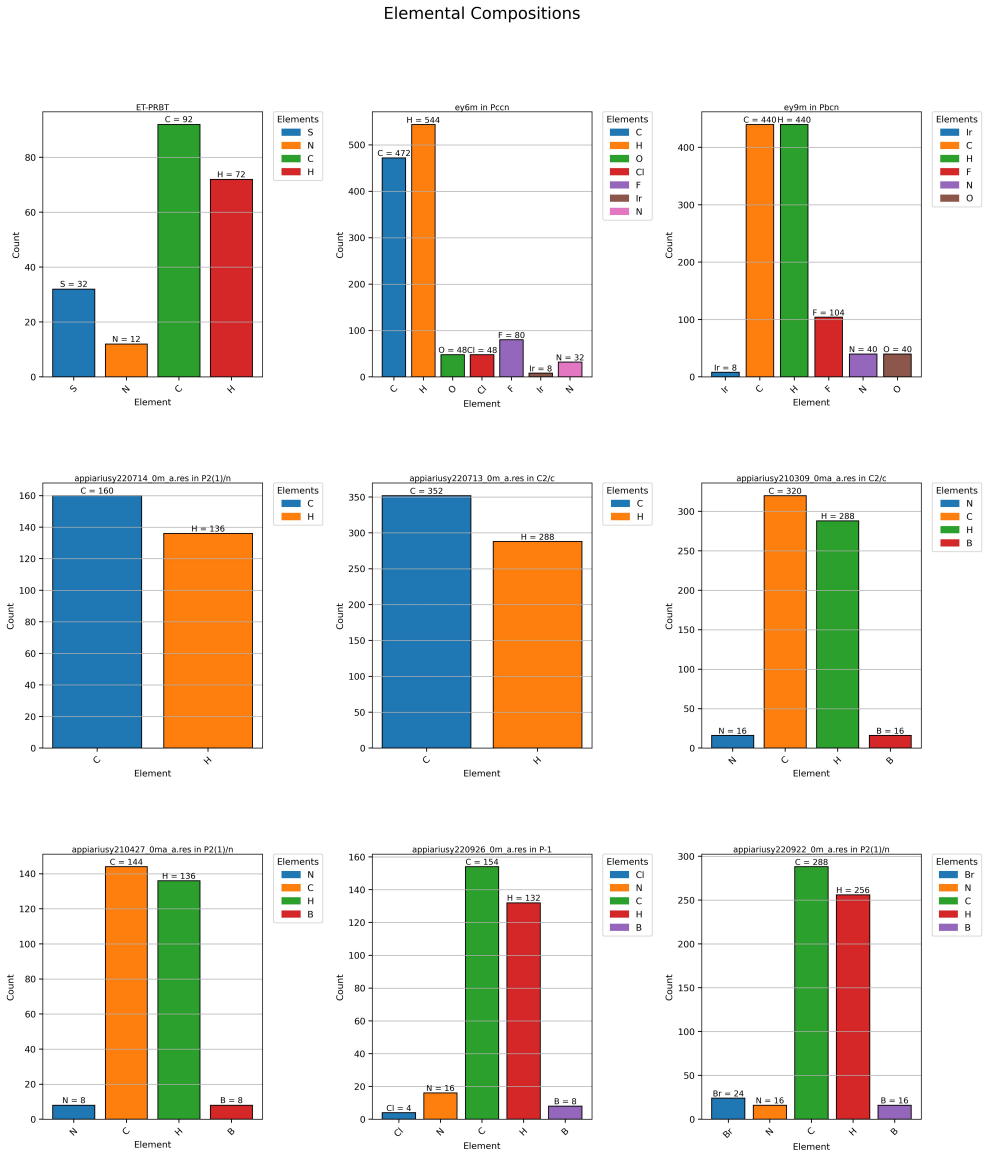


Figure 3.5: Visualization of the atomic counts of elements present in each structure.

We also graphed unit cell dimensions and angles for different CIF samples to enable structural feature engineering. The following bar plots indicate axis lengths (a, b, c), and the following charts indicate angular measurements (α, β, γ). These visual checks confirmed the variation in geometric arrangements, which made it reasonable to include these descriptors in our model training pipeline.

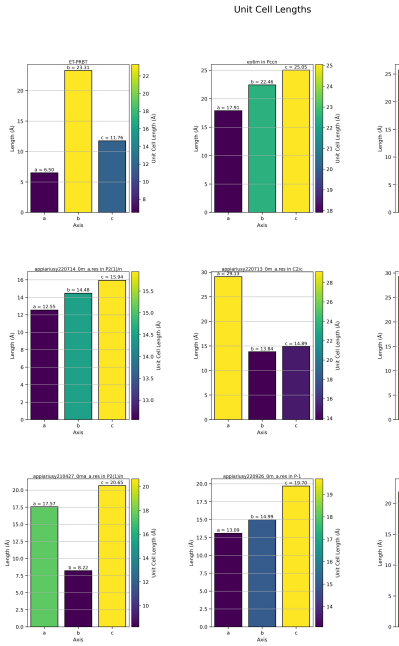


Figure 3.6: Bar plots of a, b, and c show variation in unit cell sizes.

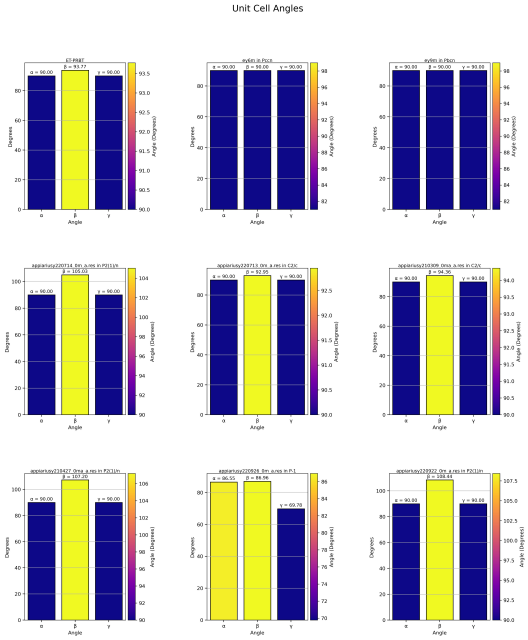


Figure 3.7: Charts of α, β, and γ angles across the same structures. Most entries are orthogonal (90°), with some distortions highlighted in color.

Radial Distribution Functions (RDFs) were plotted in order to study atomic pair distributions as a function of distance in both crystals. These g(r) curves provide short- and long-range order, revealing local structure regularities that are not captured by cell parameters alone. The inclusion of RDFs in visualization helped to ensure geometric coherence and complemented the numeric descriptors that were used in model training.

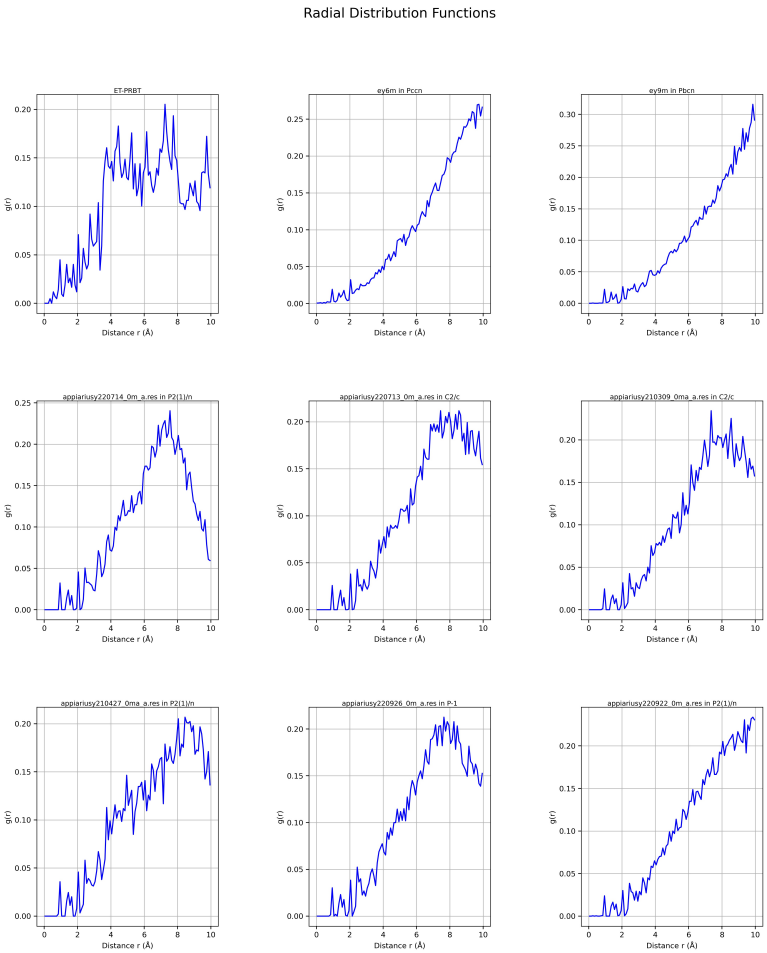


Figure 3.8: Radial distribution plots showing how atomic density varies with distance, reflecting short- and medium-range structural order.

