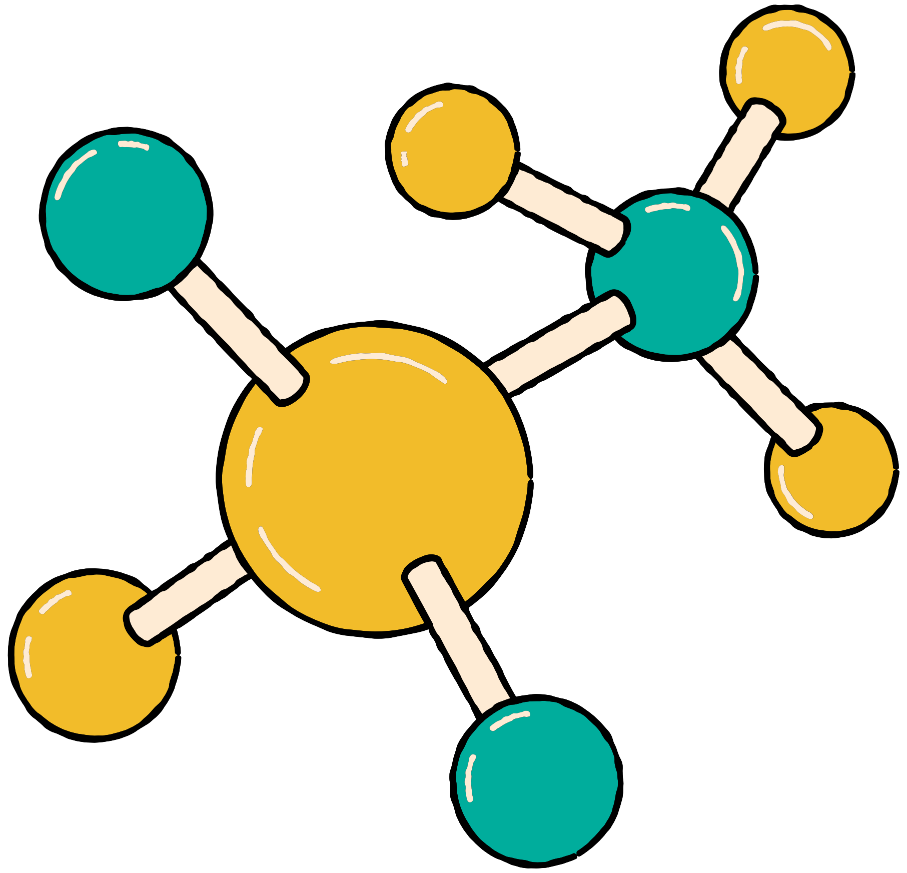
STANFORD RNA 3D FOLDING



Group 20

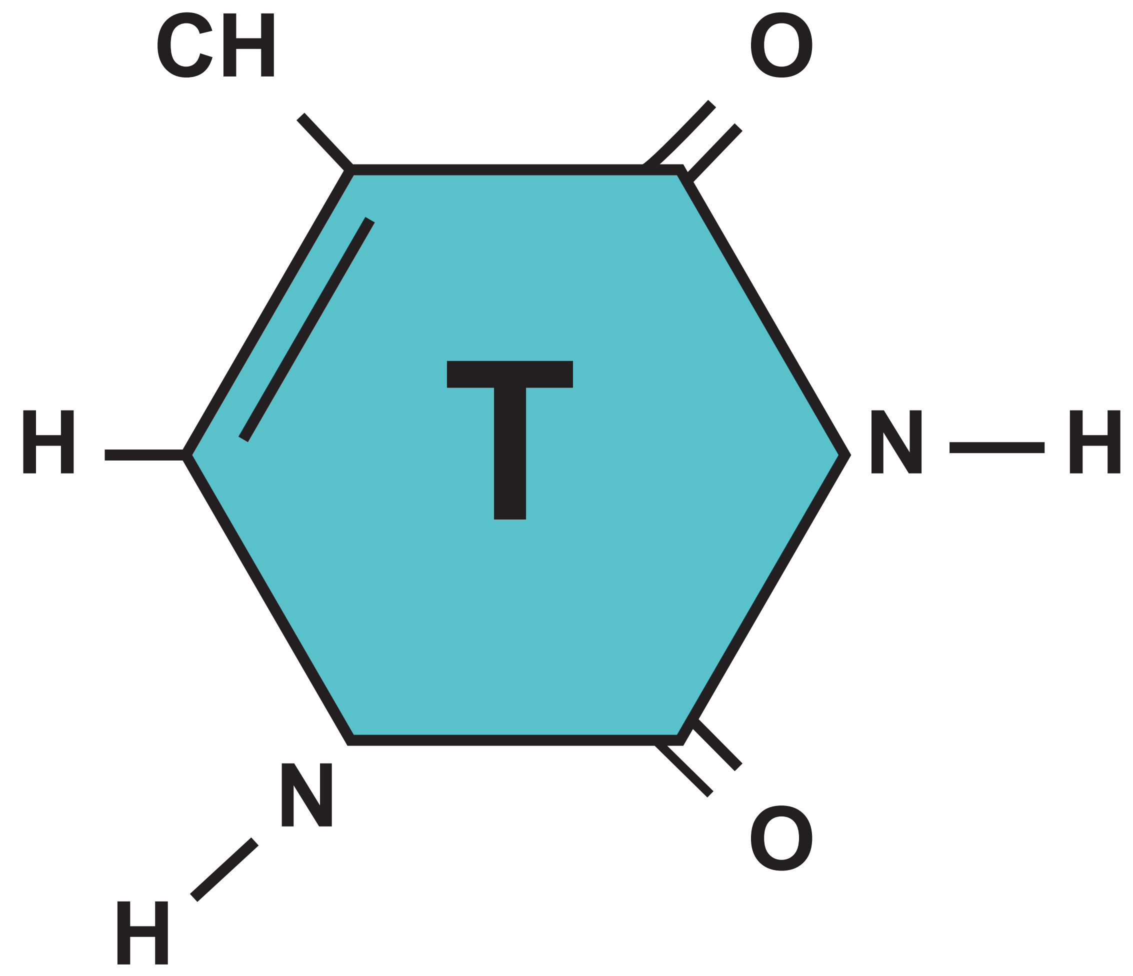
WANG Zhixuan
CHAN Yuk Yee

TABLE OF CONTENTS

1. **Introduction**
2. **Data Description**
3. **Basic Machine Learning Algorithm Method**
4. **Deep Learning Method**
5. **Conclusion**

INTRODUCTION

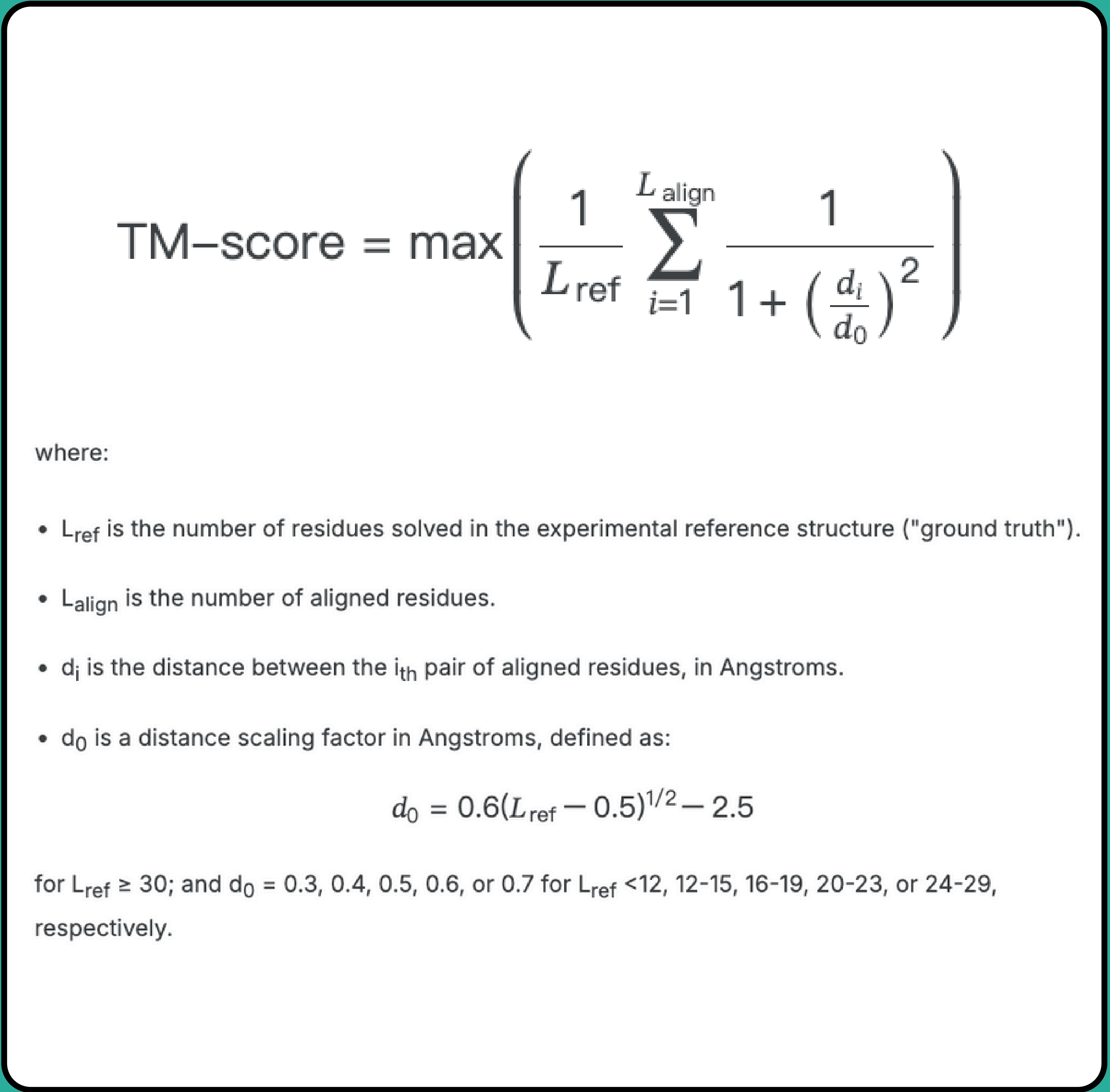
- **Problem statement:** Predicting RNA 3D structure from sequence data
- **Significance:** RNA's crucial role in biological processes
- **Challenge:** Computational prediction as an efficient alternative to experimental methods
- **Our approach:** Traditional ML pipeline and deep learning approach



THE STANFORD RNA 3D FOLDING CHALLENGE

- **Input**: RNA nucleotide sequences (A, C, G, U)
- **Output**: 3D coordinates for each nucleotide
- **Evaluation metric**: TM-score (Template Modeling score)
- **Dataset composition**:

1. Training: 844 sequences
2. Validation: 12 sequences
3. Test: 12 sequences



$$\text{TM-score} = \max\left(\frac{1}{L_{\text{ref}}}\sum_{i=1}^{L_{\text{align}}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right)$$

where:

- L_{ref} is the number of residues solved in the experimental reference structure ("ground truth").
- L_{align} is the number of aligned residues.
- d_i is the distance between the i_{th} pair of aligned residues, in Angstroms.
- d_0 is a distance scaling factor in Angstroms, defined as:

$$d_0 = 0.6(L_{\text{ref}} - 0.5)^{1/2} - 2.5$$

for $L_{ref} \geq 30$; and d_0 = 0.3, 0.4, 0.5, 0.6, or 0.7 for L_{ref} <12, 12-15, 16-19, 20-23, or 24-29, respectively.

