Machine Learning in Business

Protein Structure: Machine Learning nPre

By: Lilith Froude, Nitin Devasahayam, David Cho, Tarek Elbissat, Pratiksha Theodore

Why ML for Protein Analysis?

1

Massive Bottleneck

High cost, slow experimental methods for protein 3D structure.

2

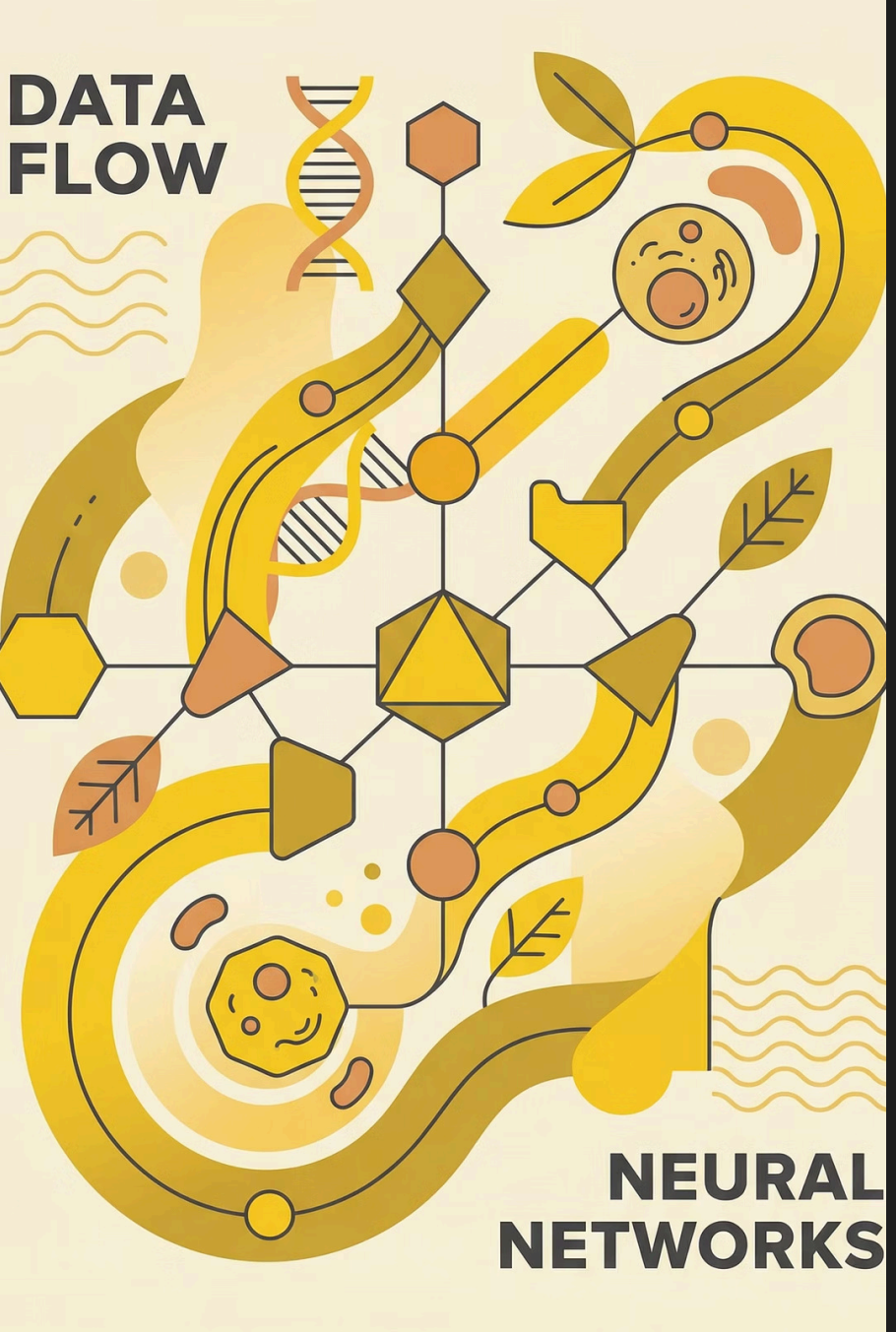
Drug Discovery

Labs need fast, reliable prediction for new drug candidates.