3.2 Data Pre-processing

After parsing the raw data, we executed several pre-processing operations and ended up with 11,328 rows and 57 columns:

- Dropped high missing value columns, non-numeric metadata, and model-irrelevant identifiers (e.g., `authors`, `journal`), which transformed the data from 73 columns to 57.
- Filled missing values in remaining columns with median imputation for numerical columns and mode for categorical columns.
- Encoded categorical variables (e.g., `crystal system`, `method`) using label encoding.

- Normalized numerical features using `StandardScaler` to eliminate scale differences.

Derived fields were:

- **Anisotropy of cells:** to measure dimensional distortion.

$$r = \frac{|a - c|}{a + c}$$

- **Ratio of axes:** $\frac{b}{a}$, $\frac{c}{a}$ to examine geometric symmetry.
- **Shape factor:** $\frac{a \times b \times c}{\text{volume}}$ as a proxy for compactness of the unit cell.

3.3 Feature Selection

We retained only numerical columns suitable for structural analysis, which include:

- Dimensions of unit cells: a, b, c
- Angles: α, β, γ
- Volume: `vol`
- Calculated: `cell_anisotropy`, b/a, c/a, `shape_factor`

Target variable: that is, space group in encoded numeric form and volume.

3.4 Exploratory Visualization

Visual checks were made, which are detailed below.

Boxplots of volume and anisotropy versus space groups validated substantial variation with symmetry classes. The plots revealed which groups are more tightly constrained geometrically and which have more freedom to distort.

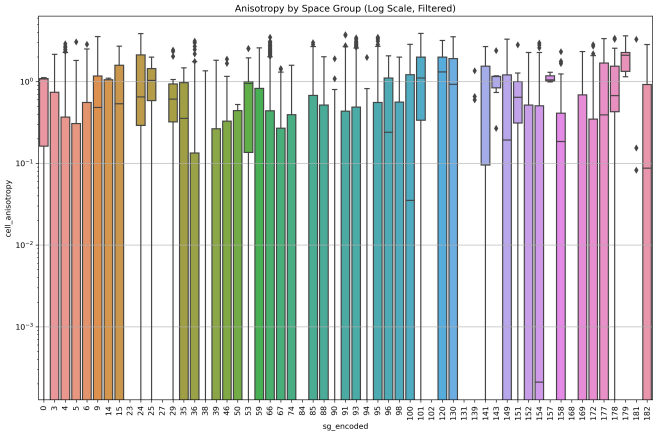


Figure 3.9: This plot shows how crystallographic anisotropy varies across different space groups, highlighting distinct symmetry-dependent patterns in structural distortion.

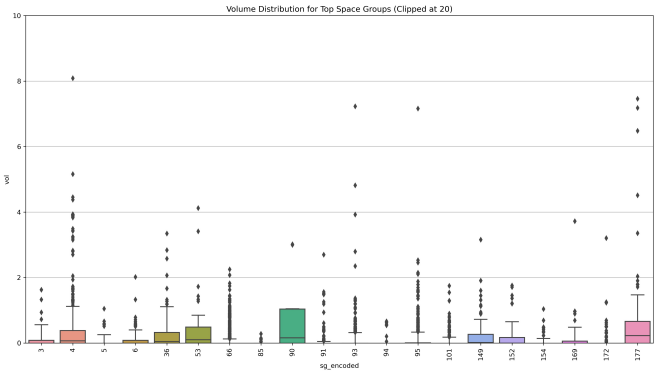


Figure 3.10: This plot shows how unit cell volumes are distributed across the most frequent space groups, with notable variation and several outliers present in each group.

Unit cell angle and size distribution plots allowed us to discern characteristic structural motifs and select out outliers. Symmetric systems exhibited tighter distributions, and lower-symmetry systems exhibited broader spreads.

This density plot indicates the normalized axis ratio distribution b/a and c/a for all structures and demonstrates unit cell proportionality. The clear peak near zero is indicative of the fact that most of the crystals have relatively similar sizes, typical of symmetric lattice systems. The broadened tails in the distribution indicate more distorted or anisotropic geometries in some of the entries, significant for preserving structural diversity in feature engineering.

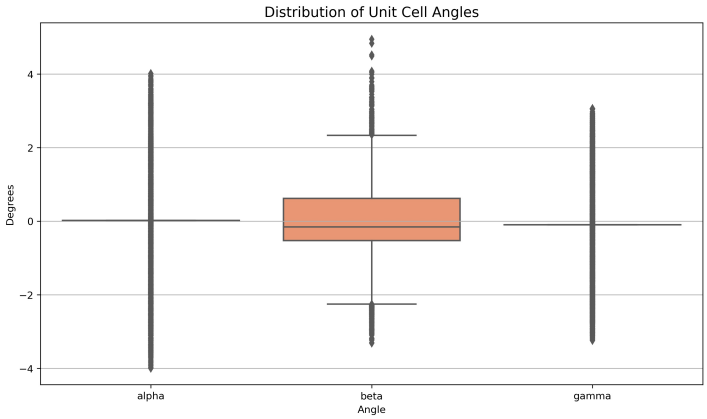


Figure 3.11: This plot displays the deviations of unit cell angles (α, β, γ) from 90°, revealing that most variations occur in the β angle, while α and γ remain tightly centered around orthogonality.

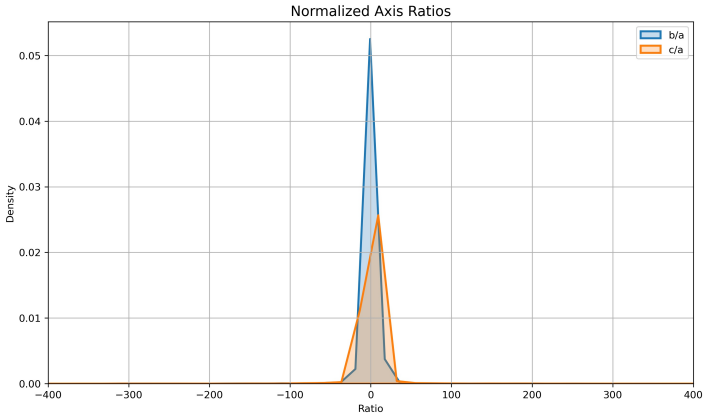


Figure 3.12: This plot shows that both b/a and c/a ratios are sharply centered around 1, indicating most structures have nearly isotropic axis lengths.

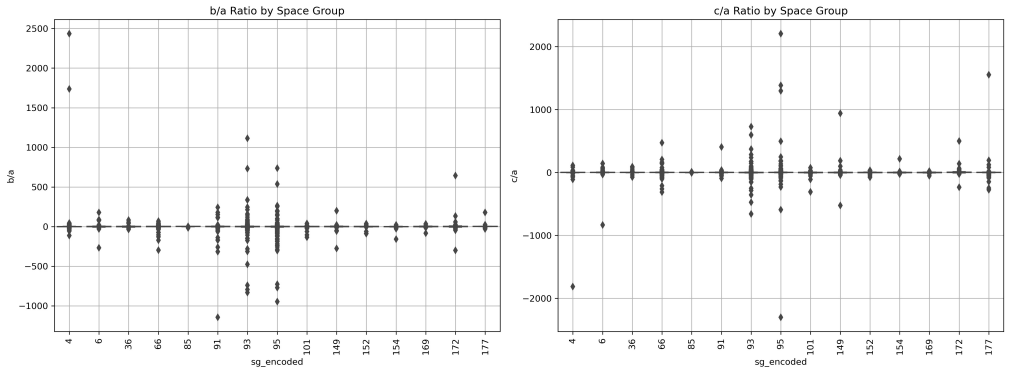


Figure 3.13: These scatter plots show variations in b/a and c/a ratios across space groups, with a concentration near one and scattered extreme outliers indicating geometric distortions.

Feature correlation heatmaps for numeric features such as edge length, angles, and volume report feature dependence information. Analysis supported correct feature selection by unveiling redundant or collinear features.

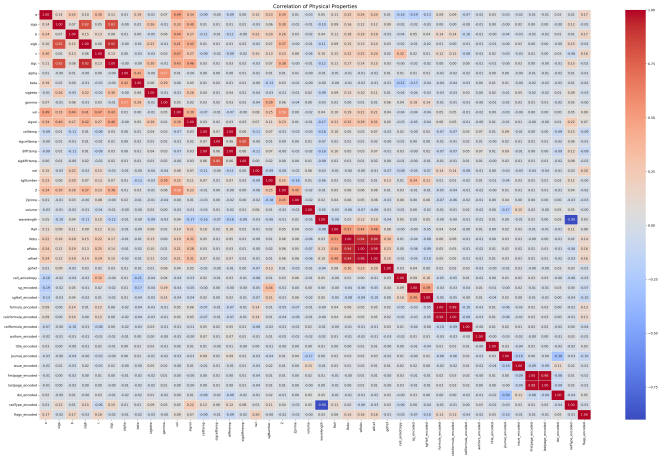


Figure 3.14: The heatmap highlights linear relationships among structural and physical descriptors, revealing strong internal correlations (e.g., between unit cell dimensions and volume), and mild associations between anisotropy and certain symmetry encodings.