DATA EXPLORATION

- Nucleic acids are macromolecules that exist as polymers called polynucleotides. As indicated by the name, each polynucleotide consists of monomers called nucleotides. A nucleotide, in general, is composed of three parts:
- A five-carbon sugar (a pentose)
- A nitrogen-containing (nitrogenous) base
- One to three phosphate groups

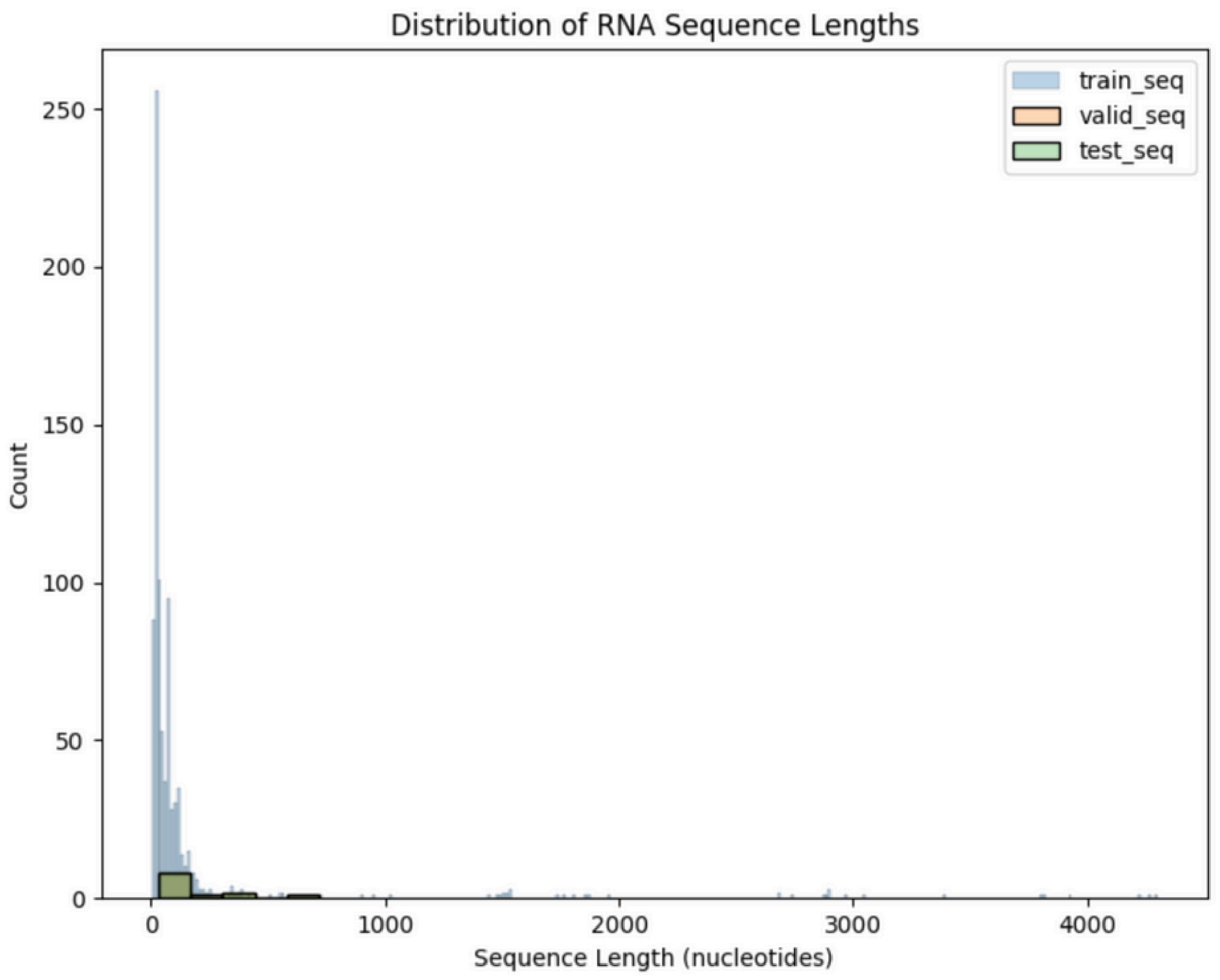
Consider the nitrogenous bases. Each nitrogenous base has one or two rings that include nitrogen atoms. There are two families of nitrogenous bases:

Pyrimidines: A pyrimidine has one six-membered ring of carbon and nitrogen atoms.
Cytosine C
Thymine T
Uracil U

Purines: Purine is larger, with a six-membered ring fused to a five-membered ring.
Adenine A
Guanine G

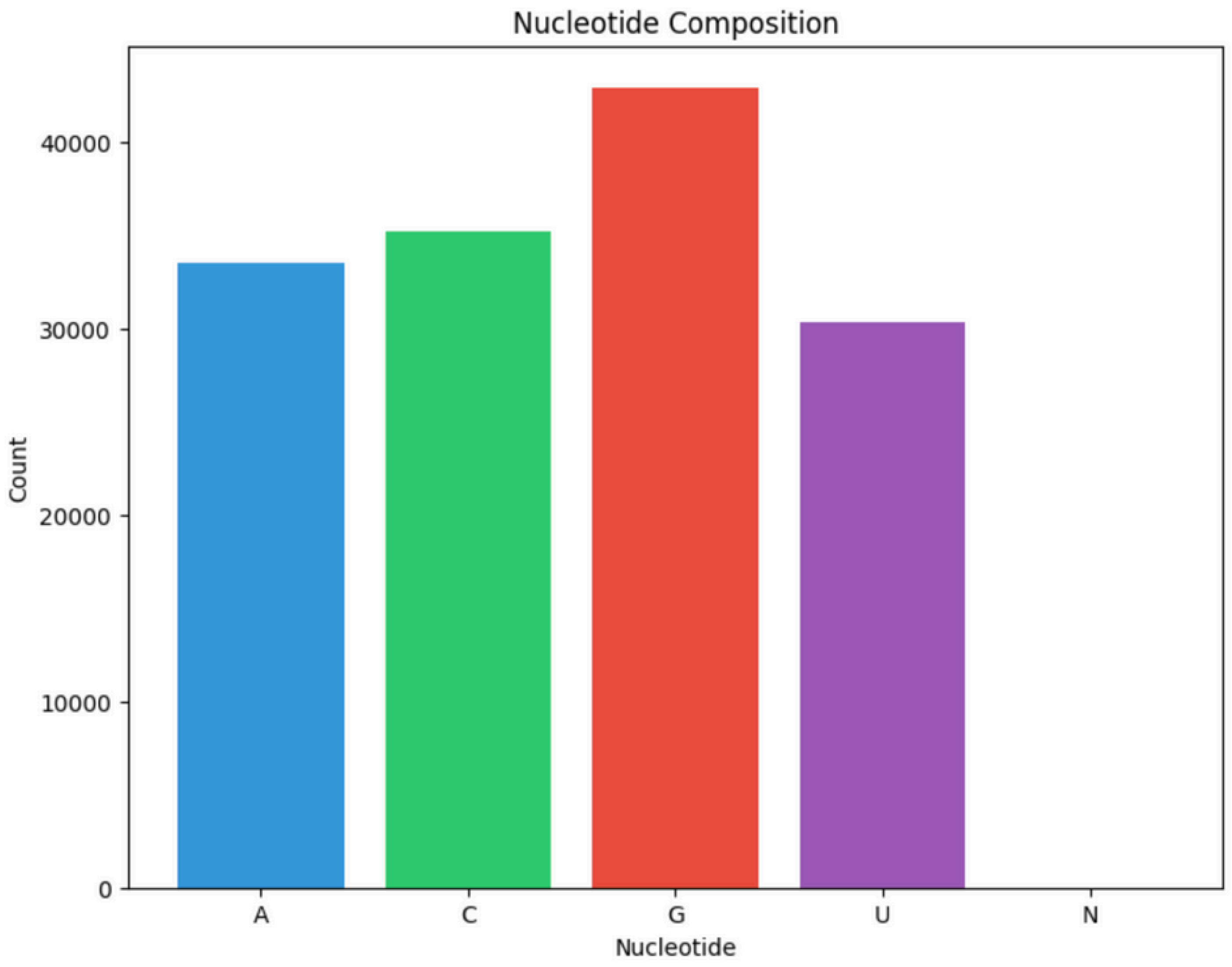
DATA EXPLORATION

- The histogram shows a highly skewed distribution of sequence lengths (3 to 4298 nucleotides)
- Most sequences in the dataset are relatively short (under 200 nucleotides), with a long tail of longer sequences



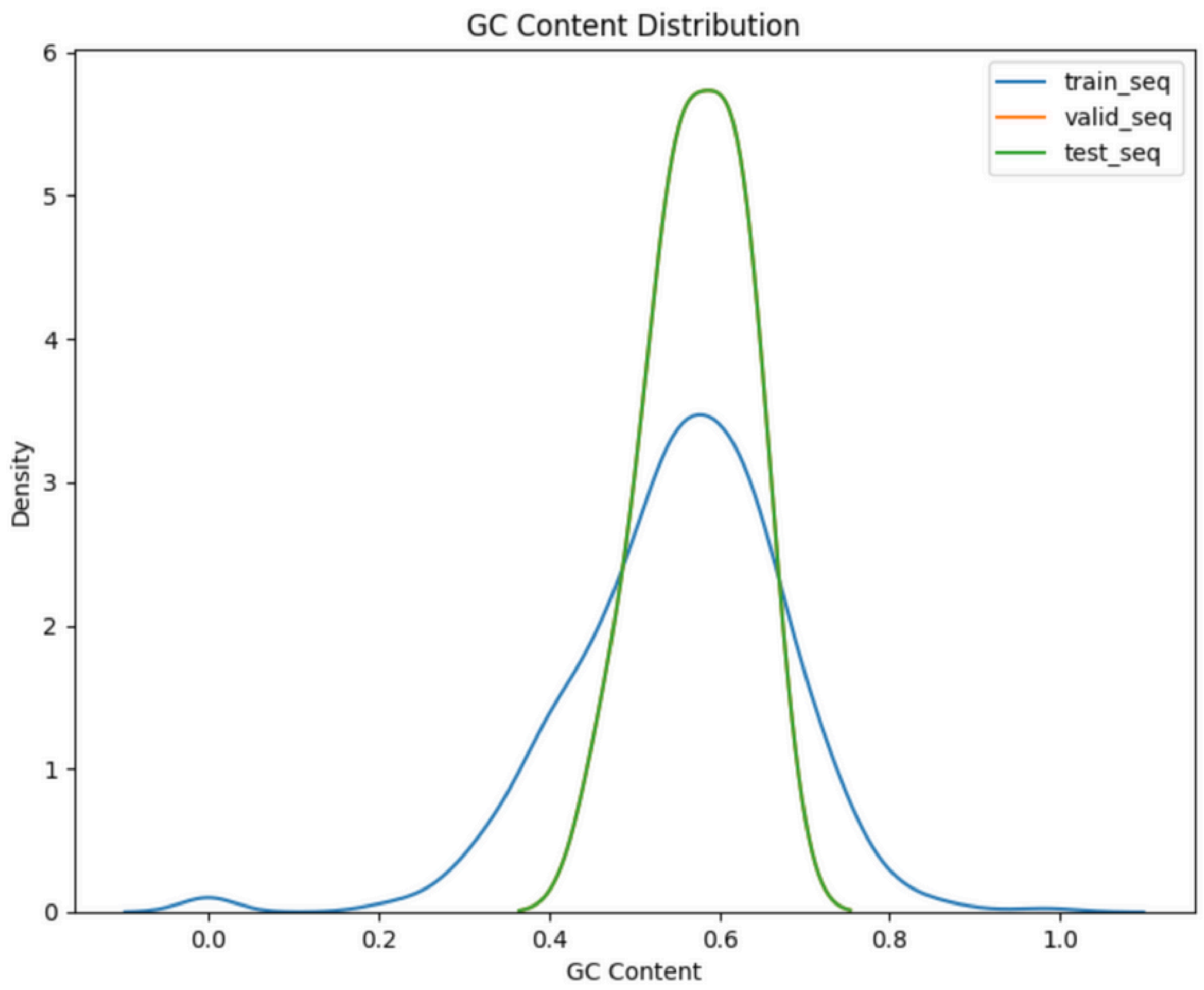
DATA EXPLORATION

- The nucleotide composition chart reveals a slight predominance of G and C nucleotides