3

Clinical Oncology

Doctors need to triage 'driver' vs. 'passenger' mutations in cancer.



ML Solution: Computational Triage

Our system predicts protein structure and stability from amino acid sequences.

BiLSTM

Predicts secondary structure (helix, sheet, coil).

Random Forest

Classifies proteins as stable or unstable.

Isolation Forest

Detects anomalous proteins (mutation damage).

Data Source: PISCES Dataset

Experimentally determined protein structures from the Protein Data Bank (PDB).

1

Training Dataset

15,079 protein chains

2

Testing Dataset

Separate PISCES PC25 subset

3

Avg. Sequence Length

249.66 amino acids

4

Experimental Method

100% X-ray Crystallography

Key Data Characteristics

Sequence Length Range	1 to 2,128 amino acids
Resolution Range	0.48 Å to 2.5 Å (lower = better quality)
Experimental Method	100% X-ray Crystallography

Key Data Fields

seq	Amino acid sequence (20 standard amino acids)
sst3	3-state secondary structure labels (C=Coil, H=Helix, E=Sheet)
sst8	8-state DSSP secondary structure classification
len_x	Sequence length in amino acids
resol	Crystal resolution in Angstroms
rfac & freerfac	Model accuracy metrics

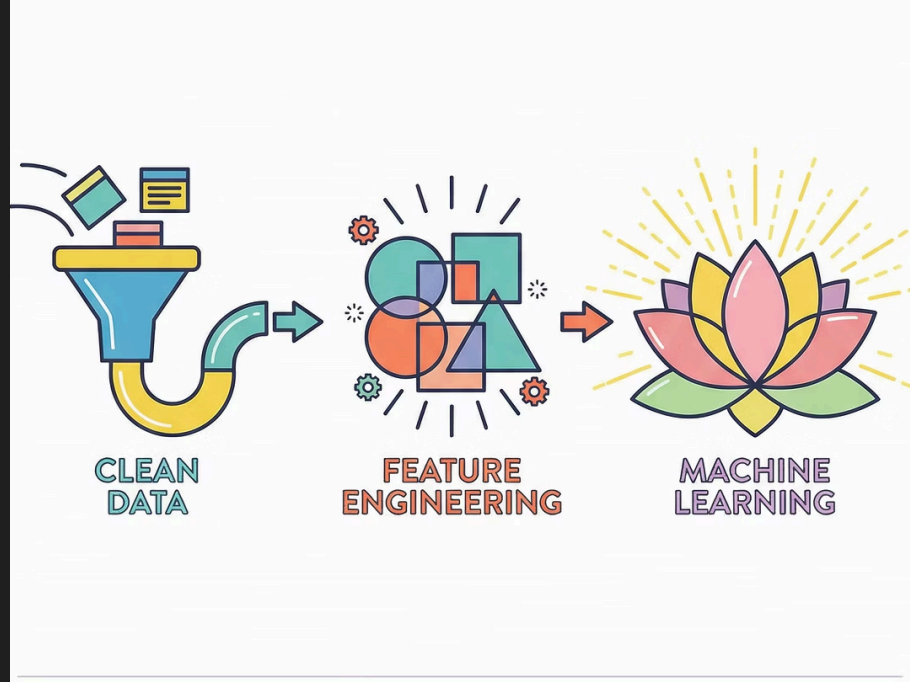
Data Preparation

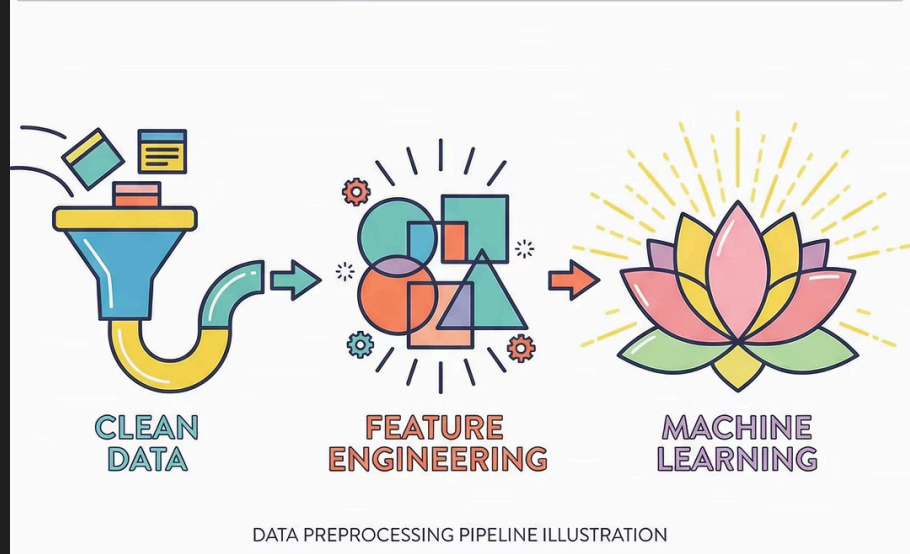
BiLSTM (Sequence-to-Sequence)

- Vocabulary Creation
- Sequence Encoding
- Padding/Truncating (MAX_SEQ_LEN = 534)
- One-hot encoded sst3 labels
- 85% Train / 15% Validation Split