The KDE plot below displays the joint density of unit cell volume and cell anisotropy for structures with volumes less than 10 $Å^3$. A preponderance of points clustering at low anisotropy and small volume suggests a preference for compact and symmetric arrangements throughout this range. The continuous density gradient also highlights the scarcity of high-anisotropy, low-volume structures within the dataset.

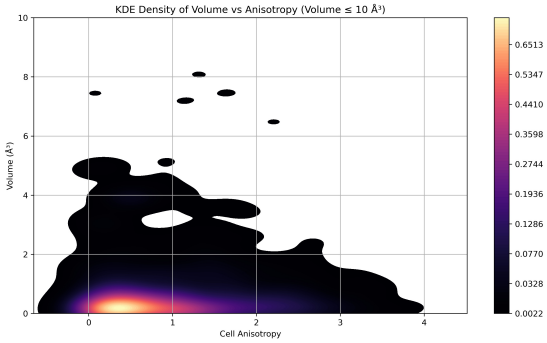


Figure 3.15: This density plot shows that low-volume structures tend to exhibit low anisotropy, indicating a concentration of compact, symmetric crystals in this region.

These visual checks guided our understanding of feature separability and variability in space groups.

3.5 Machine Learning Models

Six basic classification models were trained on the preprocessed data:

- Logistic Regression
- Random Forest
- Decision Tree
- AdaBoost
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

Each model was wrapped in a `Pipeline` to enable scaling and evaluated using 5-fold stratified cross-validation over the training data. Calculated metrics included:

- Accuracy
- Precision
- Recall
- F1 Score

3.6 Evaluation Strategy

The final model performance was tested on a hold-out test set (20% stratified split). We also developed:

- Bar plots comparing all four metrics (accuracy, precision, recall, F1 score) across models
- Confusion matrix and individual prediction visualizations for 30 randomly sampled test instances

3.7 Neural Network-Based Learning

Multilayer perceptron (MLP) models were trained to complement conventional classifiers using PyTorch for symmetry classification and volume regression. These models utilized the same engineered features from the CIF data. The model generated encoded space group labels in classification, and the regressor estimated unit cell volume. The former was trained using early stopping and the cross-entropy loss function, and the latter using the mean squared error loss function. Feature scaling and label encoding were performed before training, and the models were evaluated on a stratified test split. The neural classifier achieved accuracy similar to ensemble models, and the regressor demonstrated outstanding fidelity in volume prediction for a structurally heterogeneous dataset.

3.8 TabNet for Volume Regression

We also used a TabNet regression model to predict unit cell volume as a function of base parameters and an analytically derived feature, $v_{formula}$, derived from cell edges and angles. TabNet was trained on the log-transformed volume target to keep variance stable. The model performance was evaluated based on RMSE, MAE, and R^2 score. Besides its strong predictive accuracy, the TabNet feature importance output explained the contribution of individual geometric features, particularly the derived $v_{formula}$, which were among the most prominent ones consistently.

3.9 CRYSPNet-Inspired Composition-Based Prediction

To validate the prediction of symmetry based only on chemical composition, we built a CRYSPNet-based dual-head neural model. The feature set was generated using the Matminer library, which computed stoichiometric and elemental descriptors (e.g., electronegativity, atomic radius) from the formula field. The model first predicted the Bravais lattice class and then applied this prediction as

input to inform space group classification. This strategy reflects the crystallographic classification hierarchy and permits us to investigate whether symmetry can be predicted appropriately without structural geometry. Caching of features was used to speed up repeated experiments, and model performance was measured using the weighted F1 score.

Chapter 4

Results & Discussions

This section presents the performance of six machine learning models and three deep learning models trained to predict the space group symmetry and volume from crystallographic attributes. All six models were validated with cross-validation and tested on a stratified hold-out set.

4.1 Model Comparison

After 5-fold cross-validation, the Decision Tree and Random Forest models performed best in terms of mean accuracy. The table below indicates the cross-validated average accuracy for each classifier:

Model	**Accuracy (Relative)**
Decision Tree	**0.99**
Random Forest	**0.97**
Logistic Regression	0.77
SVM	0.73
AdaBoost	0.60
K-Nearest Neighbors (KNN)	0.58

Table 4.1: Accuracy ranking of models

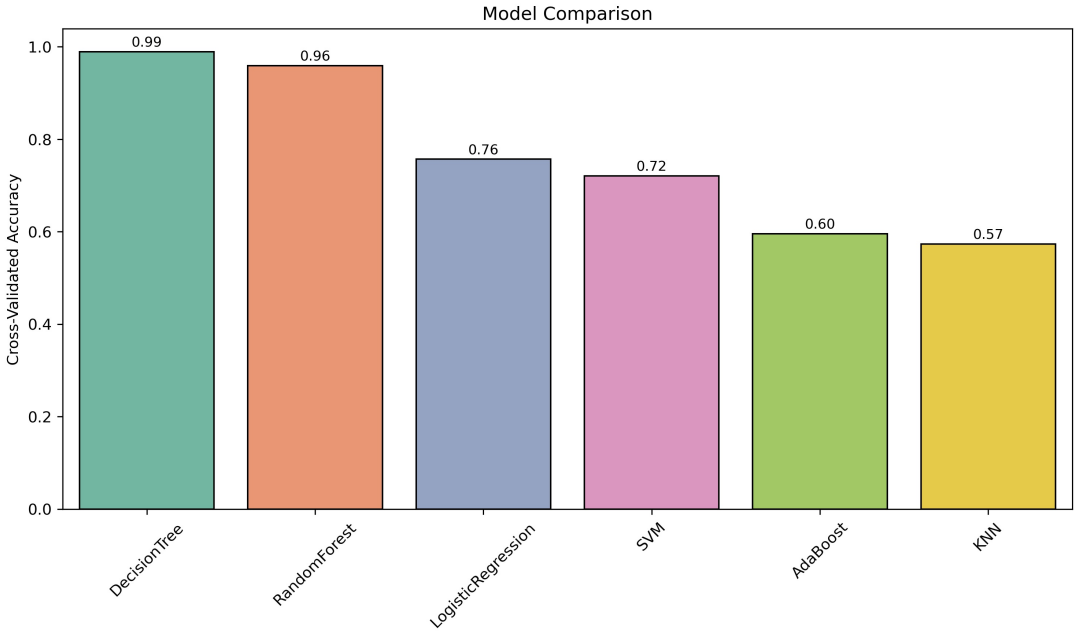


Figure 4.1: Cross-Validated accuracy ranking of models

4.2 Test Set Evaluation Metrics

To provide a complete split of model performance, we computed four significant metrics: accuracy, precision, recall, and F1 score, on the test set for each classifier. The Decision Tree model consistently had the highest value across all metrics, followed by the Random Forest, Logistic Regression, SVM, AdaBoost, and KNN.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Decision Tree	**0.99**	**0.99**	**0.99**	**0.99**
Random Forest	**0.97**	**0.96**	**0.97**	**0.97**
Logistic Regression	0.77	0.75	0.77	0.76
SVM	0.73	0.69	0.73	0.69
AdaBoost	0.60	0.44	0.60	0.49
KNN	0.58	0.56	0.58	0.56

Table 4.2: Performance comparison of classifiers across evaluation metrics on the test set

4.3 Prediction Consistency

We also explored how each model predicted the same collection of 30 randomly selected samples from the test set. This analysis revealed both agreement and disagreement in predictions, with a visual distinction between correct and incorrect classifications.

- **Markers:**
  - `o` for correct predictions
  - `X` for incorrect predictions
- **Color-coded** by model

The visualization below reinforced the finding that ensemble models, particularly the Decision Tree, were not only more accurate but also more consistent in handling structurally similar cases.