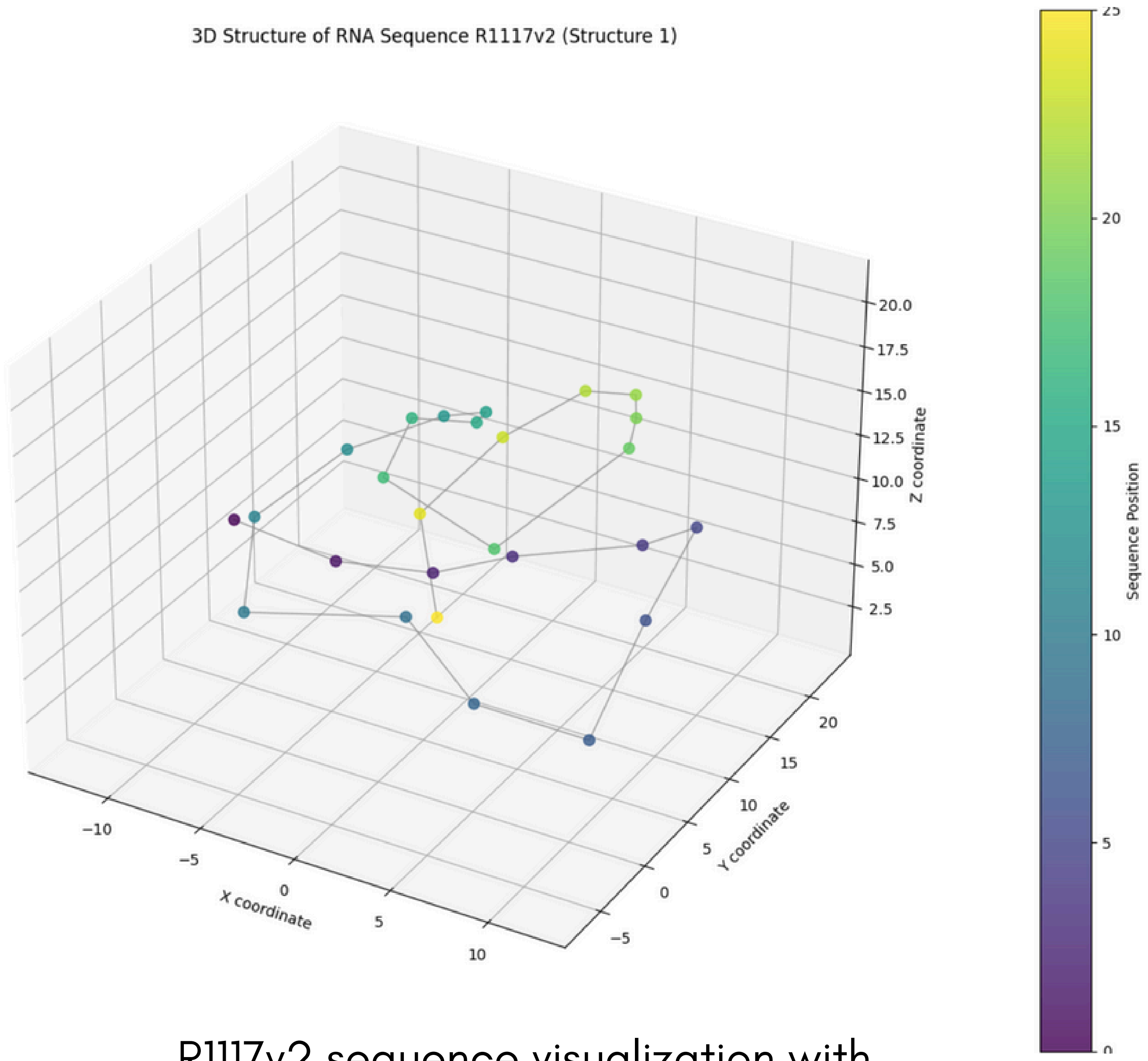
DATA EXPLORATION

- GC content is centered around 0.5-0.6, which aligns with typical RNA sequences
- Note how the training, validation, and test distributions appear similar, suggesting the test set is representative



3D STRUCTURE VISUALIZATION

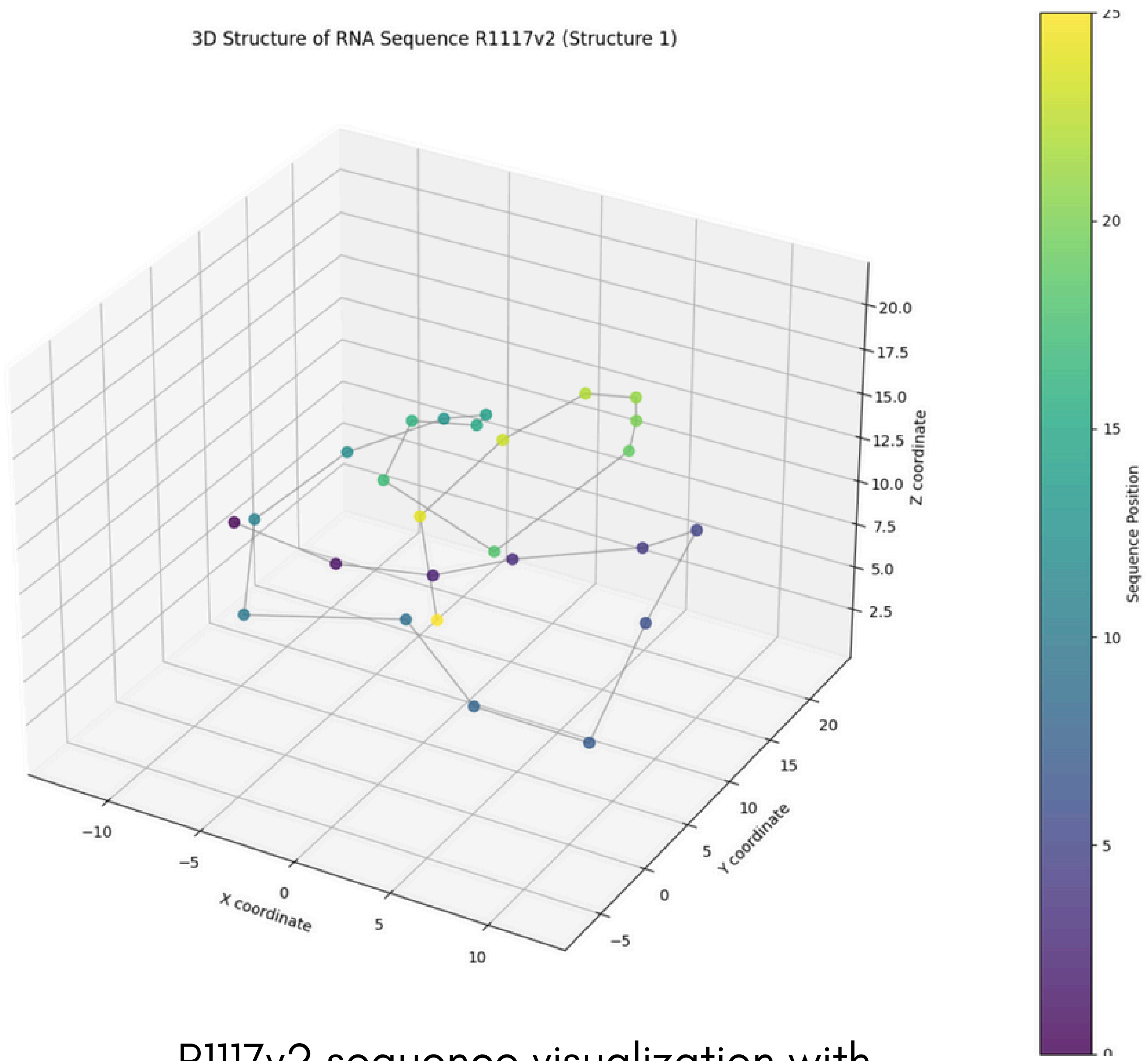
- The visualization shows the 3D coordinates with backbone represented as gray lines and nucleotides as colored points
- The colorbar represents sequence position, showing the progression through the RNA chain
- The specific coordinate ranges: X: 24.66, Y: 29.83, Z: 19.81
- The structure contains 26 valid coordinates out of 30 positions



R1117v2 sequence visualization with sequence length of 30 nucleotides

3D STRUCTURE VISUALIZATION

- Point out structural features like the compact folding pattern and potential stem-loop structures
- This visualization demonstrates the spatial complexity we aim to predict - each nucleotide must be correctly positioned in 3D space
- Predicting these coordinates accurately requires capturing both local sequence patterns and global folding principles



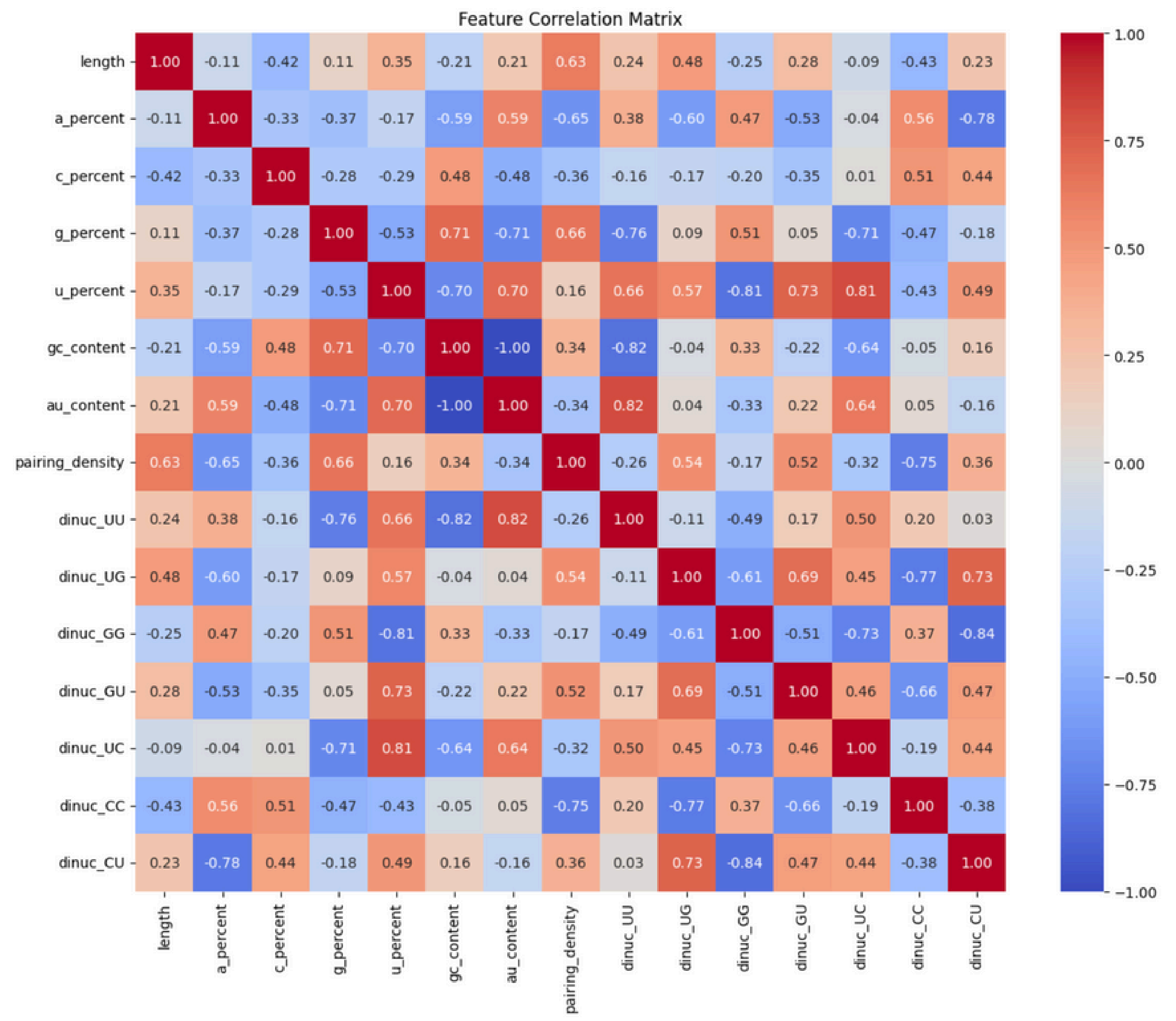
R1117v2 sequence visualization with sequence length of 30 nucleotides