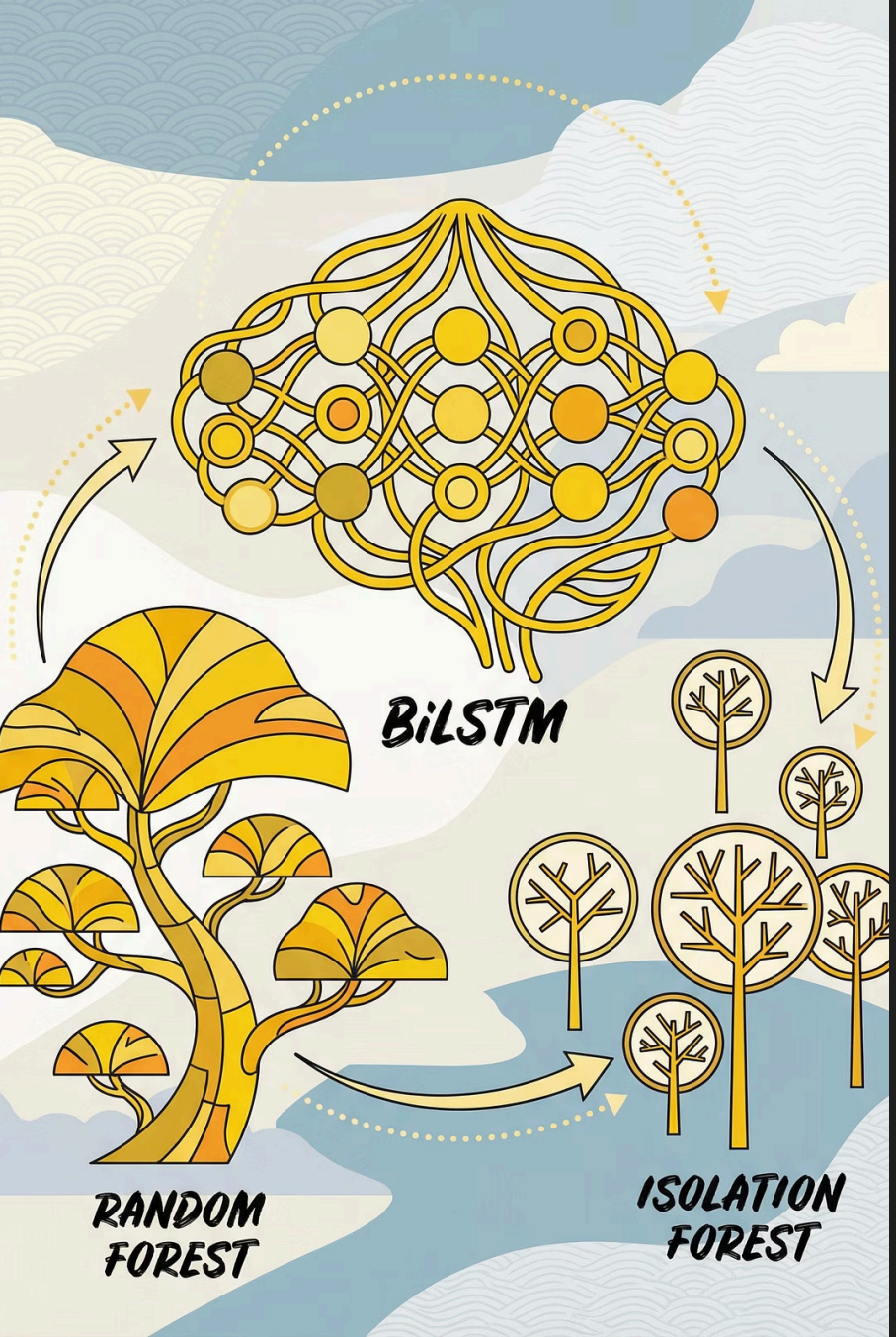
Random Forest / Isolation Forest

- Feature Engineering (structural ratios)
- Amino Acid Composition
- Stability Label Creation (engineered)
- Missing Value Imputation
- Feature Scaling





DATA PREPROCESSING PIPELINE ILLUSTRATION