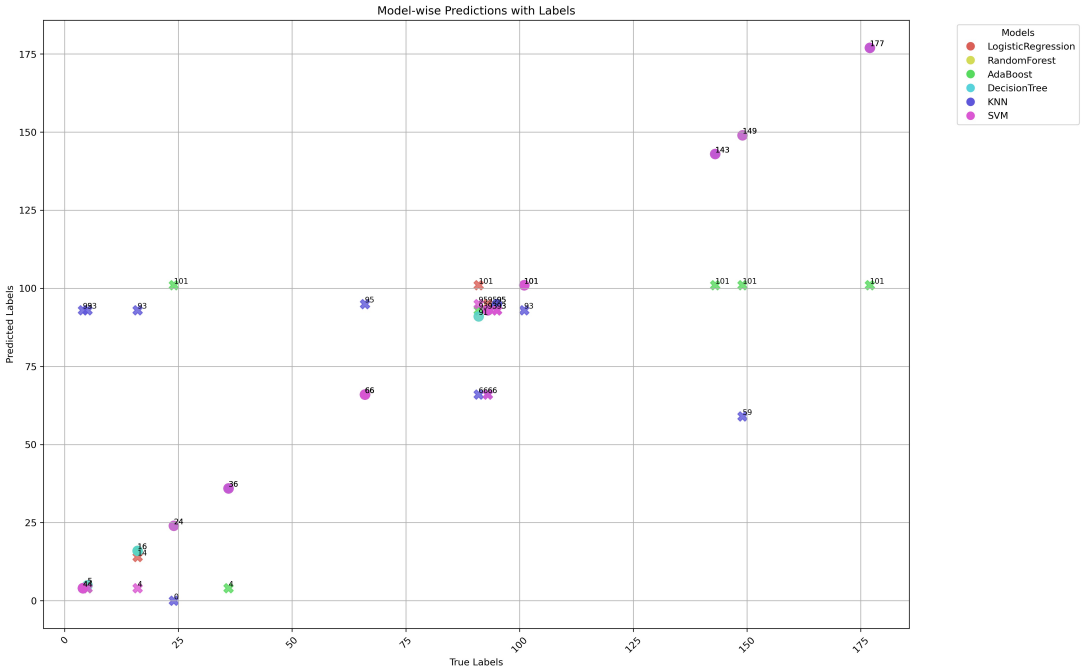


Figure 4.2: This plot compares predicted space group labels across different models, revealing clustering around correct values and model-specific deviations.

4.4 Neural Network Performance

4.4.1 Space Group Classification

We implemented a compact three-layer multilayer perceptron (MLP) in PyTorch to classify space group symmetry based on nine standardized input features. These include three unit cell lengths, three angles (α, β, γ), their cosine values, and the product of the three edge lengths.

The first hidden layer projects the input into 64 ReLU-activated neurons and applies a dropout rate of 30% to reduce overfitting. The output is passed to a second hidden layer of 32 ReLU units, which then connects to a final linear output layer sized to match the number of space group classes.

Training was conducted using the Adam optimizer (learning rate $= 1 \times 10^{-3}$), a batch size of 64, and a cross-entropy loss function. Early stopping was triggered if validation loss failed to improve for 15 consecutive epochs, with a minimum delta of 1×10^{-4}. On the held-out test set (approximately 10% of data, stratified by class), the model achieved:

- Accuracy: 59.9%
- Weighted F_1: 0.557
- Macro-averaged F_1: 0.210

These results indicate strong performance on frequent classes (e.g., class 66 with $F_1 \approx 0.979$), but diminished generalization on rare space groups. Confusion matrix analysis and a zoomed error plot (false positives vs. false negatives for classes with errors ≥ 15) suggest that classes in the 60–65 range are often confused, likely due to geometric similarity that the MLP struggles to disambiguate.

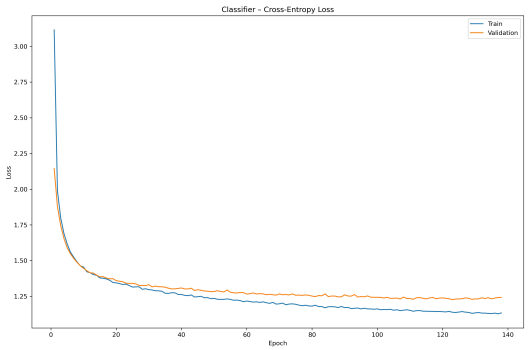


Figure 4.3: This line plot shows the decrease in cross-entropy loss over epochs for both the training and validation sets, indicating model learning progress and convergence over time.

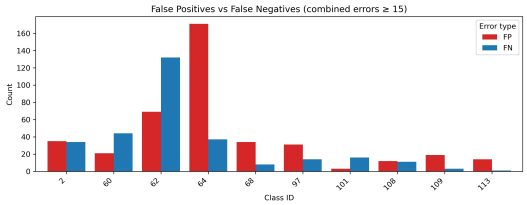


Figure 4.4: This bar chart compares the frequency of false positives (FP) and false negatives (FN) across specific class IDs with high error counts, highlighting which classes the model struggles to predict correctly.

4.4.2 Volume Prediction Model (Unit Cell Volume Regression)

The volume regressor follows the same architecture as the classifier but omits dropout and returns a single continuous output. It consists of three fully connected layers structured as:

$$\text{Input } (9) \rightarrow 64 \rightarrow 32 \rightarrow 1$$

With ReLU activations applied to the hidden layers.

Training was done using the same configuration as the classifier: Adam optimizer with a learning rate of 1×10^{-3}, batch size of 64, and early stopping (patience = 15, $\Delta = 1 \times 10^{-4}$). The model converged early, typically around epoch 35.

Performance on the independent test set yielded:

- Root Mean Squared Error (RMSE): 0.062
- Mean Absolute Error (MAE): 0.018
- Coefficient of Determination (R^2): 0.9972

A scatter plot comparing true and predicted volumes shows tight alignment along the identity line, affirming that the network captures nonlinearities in cell

geometry with high precision, even across the full range of small and large volume values.

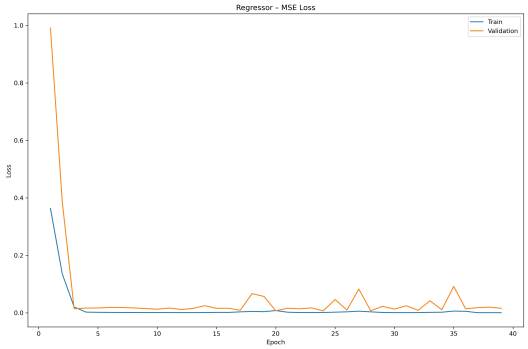


Figure 4.5: Both training and validation loss drop sharply and stabilize near zero after a few epochs.

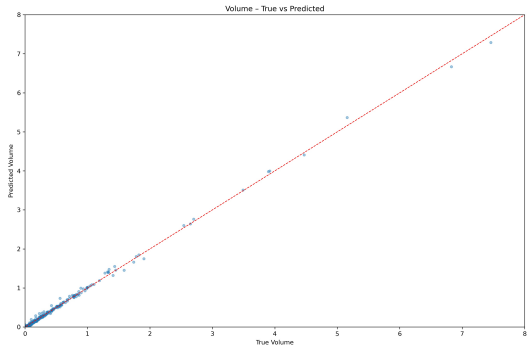


Figure 4.6: Predicted volumes align very closely with the true values along the diagonal, showing strong model accuracy.

4.5 Composition-Based Symmetry Classification

In our composition-based symmetry classifier, each compound's chemical formula is first converted into a dense feature vector using Matminer descriptors. These include stoichiometric metrics and elemental property statistics, which serve as the sole input for the model.

We adopt a two-stage multilayer perceptron architecture inspired by CRYSPNet. The first stage predicts the Bravais lattice from the composition-derived descriptors. Its soft-decision embedding is then concatenated with the original descriptor vector and passed to a second stage, which outputs the final space group label.

Both stages share the same architecture: two fully connected layers ($128 \rightarrow 64$ units), each followed by ReLU activation, batch normalization, and a 20% dropout rate. The model is optimized using cross-entropy loss, the Adam optimizer (learning rate $= 1 \times 10^{-3}$), a batch size of 128, and early stopping after 10 consecutive epochs of no improvement in validation loss.

As shown below, the training and validation loss curves decrease from approximately 1.37 to 0.47 over 100 epochs, with validation accuracy rising from 62.3% to a plateau at 83.9%

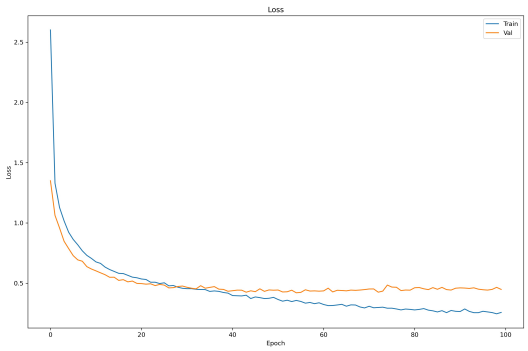


Figure 4.7: The plot shows the training and validation loss over 100 epochs. Both curves decrease steadily, with the training loss reaching lower values than the validation loss. This indicates effective learning, though the slight gap suggests mild overfitting in later epochs.