FEATURE ENGINEERING

Key features extracted from RNA sequences:

- Sequence composition (nucleotide frequencies)
- Sequence patterns (dinucleotide frequencies)
- Structural indicators (potential base pairing)
- Position-based features

These correlations helped guide our feature selection, letting us identify redundant features while retaining complementary ones

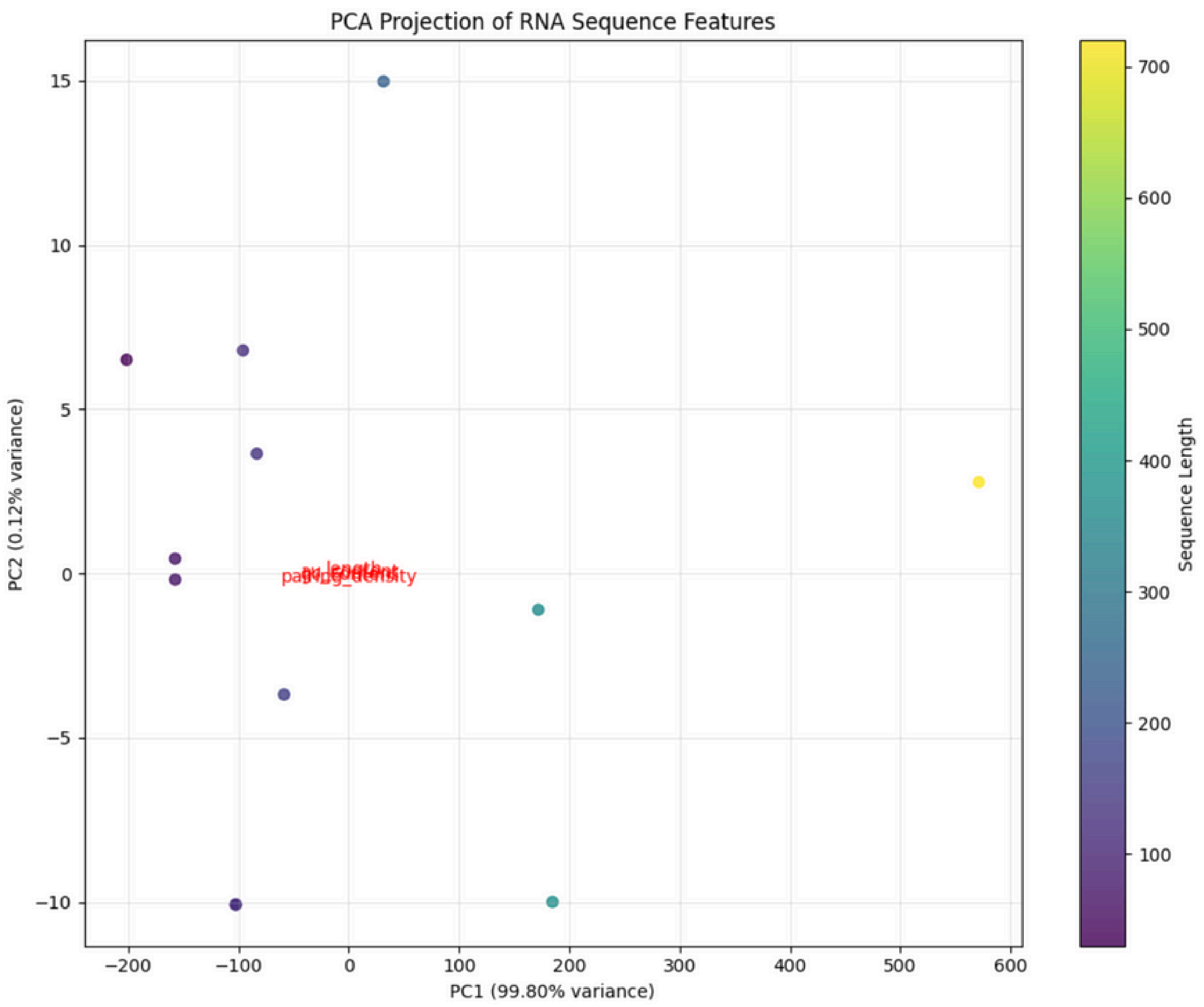


Feature Correlation Matrix

	length	a_percent	c_percent	g_percent	u_percent	gc_content	au_content	pairing_density	dinuc_UU	dinuc_UG	dinuc_GG	dinuc_GU	dinuc_UC	dinuc_CC	dinuc_CU
length	1.00	-0.11	-0.42	0.11	0.35	-0.21	0.21	0.63	0.24	0.48	-0.25	0.28	-0.09	-0.43	0.23
a_percent	-0.11	1.00	-0.33	-0.37	-0.17	-0.59	0.59	-0.65	0.38	-0.60	0.47	-0.53	-0.04	0.56	-0.78
c_percent	-0.42	-0.33	1.00	-0.28	-0.29	0.48	-0.48	-0.36	-0.16	-0.17	-0.20	-0.35	0.01	0.51	0.44
g_percent	0.11	-0.37	-0.28	1.00	-0.53	0.71	-0.71	0.66	-0.76	0.09	0.51	0.05	-0.71	-0.47	-0.18
u_percent	0.35	-0.17	-0.29	-0.53	1.00	-0.70	0.70	0.16	0.66	0.57	-0.81	0.73	0.81	-0.43	0.49
gc_content	-0.21	-0.59	0.48	0.71	-0.70	1.00	-1.00	0.34	-0.82	-0.04	0.33	-0.22	-0.64	-0.05	0.16
au_content	0.21	0.59	-0.48	-0.71	0.70	-1.00	1.00	-0.34	0.82	0.04	-0.33	0.22	0.64	0.05	-0.16
pairing_density	0.63	-0.65	-0.36	0.66	0.16	0.34	-0.34	1.00	-0.26	0.54	-0.17	0.52	-0.32	-0.75	0.36
dinuc_UU	0.24	0.38	-0.16	-0.76	0.66	-0.82	0.82	-0.26	1.00	-0.11	-0.49	0.17	0.50	0.20	0.03
dinuc_UG	0.48	-0.60	-0.17	0.09	0.57	-0.04	0.04	0.54	-0.11	1.00	-0.61	0.69	0.45	-0.77	0.73
dinuc_GG	-0.25	0.47	-0.20	0.51	-0.81	0.33	-0.33	-0.17	-0.49	-0.61	1.00	-0.51	-0.73	0.37	-0.84
dinuc_GU	0.28	-0.53	-0.35	0.05	0.73	-0.22	0.22	0.52	0.17	0.69	-0.51	1.00	0.46	-0.66	0.47
dinuc_UC	-0.09	-0.04	0.01	-0.71	0.81	-0.64	0.64	-0.32	0.50	0.45	-0.73	0.46	1.00	-0.19	0.44
dinuc_CC	-0.43	0.56	0.51	-0.47	-0.43	-0.05	0.05	-0.75	0.20	-0.77	0.37	-0.66	-0.19	1.00	-0.38
dinuc_CU	0.23	-0.78	0.44	-0.18	0.49	0.16	-0.16	0.36	0.03	0.73	-0.84	0.47	0.44	-0.38	1.00

FEATURE ENGINEERING

The PCA visualization demonstrates how our feature set effectively separates RNA sequences, with PC1 capturing primarily length-related variance and PC2 capturing composition differences



BASIC MACHINE LEARNING ALGORITHM METHOD

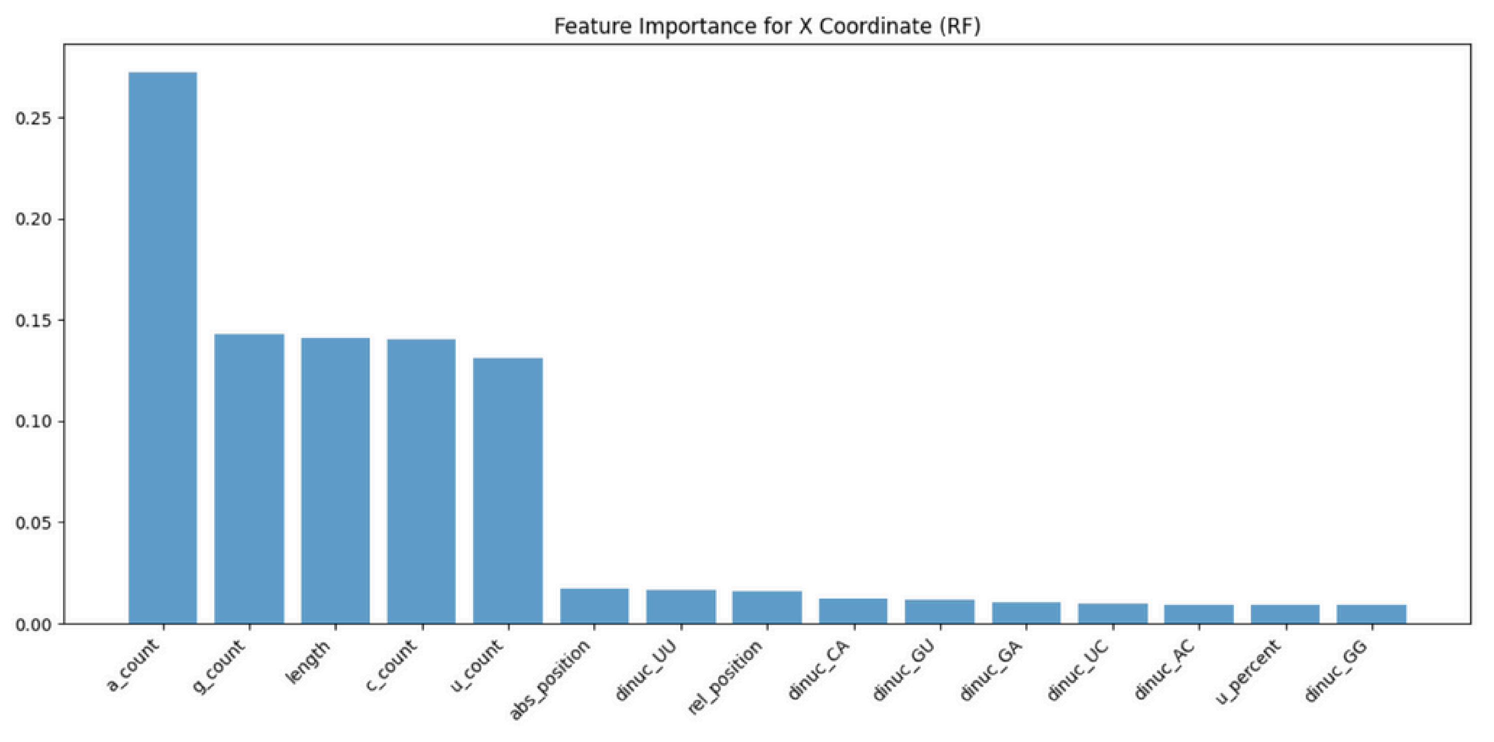
Implemented and compared multiple algorithms for coordinate prediction:

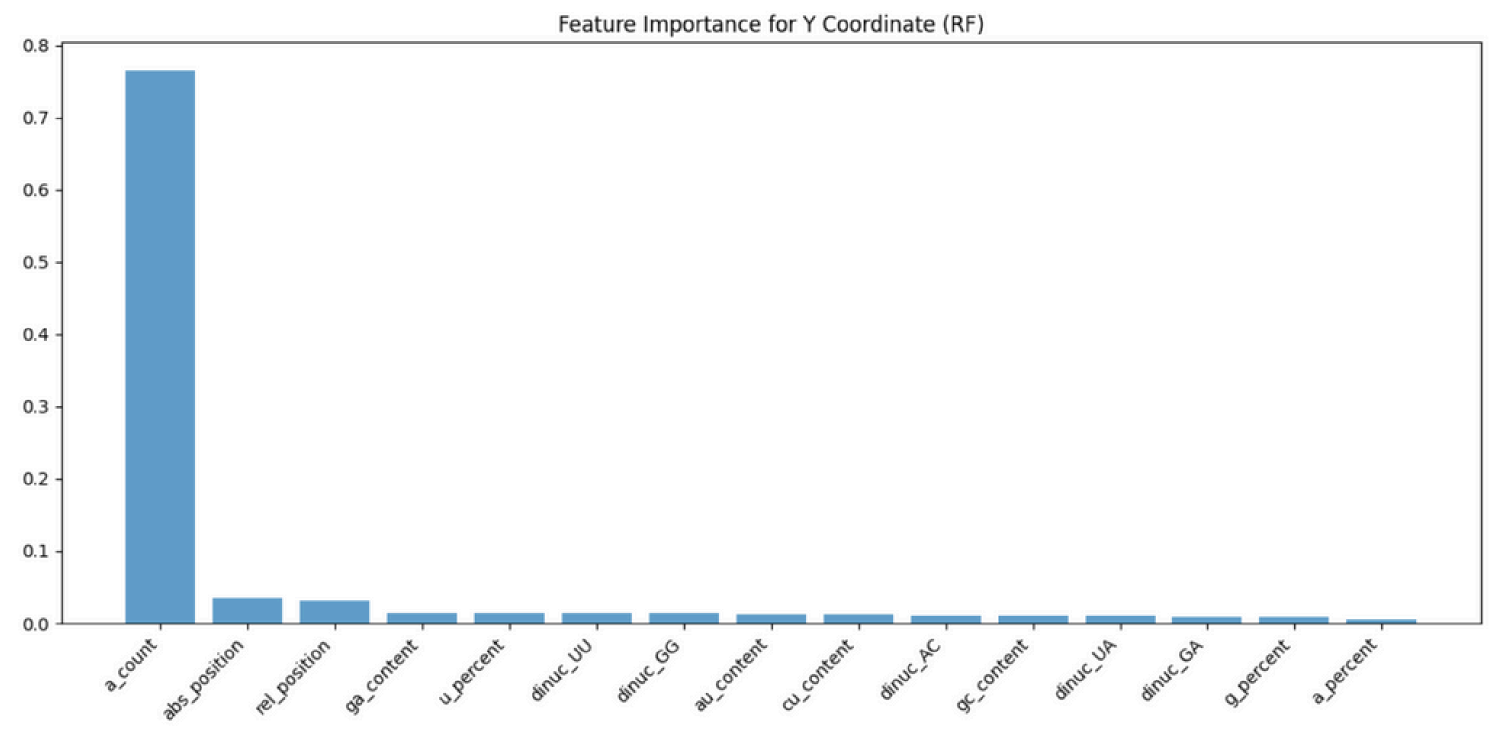
- Random Forest
- Gradient Boosting
- Ridge Regression
- Support Vector Regression
- K-Nearest Neighbors

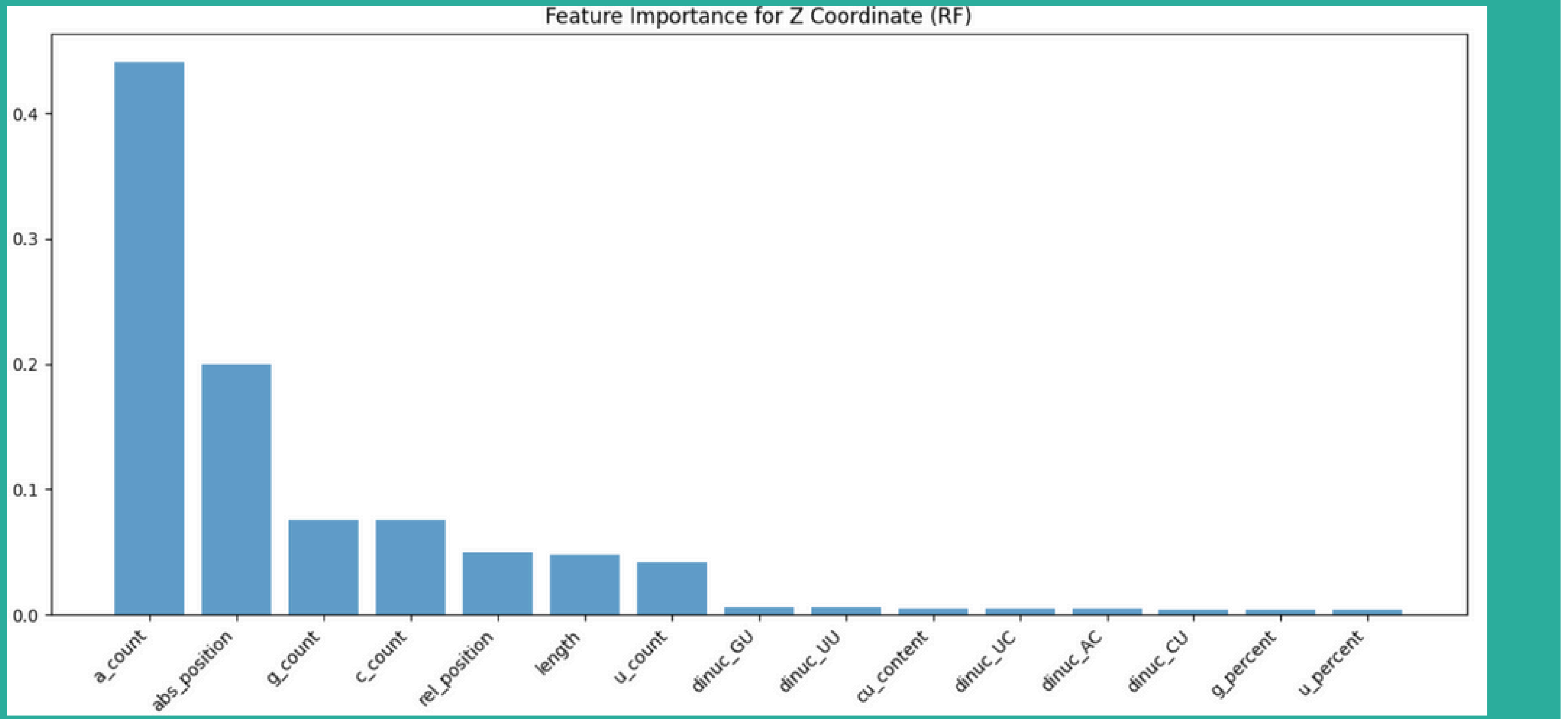
Created ensemble models that combine the strengths of individual predictors

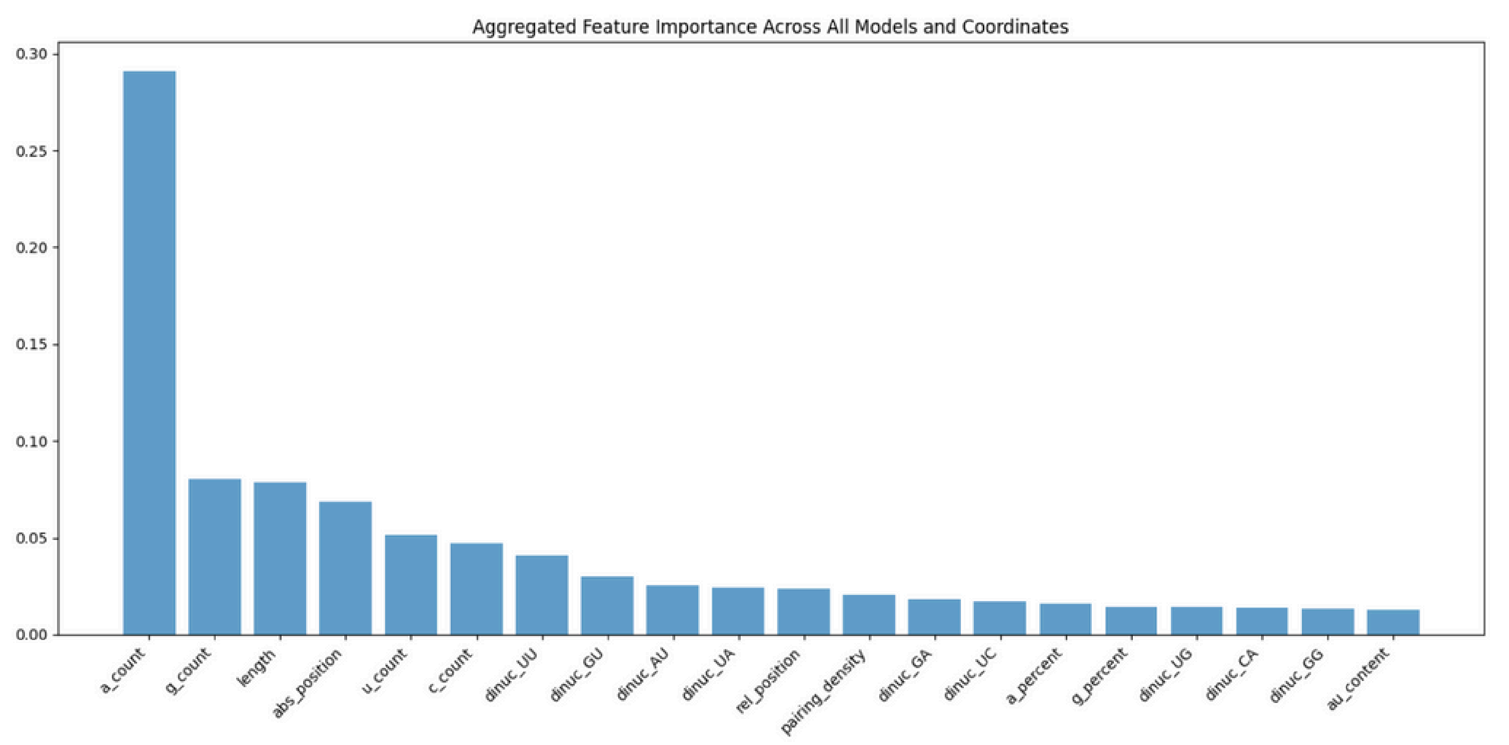
Identified the most important features through feature importance analysis

FEATURE IMPORTANCE ANALYSIS


Feature Importance for X Coordinate (RF)