Modeling Approach

1

BiLSTM

Sequence-to-sequence labeling for secondary structure. Captures bidirectional context.

2

Random Forest

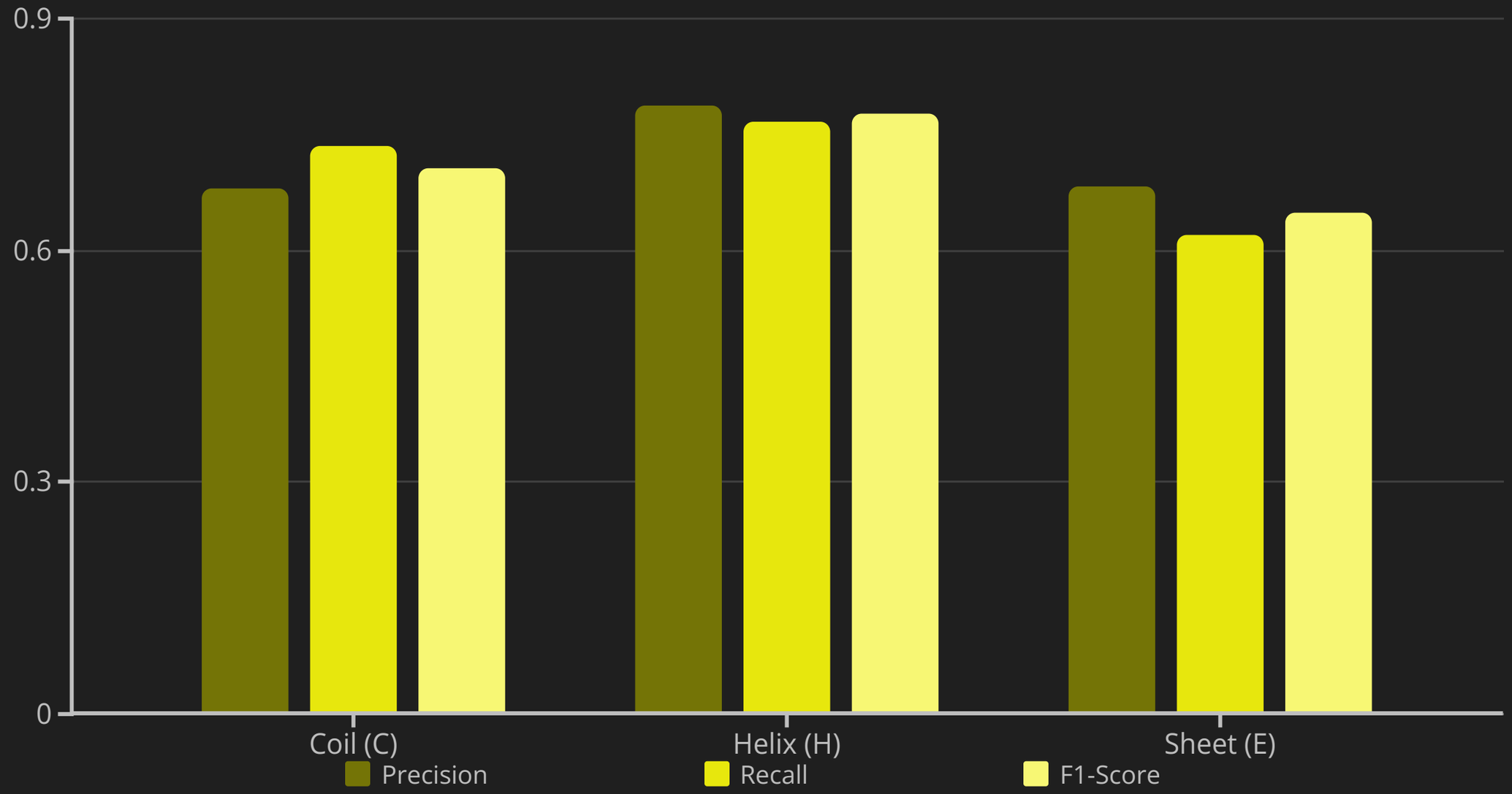
Ensemble method for stability classification. Handles tabular feature data.