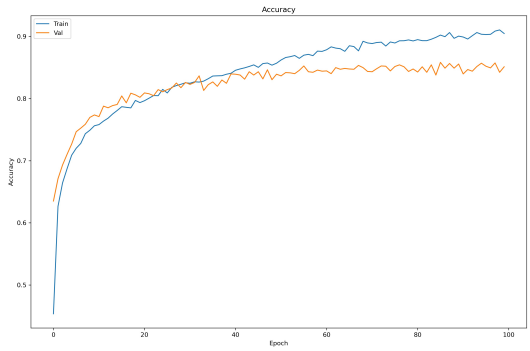


Figure 4.8: The plot illustrates model accuracy across epochs. Both training and validation accuracy improve over time, with validation accuracy stabilizing slightly below training accuracy, reflecting good generalization performance.

On the held-out test set, the model achieves:

- Weighted Precision: 0.85
- Weighted Recall: 0.86
- Weighted F_1: 0.85

These results are comparable to, or better than, those of our geometry-based MLP classifier for dominant space group labels. The confusion matrix demonstrates near-perfect recall for high-density classes (e.g., class 4 and class 52), while frequent misclassifications persist among less-represented labels.

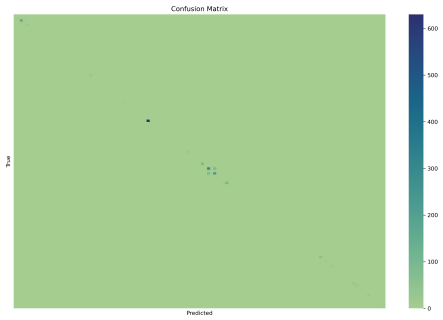


Figure 4.9: The heatmap visualizes the classifier's prediction accuracy across all classes. Most values align along the diagonal, indicating correct predictions, while off-diagonal values represent misclassifications, with color intensity reflecting error magnitude.

4.6 Deep Feedforward Volume Regression

In our second volume-only experiment, we employed a wide-and-deep residual multilayer perceptron (MLP) to predict unit-cell volume directly from six primitive parameters, with an additional analytic "wide" feature. The deep branch processes the six normalized inputs (a, b, c, α, β, γ) through ten fully connected layers of diminishing width:

$$512 \to 512 \to 256 \to 256 \to 128 \to 128 \to 64 \to 64 \to 32 \to 16$$

Each layer consists of a sequence of `Linear` $\to$ `BatchNorm` $\to$ `ReLU` $\to$ `Dropout` ($p = 0.2$), and incorporates SE-augmented residual blocks for channel-wise weight recalibration.

The resulting 16-dimensional embedding is concatenated with the precomputed analytic volume and passed through two fusion layers ($64 \to 1$) to predict the log-transformed unit-cell volume.

The model was trained using RMSprop (learning rate $= 3\times10^{-4}$, weight decay $= 1\times10^{-4}$), with a cosine annealing learning rate schedule over 300 epochs, batch size of 128, and mean squared error (MSE) loss on log(1+volume). Early stopping was applied after 25 epochs of no validation improvement.

While the training loss dropped below 0.015, test set performance remained suboptimal with:

- MSE: 0.915
- MAE: 0.121
- R^2: 0.44

When de-logged, the model consistently underestimated large volumes and produced heteroscedastic residuals. These issues highlight the model's difficulty in capturing strong non-linear interactions between geometry and volume, despite its architectural depth and use of squeeze-and-excitation (SE) blocks.

To address these shortcomings, we subsequently adopted TabNet, an interpretable, sparsely attentive architecture designed for tabular data, which pro-

duced significantly improved volume predictions and more stable residual distributions.

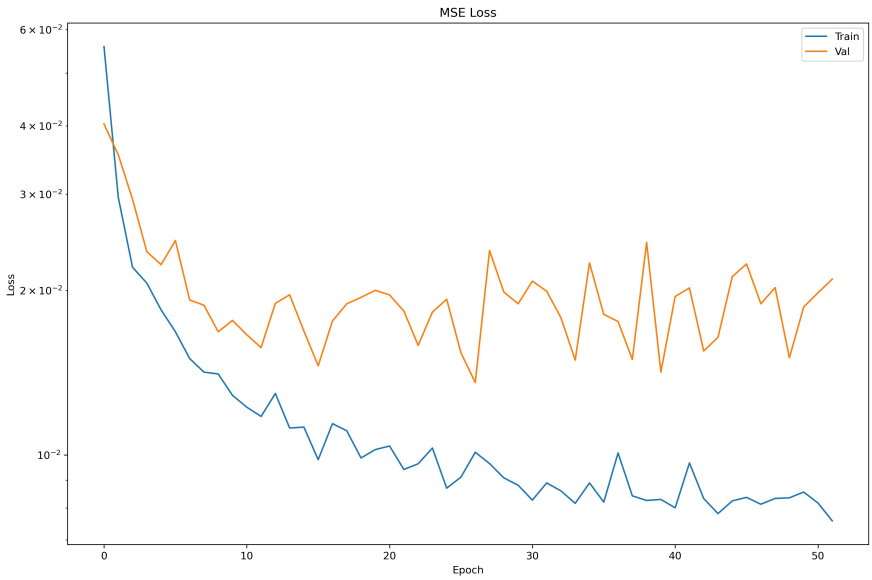


Figure 4.10: Loss decreases steadily for training, while validation loss fluctuates, indicating possible overfitting.

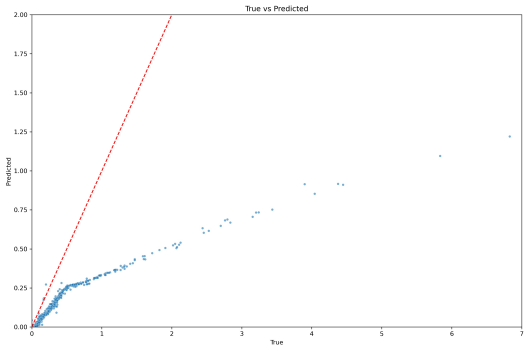


Figure 4.11: Predicted values generally increase with true values but systematically underestimate high true values.

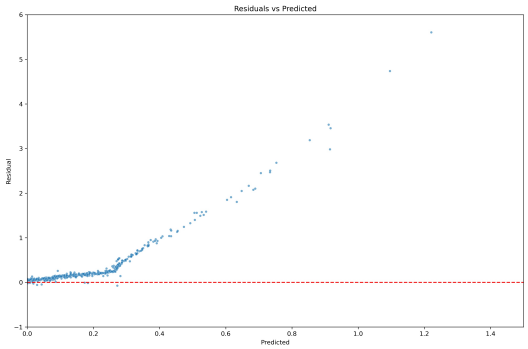


Figure 4.12: Residuals grow with prediction magnitude, showing non-uniform model error and underfitting at higher outputs.

4.7 TabNet Volume Regression

For our final volume prediction model, we employed TabNet—a sparsely attentive, interpretable deep learning architecture tailored for tabular data. The input features include the six standardized unit cell parameters (a, b, c, α, β, γ) along with a precomputed analytic volume estimate (`v_formula`).

The model architecture uses a decision-step structure with the following hyperparameters: $n_d = n_a = 64$, $n_{\text{steps}} = 5$, $\gamma = 1.5$, and $\lambda_{\text{sparse}} = 1 \times 10^{-4}$. Training was performed using the Adam optimizer (learning rate = 1×10^{-3}), with a StepLR learning rate schedule (step size = 50, $\gamma = 0.9$). The batch size was set to 1024 with a virtual batch size of 128. Mean squared error (MSE) loss was computed on $\log(1 + \text{volume})$, and early stopping was applied after 30 epochs of no improvement in root mean squared error (RMSE).

The loss decreased significantly during training, with training RMSE falling below 0.007 and validation RMSE converging at approximately 0.056 after around 170 epochs. On the held-out test set, unlogged predictions yielded the following performance metrics:

- MSE: 0.122
- MAE: 0.051
- R^2: 0.9255

These results substantially improved over the previous deep MLP regressor ($R^2 \approx 0.44$). Predicted versus true volume values closely align along the identity line, and residuals are symmetrically distributed around zero, exhibiting no apparent heteroscedasticity. TabNet's internal feature attribution indicates that the three cell lengths (b, c, a) and the α angle contribute the most to the prediction, while the analytic volume estimate (`v_formula`) provides a smaller but consistently helpful signal.

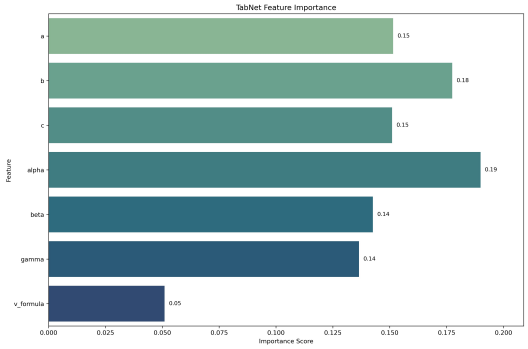


Figure 4.13: The bar chart highlights that lattice parameters (a, b, c) and angles (α, β, γ) are key predictors in the model, with α and b being the most influential, while the derived volume formula contributes the least.