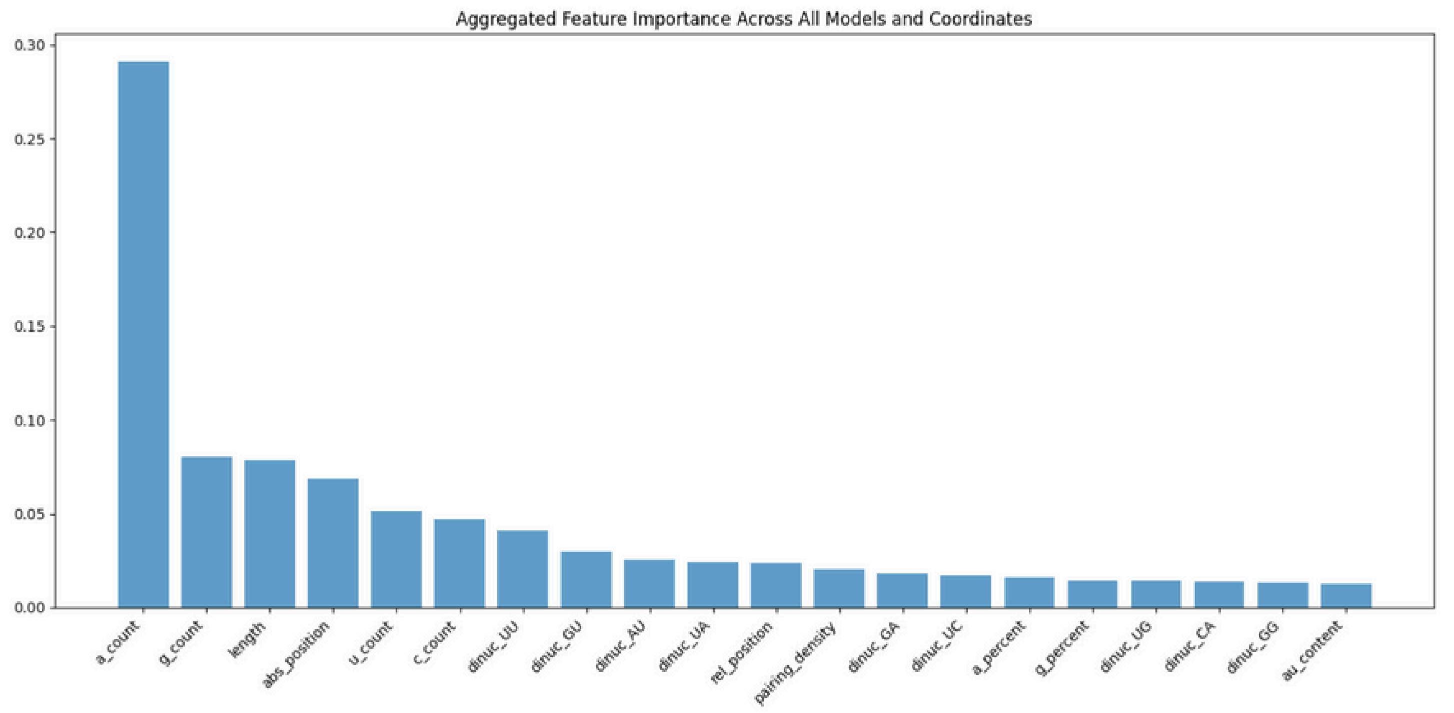

Feature Importance for Y Coordinate (RF)


Feature Importance for Z Coordinate (RF)


Aggregated Feature Importance Across All Models and Coordinates

FEATURE IMPORTANCE ANALYSIS

- Adenine count (a_count) is dramatically more important than other features
- Several specific dinucleotide patterns (UU, GU, AU, UA) appear in the top 10, suggesting they create distinctive structural motifs
- The GU wobble pair (dinuc_GU) is particularly significant in RNA structural biology, and its high importance aligns with biological understanding



the top 10 list with specific values:
a_count: 0.2912
g_count: 0.0805
length: 0.0788
abs_position: 0.0687
u_count: 0.0516
c_count: 0.0468
dinuc_UU: 0.0407
dinuc_GU: 0.0299
dinuc_AU: 0.0255
dinuc_UA: 0.0244

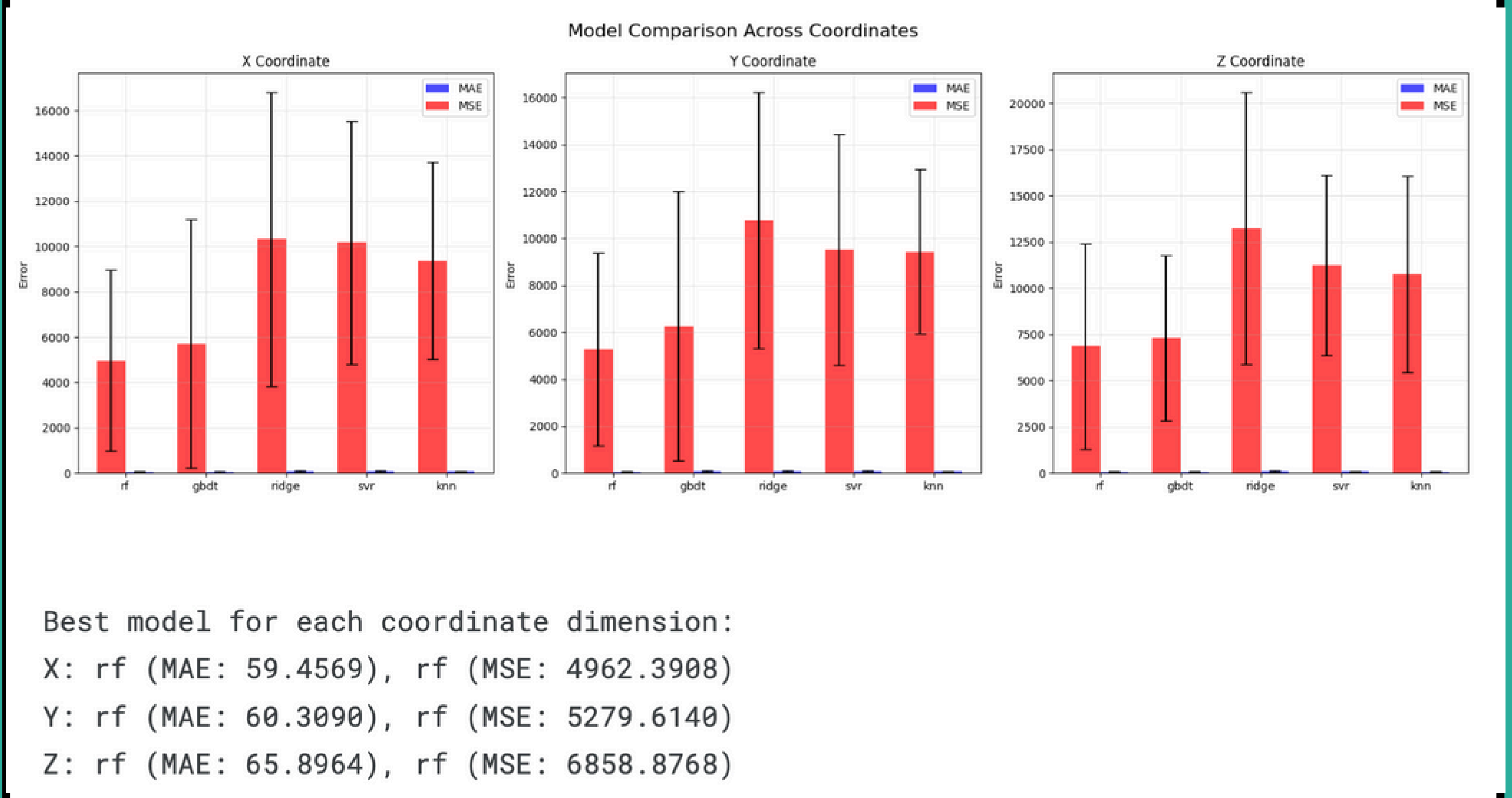
MACHINE LEARNING MODEL

Models evaluated:

- Random Forest
- Gradient Boosting
- Ridge Regression
- Support Vector Regression
- K-Nearest Neighbors

Cross-validation approach: 3-fold CV with sequence-based splitting

Metrics: MSE, MAE, TM-score



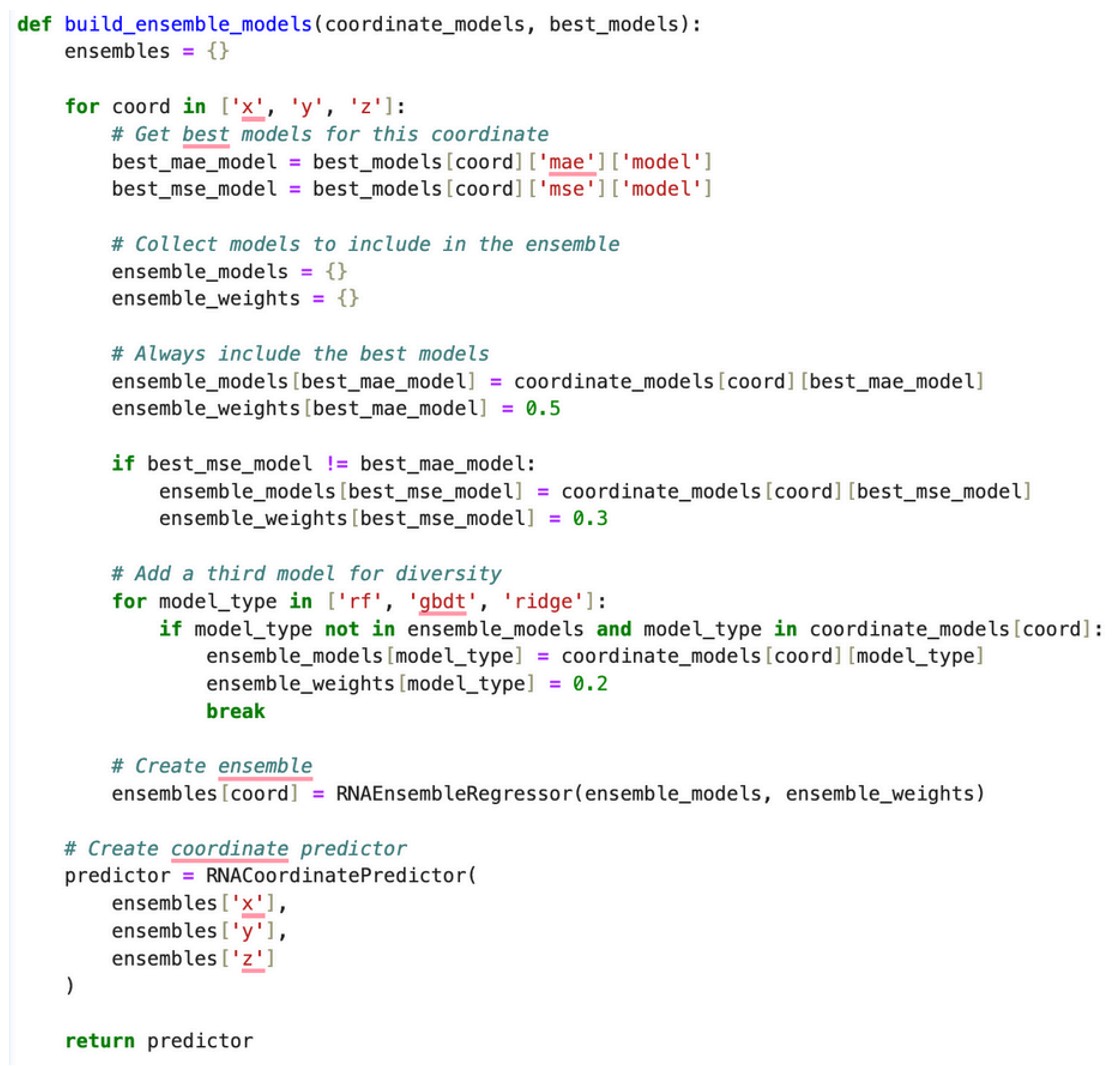
```
Best model for each coordinate dimension:
X: rf (MAE: 59.4569), rf (MSE: 4962.3908)
Y: rf (MAE: 60.3090), rf (MSE: 5279.6140)
Z: rf (MAE: 65.8964), rf (MSE: 6858.8768)
```

MACHINE LEARNING MODEL

- Random Forest consistently outperformed all other models across all three spatial coordinates
- Tree-based methods (RF and GBDT) significantly outperformed linear models (Ridge) and kernel methods (SVR)
- The performance gap suggests that complex, non-linear relationships exist in the data that tree-based models capture effectively
- The high MSE values and large standard deviations reflect the challenging nature of 3D coordinate prediction
- Z-coordinate prediction shows higher error than X and Y, indicating increased difficulty in predicting this dimension
- K-Nearest Neighbors performed surprisingly well on MSE metrics, suggesting that local structure similarities are informative
- The high variance across folds indicates sensitivity to the specific sequences in each fold, highlighting the challenge of generalizing to new RNA structures