3

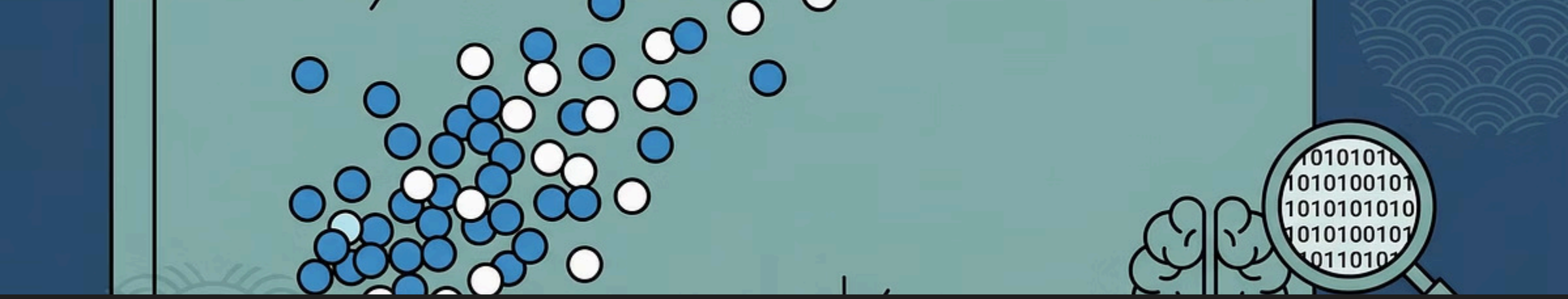
Isolation Forest

Cluster-aware anomaly detection. Identifies outliers within structural families.

BiLSTM Performance: 71.98% Q3 Accuracy



Helix prediction is most accurate, Sheet prediction is most challenging due to biological complexity.



Random Forest & Isolation Forest

Random Forest (Stability)