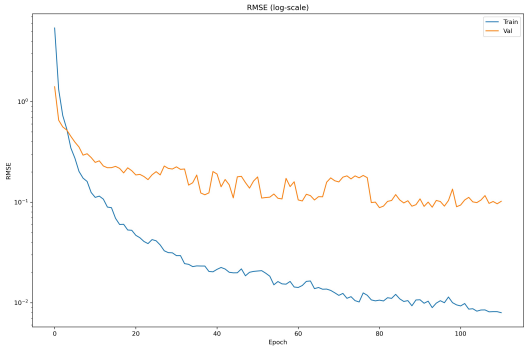


Figure 4.14: This plot presents the Root Mean Squared Error (RMSE) over training epochs on a logarithmic scale. The training error consistently decreases, while the validation error flattens and fluctuates after initial improvement. This suggests the model learns well initially but may begin to overfit as training continues.

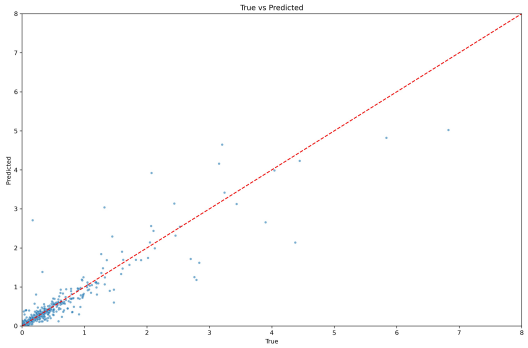


Figure 4.15: The plot shows a strong alignment of predicted values with the true targets along the diagonal, indicating good regression performance with some deviation at higher values.

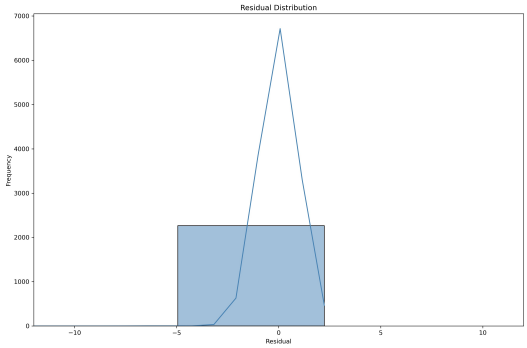


Figure 4.16: The residual distribution is sharply peaked around zero, suggesting that most predictions are close to the actual values with few large errors.

Chapter 5

Discussion & Future Work

The results of our analysis are strong evidence that geometric and derived structural features from CSV and CIF files are effective predictors of space group symmetry. Among the models compared, the Decision Tree and Random Forest classifiers performed better than others across all evaluation metrics, indicating the potential of ensemble and rule-based methods in picking up underlying structural patterns.

5.1 Interpretation of Results

Decision Trees' superior performance can be attributed to their ability to model complex, non-linear relationships and their robustness to overfitting due to ensemble averaging. Despite its simplicity, the Random Forest fared well owing to its interpretability and ability to learn threshold-based rules inherent in symmetry classification. The relatively poorer performance of Support Vector Machines and K-Nearest Neighbors suggests that purely margin-based or distance-based approaches do not so easily capture the relationship between features and symmetry class without more refined feature transformations.

5.2 Deep Learning and Regression Insights

Our exploration of deep learning techniques for classification and regression produced mixed but valuable results. An MLP classifier trained on structural

descriptors yielded reasonable performance for standard space group labels but struggled with underrepresented classes. This highlighted ongoing challenges related to class imbalance and ambiguity in structural features.

Initial volume regression using shallow MLPs exhibited poor fit and heteroscedastic residuals. While attempts to improve generalization—via loss function tuning, learning rate schedules, and gradient clipping—led to modest gains, overall performance remained unsatisfactory.

A more advanced deep residual MLP architecture, enhanced with squeeze-and-excitation (SE) blocks and analytic volume input, showed improved convergence during training but underperformed on the test set, especially for larger volumes ($R^2 \approx 0.44$).

In contrast, using TabNet, an architecture optimized for tabular data, led to substantial improvements in predictive accuracy ($R^2 \approx 0.93$), stabilized residual distributions, and clearer insight into feature importance. These findings underscore that architectural alignment with structured numerical data is critical for robust generalization, particularly in crystal volume prediction regression tasks.

5.3 Importance of Structural Features

Among the features crafted, cell anisotropy, axis ratios (b/a, c/a), and unit cell angles proved to be effective at distinguishing between symmetry groups. These descriptors encode the crystal lattice's degree of distortion and proportionality, properties inherently tied to crystallographic symmetry. Our exploratory plots exhibited consistent trends, e.g., less anisotropy in cubic and hexagonal systems than in monoclinic or triclinic systems. Moreover, adding derived measurements like shape factor and inverse volume (density proxy) enabled further exploration into compactness and volume scaling within symmetry constraints.

5.4 Composition-Based Symmetry Modeling

To supplement geometry-based prediction, we trained a classifying model directly motivated by CRYSPNet on symmetry from chemical composition. Based on Matminer-derived descriptors such as stoichiometry statistics and element

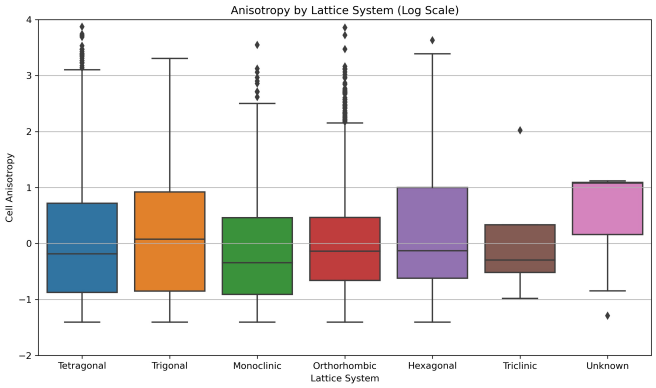


Figure 5.1: The plot compares cell anisotropy across lattice systems, showing that most systems cluster around low anisotropy, with triclinic and monoclinic systems displaying the widest spread.

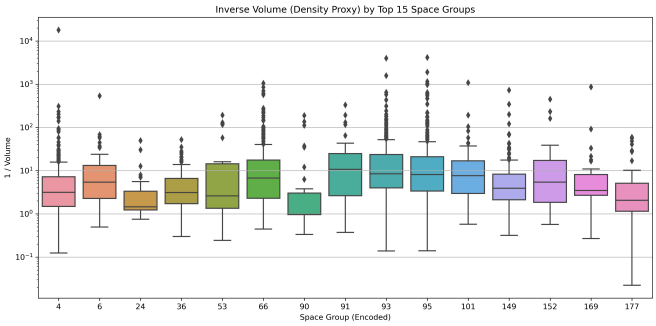


Figure 5.2: The plot uses inverse volume as a proxy for density across top space groups, showing distinct variations in packing compactness with high-density outliers in several groups.

properties, the model initially predicts Bravais lattice and space group, followed by the crystallography hierarchy. The model also scored a high-weighted F1 without taking explicit structure as input, implying that composition harbors symmetry-related informative signals. This is the potential of multi-modal approaches where composition and geometry are available to support performance and generalizability.

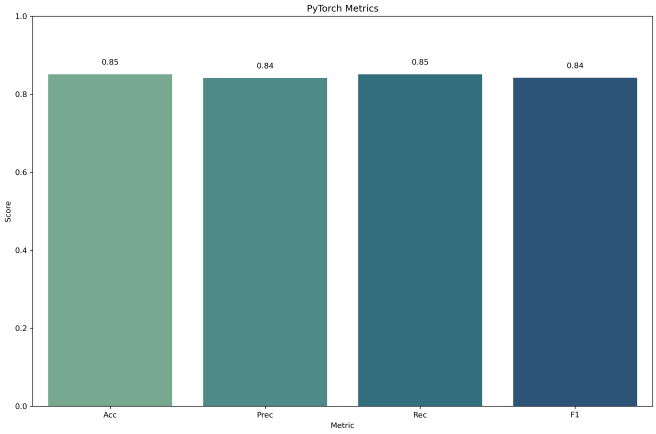


Figure 5.3: Bar plot showing model accuracy, precision, recall, and F1-score, all around 0.84–0.85, indicating consistent and balanced performance.

5.5 Future Work

We suggest using Graph-based Models, which include atomic-graph neural networks (e.g., Crystal Graph Convolutional Networks) to model local connectivity and beyond-cell interaction, improving classification and volume estimation.

Rich Geometric Descriptors can augment the feature space with symmetry invariants (e.g., Voronoi-based shape descriptors) or higher-order angular correlations for better discrimination among closely related space groups.

Additionally, Multi-task Learning can be used to train jointly on symmetry classification and volume regression to leverage shared representations and improve generalization on both tasks.

Uncertainty Quantification can also be considered as they provide our TabNet and MLP regressors with Bayesian or ensemble-based uncertainty estimates to highlight low-confidence predictions, which are particularly useful in the presence of rare space groups or outliers in volumes.