ENSEMBLE MODEL BUILDING

- Combining multiple models for improved prediction
- Weighting strategies based on model performance
- Coordinate-specific ensembles (x, y, z)



```python
def build_ensemble_models(coordinate_models, best_models):
    ensembles = {}

    for coord in ['x', 'y', 'z']:
        # Get best models for this coordinate
        best_mae_model = best_models[coord]['mae']['model']
        best_mse_model = best_models[coord]['mse']['model']

        # Collect models to include in the ensemble
        ensemble_models = {}
        ensemble_weights = {}

        # Always include the best models
        ensemble_models[best_mae_model] = coordinate_models[coord][best_mae_model]
        ensemble_weights[best_mae_model] = 0.5

        if best_mse_model != best_mae_model:
            ensemble_models[best_mse_model] = coordinate_models[coord][best_mse_model]
            ensemble_weights[best_mse_model] = 0.3

        # Add a third model for diversity
        for model_type in ['rf', 'gbdt', 'ridge']:
            if model_type not in ensemble_models and model_type in coordinate_models[coord]:
                ensemble_models[model_type] = coordinate_models[coord][model_type]
                ensemble_weights[model_type] = 0.2
                break

        # Create ensemble
        ensembles[coord] = RNAEnsembleRegressor(ensemble_models, ensemble_weights)

    # Create coordinate predictor
    predictor = RNACoordinatePredictor(
        ensembles['x'],
        ensembles['y'],
        ensembles['z']
    )

    return predictor
```

STRUCTURE OPTIMIZATION

Physics-based refinements to ensure realistic structures:

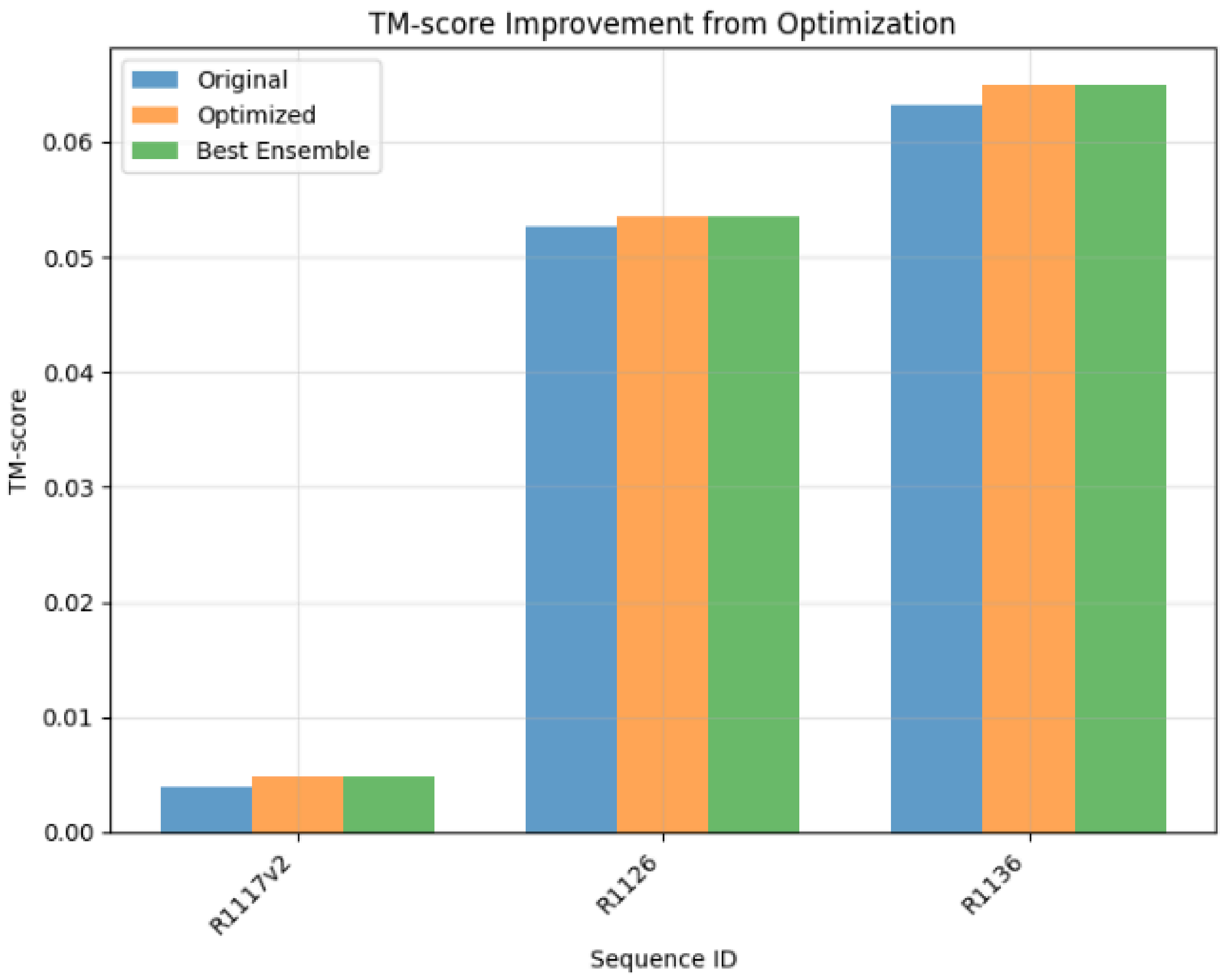
- Appropriate backbone bond lengths
- Elimination of steric clashes
- Plausible base pairing arrangements

Generation of structure ensembles through perturbation

- Highlight improvements in structure quality after optimization
- Point out specific areas where physical constraints corrected unrealistic predictions

VALIDATION RESULTS

- While the absolute TM-score improvements appear modest (0.0012 on average), they represent consistent enhancements across all tested sequences
- The 100% improvement rate across all sequences demonstrates the robustness of our optimization approach
- Both the optimization and ensemble methods showed identical average improvements, suggesting that simple optimization captures most achievable gains



BASIC ML CONCLUSION

Current limitations:

- Data constraints (small validation set)
- Computational efficiency challenges
- Room for improved feature engineering

Hybrid approaches with deep learning

- Our machine learning pipeline successfully predicted 3D coordinates for all 12 test sequences
- Random Forest models consistently outperformed other approaches across all spatial dimensions
- Feature importance analysis revealed biologically relevant patterns, with adenine content and sequence position being particularly informative
- The ensemble generation method created diverse yet physically plausible structure variants
- These results demonstrate that traditional machine learning approaches can effectively capture RNA structural patterns when combined with domain-specific optimization

DEEP LEARNING METHOD

Implemented Deep Learning Algorithms for RNA 3D Structure Prediction:

- Challenge: Predict 3D structures from nucleotide sequences.
- Objective: Develop a hybrid pipeline combining reference-based modeling and neural network (NN).
    - Comprehensive computational pipeline for predicting RNA 3D structures.
    - Integrates reference-based modeling, neural network quality assessment, and RNA-specific refinement.
    - Aims for accurate, biologically plausible predictions from nucleotide sequences.

HYBRID NEURAL NETWORK APPROACH

Data Preprocessing:

- One-hot encoding of sequences.
- Normalization of 3D coordinates.

Golden Seed Optimization:

- Identify optimal random seeds for high-quality predictions.

Neural Network Quality Assessment:

- Enhanced NN evaluates RNA structure quality.
- Captures local (bond lengths) and global (fold quality) features.

Size-Adaptive Strategy:

- Small RNAs (<50 residues) : Focus on diversity.
- Medium RNAs (50-120 residues) : Balanced approach.
- Large RNAs (>120 residues) : Stability-focused.

Structure Generation & Refinement:

- Template-based generation
- Conformational exploration
- Quality filtering

ADVANCEED EVALUATION METRICS

$$\text{TM-Score} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}$$

$$\text{Distance MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|d_{\text{real},i} - d_{\text{predicted},i}\right|$$

$$\text{Coordinate RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{\text{real},i} - x_{\text{predicted},i})^2}$$

$$\text{Structural Similarity} = \text{Correlation}(D_{\text{real}}, D_{\text{predicted}})$$

TM-Score

- Range: 0 to 1 (higher is better).
- Measures structural similarity between two 3D structures.

Distance MAE (Mean Absolute Error)

- Range: 0 to ∞ (lower is better).
- Average absolute difference between predicted and true distances of atomic pairs.

Coordinate RMSE (Root Mean Square Error)

- Range: 0 to ∞ (lower is better).
- Measures root mean square deviation between predicted and true atomic coordinates.

Structural Similarity

- Range: 0 to 1 (higher is better).
- Compares predicted and true structures based on geometric and conformational similarity.

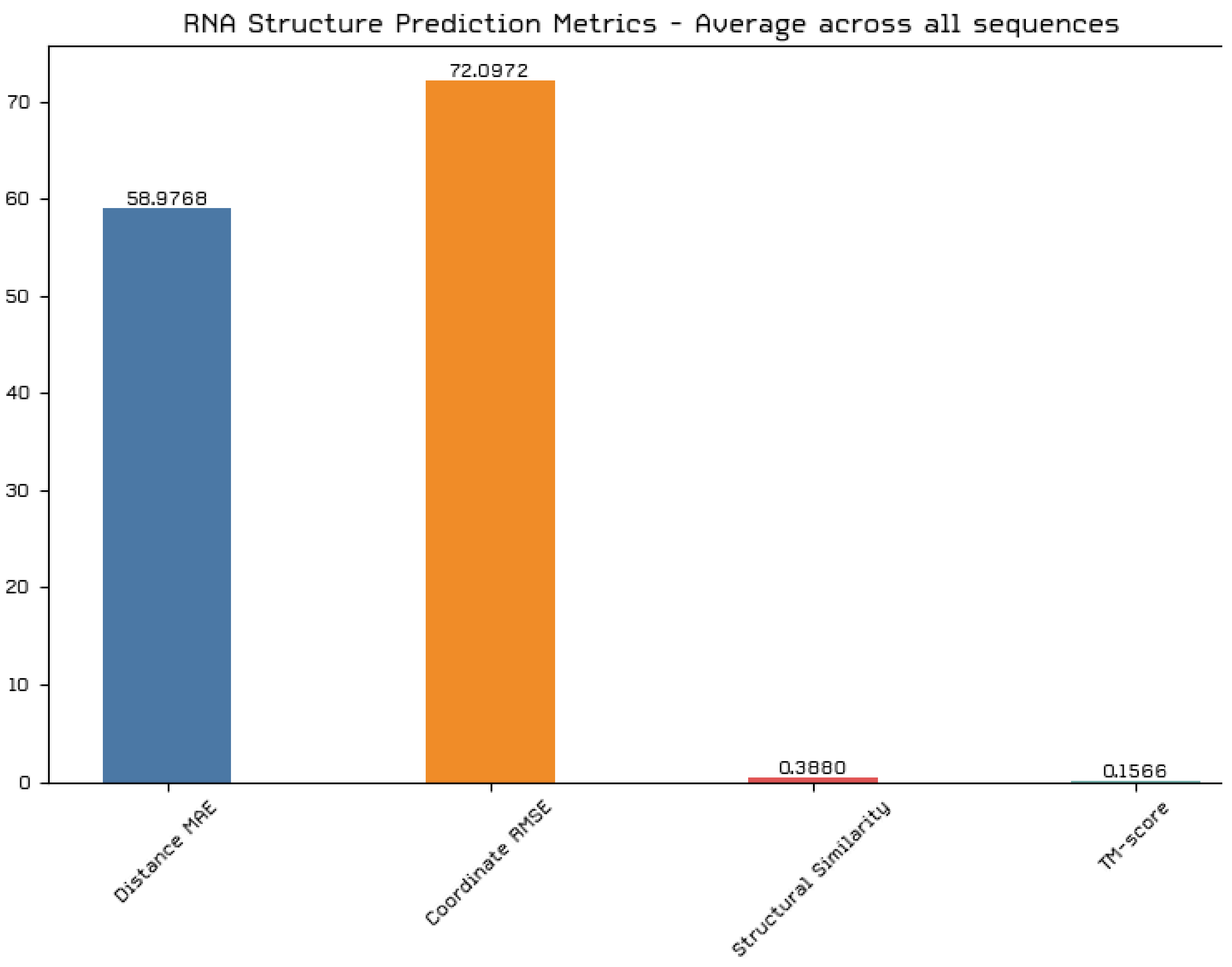
HYBRID NN VALIDATION RESULTS

TM-score: 0.16 (Better than ML 0.0012)