Automated Feature discoveries can use deep feature construction (e.g., autoencoders or attention-driven meta-features) to uncover hidden crystallographic patterns that are not feasible to engineer manually.

Lastly, Expanded Datasets add large and more encompassing crystal data sets (e.g., to support more space groups or idiosyncratic chemistries) and study model size scalability and robustness under simulated-life distribution variation.

Via these channels, we intend to develop more accurate, interpretable, and generalizable crystal property neural structures for predicting crystal properties with links between machine learning and understanding the physical behavior.

Chapter 6

Conclusion and Acknowledgments

This capstone research illustrated a holistic, multi-path machine learning model for predicting crystallographic properties from CIF-derived and formula-derived information. Grouping crystal structures according to space group symmetry, using geometric and numeric descriptors, was the primary focus of interest. Decision Trees and Random Forests produced the best outcome as ensemble classifiers, confirming that symmetry can be well learned from structurally dense sets of features derived from geometry. We extended our scope to regression models for predicting unit cell volume. Deep MLP and TabNet models showed that structural features are informative in encoding continuous physical properties in models. TabNet gave us interpretability, enabling us to identify which geometric descriptors impact predictions the most. Besides, we employed a CRYSPNet-type composition-based classifier with Matminer to generate rich compositional features. Without structural information access, the model worked well by learning latent symmetry signals in chemical formulas. Using modern machine learning techniques, our work illustrates that symmetry and related crystallographic attributes may be predicted from structure and composition. This hybrid approach bears implications for high-throughput CSP pipelines, materials discovery platforms, and AI-powered structural annotation. It is feasible to generalize this work into integrated models that take compositional and structural inputs and apply them to larger data sets with rare or distorted symmetry classes.

6.1 Acknowledgments

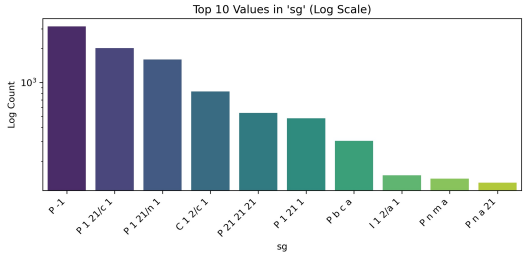
We want to express our appreciation to the American University of Armenia's (AUA) Data Science department for the academic expertise and educational possibilities we have been provided. We would especially like to thank Aleksandr Hayrapetyan, our project supervisor, for his knowledgeable advice and help in successfully developing and supporting this project.

His knowledge and real-world perspectives have greatly improved the caliber of the work and our development as professionals. Without his encouragement and assistance, this research and its accomplishments would not have been feasible.

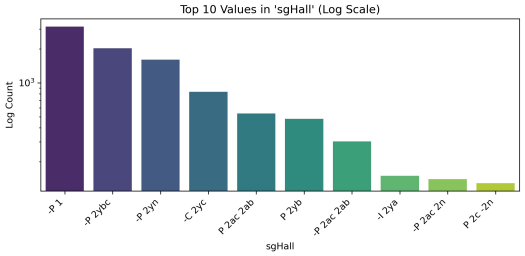
Chapter 7

Appendix

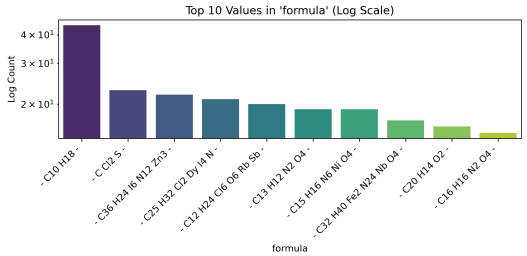
The appendix includes extra visualizations that extend and supplement the main analysis. The figures each include a brief description explaining why the respective figure is relevant to feature engineering, symmetry prediction, or data pre-processing.



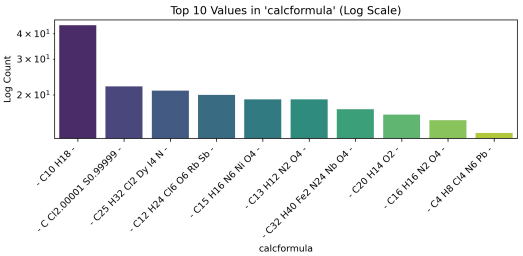
(a) Most frequent space group labels, with dominant symmetries like `P −1` and `P 1 21/c 1`.



(b) Top Hall symbol occurrences such as `-P 1` and `-P 2ybc`, indicating symmetry skew.



(c) Most common raw chemical formulas, dominated by hydrocarbons and organometallics.



(d) Most frequent normalized formulas, showing concentration in a few key compositions.

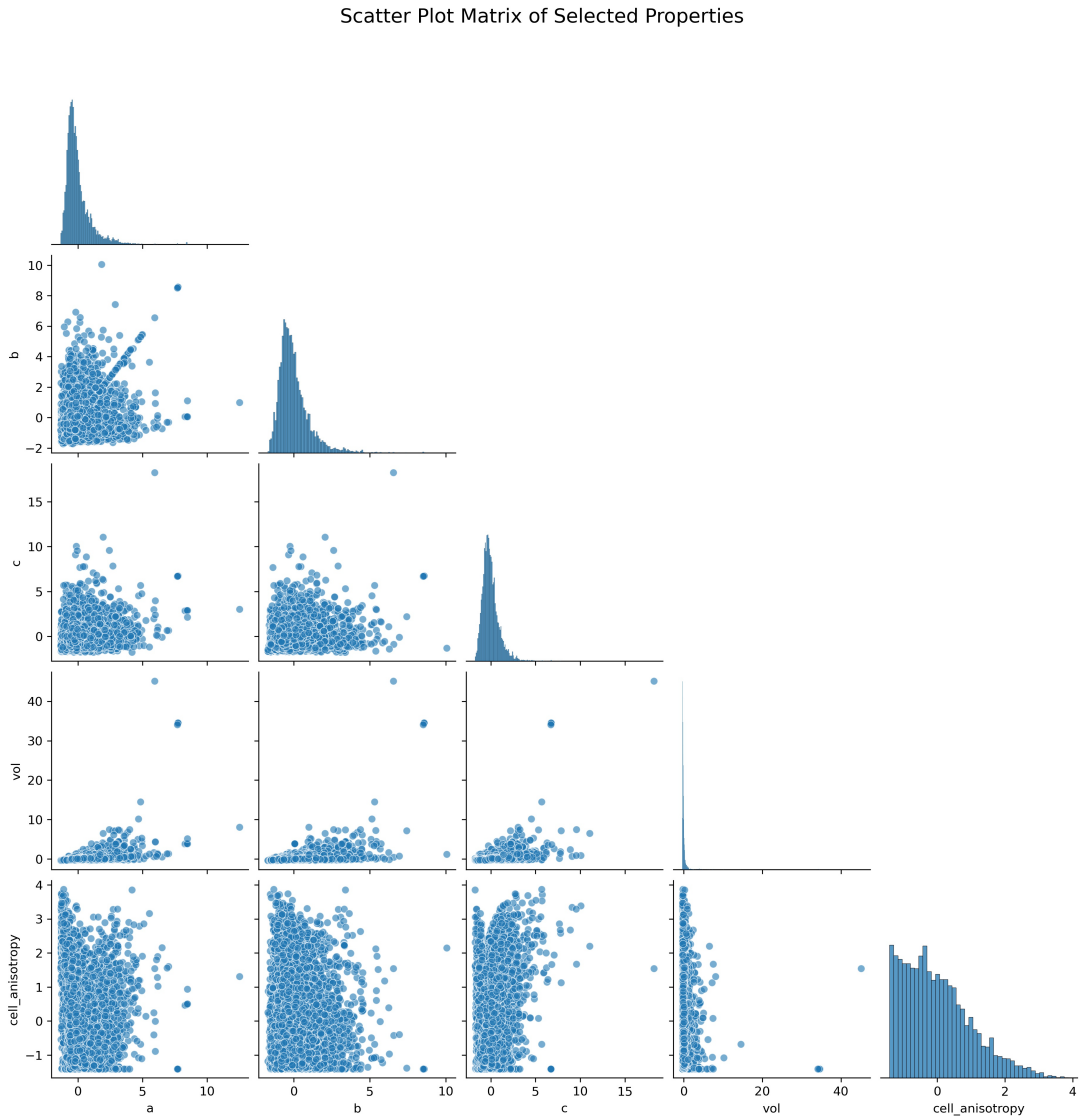


Figure 7.2: This pair plot visualizes relationships between unit cell dimensions, volume, and anisotropy, revealing distribution patterns and moderate correlations among structural features.