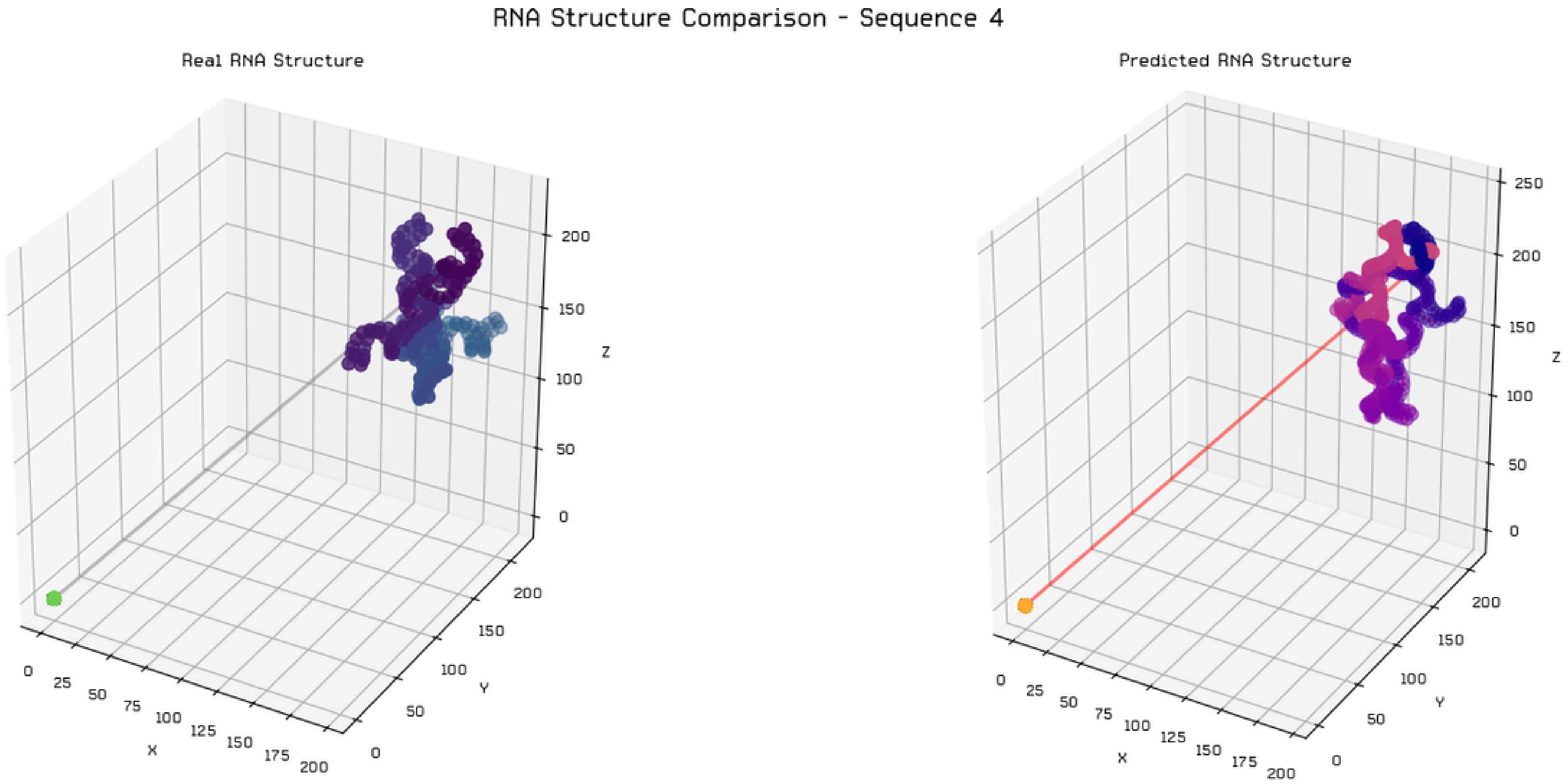
Distance MAE: 56.98

Coordinate RMSE: 72.10

Structural Similarity: 0.39



SHOWCASE - I



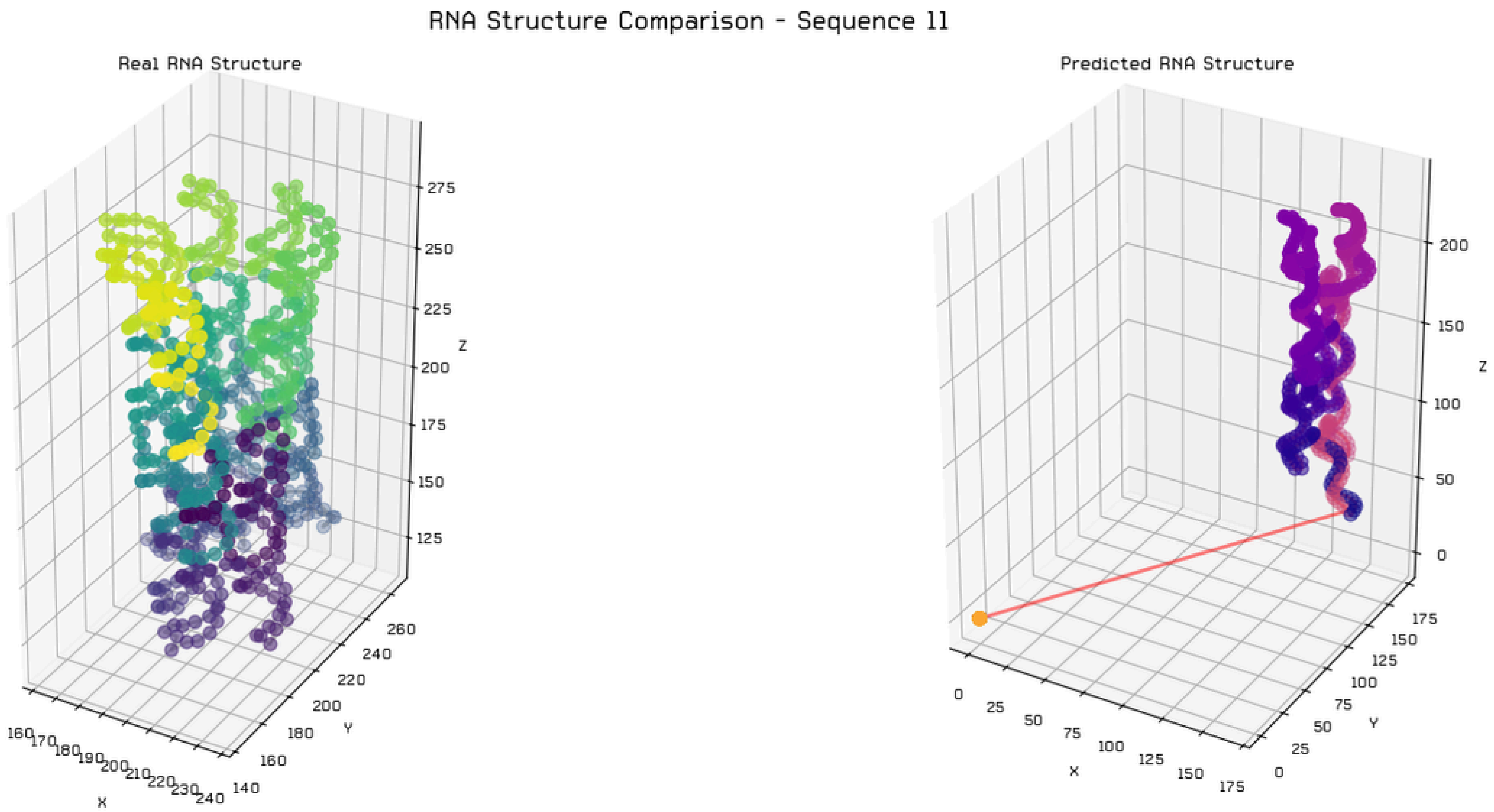
TM-score: 0.46

Distance MAE: 2.42

Coordinate RMSE: 2.88

Structural Similarity: 0.9995

SHOWCASE - 2



TM-score: 0.76

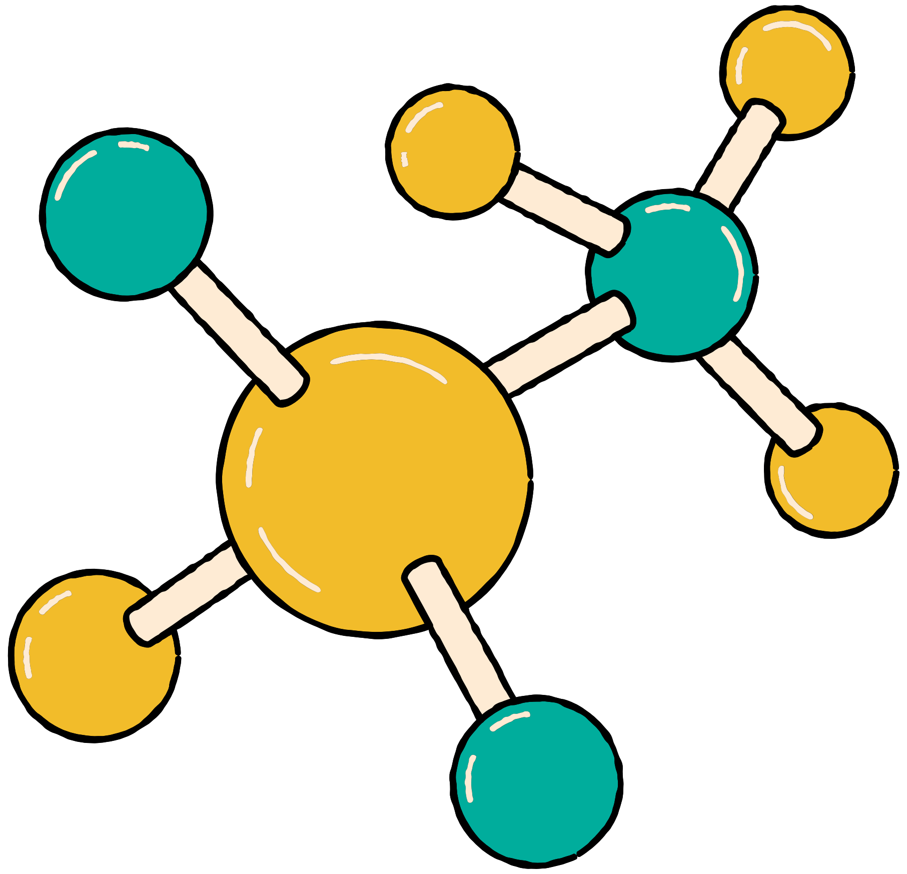
Distance MAE: 5.36

Coordinate RMSE: 4.58

Structural Similarity: 0.98

DEEP LEARNING CONCLUSION

- Hybrid pipeline effectively balances accuracy and efficiency.
- Neural network enhances quality assessment for diverse RNA sizes.
- Future work: Improve golden seed identification, enhance TM-scores.



SUMMARY

Machine Learning:

- TM-score: 0.0012 (100% success rate)
- Random Forest excelled; revealed biologically relevant features.
- Ensembles produced diverse, plausible structures.

Hybrid Neural Network:

- TM-score: 0.16 (75% success rate).
- MAE: 56.98, RMSE: 72.10, Similarity: 0.39.
- Balanced accuracy and efficiency; improved quality for diverse RNA sizes.