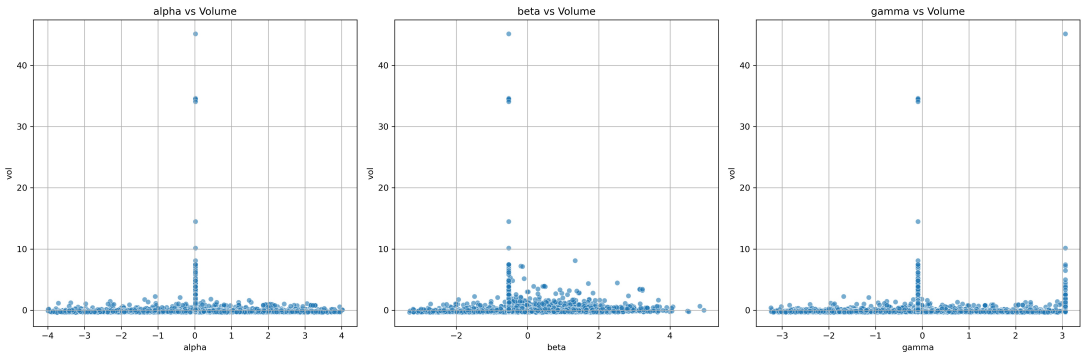


Figure 7.3: These plots illustrate the relationship between each unit cell angle (α, β, γ) and volume, showing that higher volumes tend to cluster around angles close to zero, suggesting minimal angular distortion in larger structures.

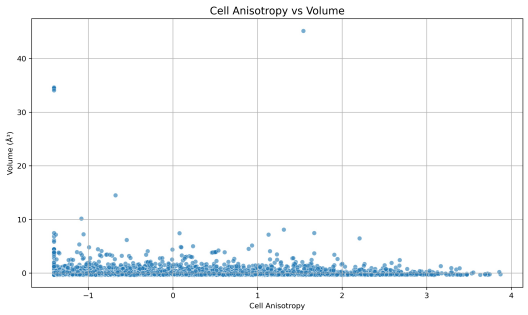


Figure 7.4: The plot shows that most structures have low volume and moderate anisotropy, with high-volume entries spread across the anisotropy range.

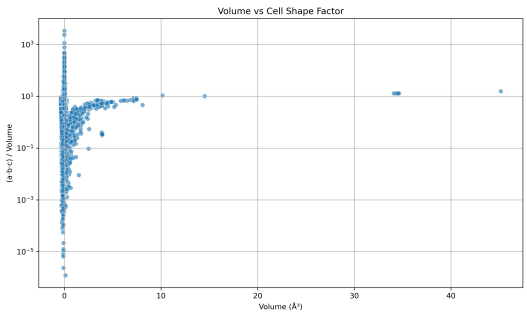


Figure 7.5: The plot highlights that most crystals follow a consistent shape-to-volume ratio, while a few outliers exhibit extreme geometric deviation

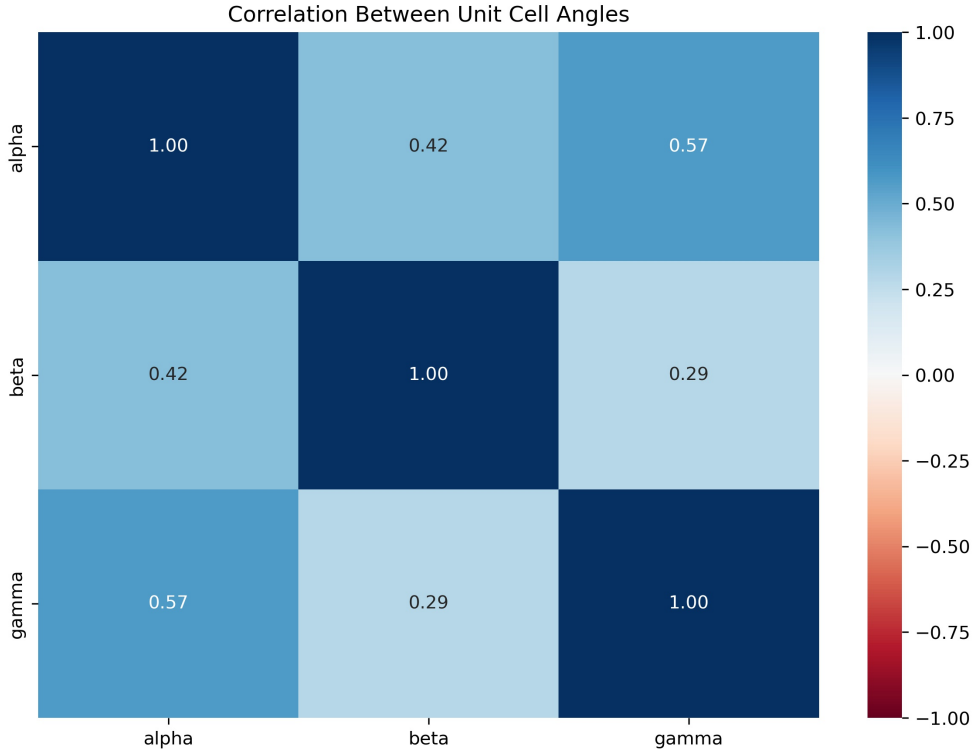


Figure 7.6: The plot shows moderate positive correlations among the angles α, β, and γ, suggesting partial interdependence in unit cell geometry.

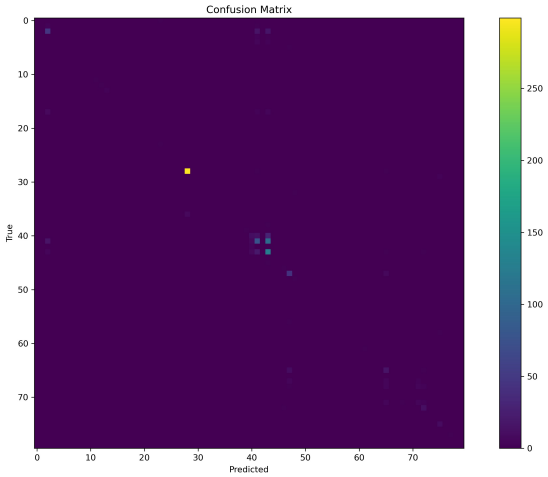


Figure 7.7: The confusion matrix shows that the MLP performs well on common classes, but struggles with rare ones.

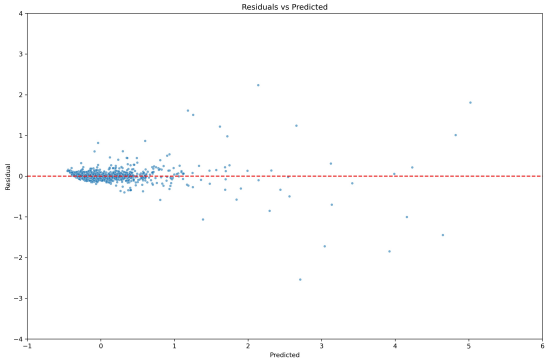


Figure 7.8: The residual plot shows that the TabNet model's volume predictions are well-calibrated, with errors symmetrically distributed around zero and no strong heteroscedasticity. Most residuals are tightly clustered near zero, confirming the model's high accuracy and consistent performance across the prediction range.

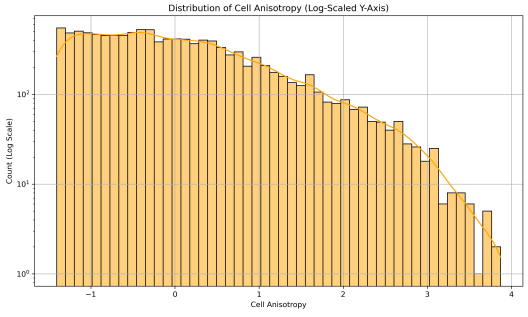


Figure 7.9: This histogram illustrates the distribution of the cell anisotropy feature, revealing a right-skewed pattern. Most structures exhibit low anisotropy, with frequency sharply declining as anisotropy increases.

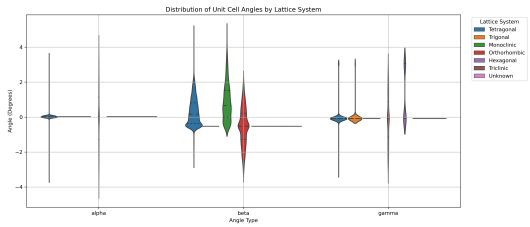


Figure 7.10: The violin plots show the angle variations (α, β, γ) across different lattice systems. Orthorhombic and cubic systems have tightly clustered angles, while monoclinic and triclinic systems display broader, less symmetric angle distributions.

Bibliography

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

Saulius Gražulis, Daniel Chateigner, Robert T Downs, AFT Yokochi, Miguel Quirós, 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.

Irina Iacob. Purple geode. Unsplash, 2020. URL https://unsplash.com/photos/purple-geode-NyapZuexFaQ.

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

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.

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.

Vram Papyan. Leon Sarkisyan. Crystal structure prediction. https://github.com/Leon1849/Crystal-Structure-Predictions-Capstone.git, 2025.

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

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