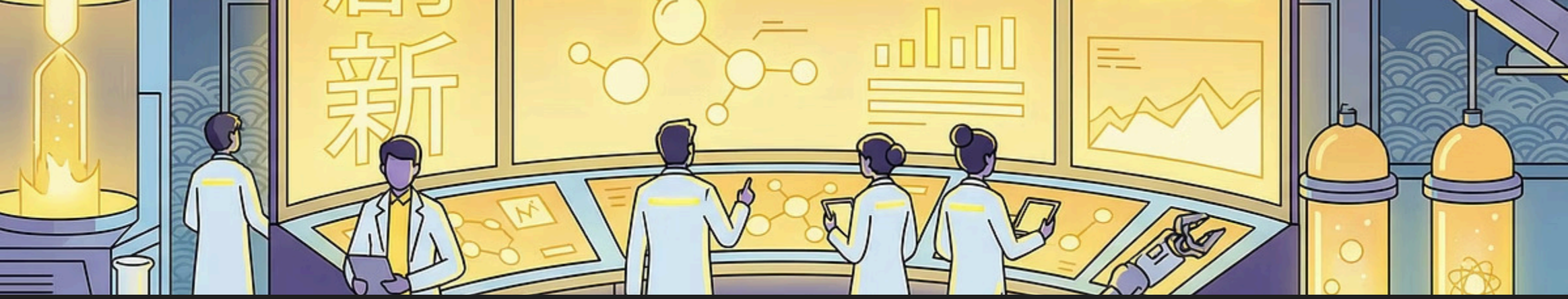
60% Accuracy, AUC-ROC 0.611.

Performs better than random, but room for improvement with experimental stability data.

Isolation Forest (Anomaly)

~5% proteins flagged as structural outliers.

Outliers show extreme coil dominance, short chains, unusual refinement stats.



Business Impact & ROI

Time Reduction

6-18 months to minutes for initial screening.

Cost Savings

$100K-$500K per structure to negligible compute.

Precision Medicine

Automated mutation scoring for personalized cancer treatment.

Increased Throughput

Thousands of predictions per day vs. 50-100/year/lab.

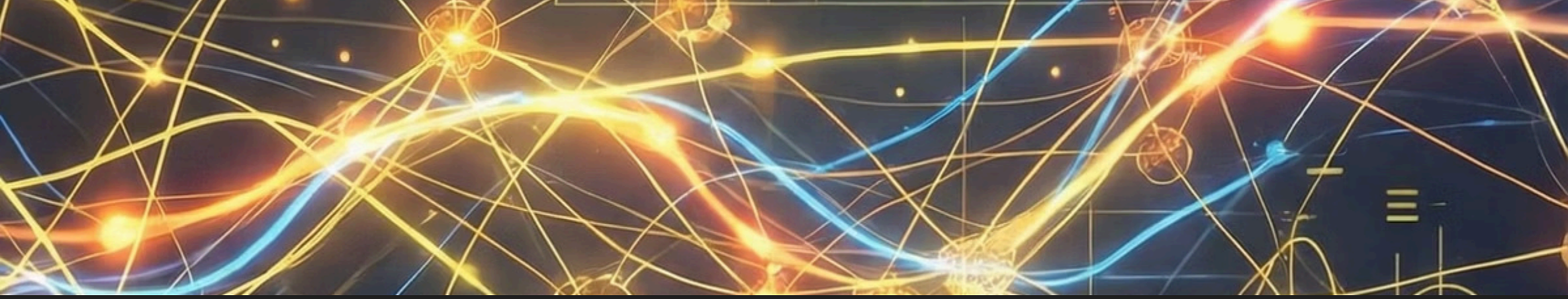
Future Work & Value

Future Enhancements

- Incorporate PSSM features
- Train on experimental stability data
- Explore Transformer architectures
- Add CNN layers for local motifs

Stakeholder Value

- **Pharma R&D:** Accelerate drug discovery
- **Clinical Oncologists:** Rapid mutation triage
- **Genomics Companies:** Scalable variant interpretation
- **Researchers:** Structural biology hypothesis generation



Measuring Impact & ROI Validation

Quantitative performance metrics from our ML models validate their impact and build a business case.

BiLSTM: Secondary Structure

71.98% overall accuracy, with 78.66% precision for helix prediction. Informs reliable computational triage.

Random Forest: Stability Prediction

60% accuracy (AUC-ROC: 0.611), outperforming random. Future use of experimental data will significantly enhance drug candidate selection ROI.

Isolation Forest: Anomaly Detection

Flags ~5% structural outliers, identifying high-risk candidates. Directly supports cost savings by early filtering of unstable proteins.

ROI depends on comparing experimental cost savings against ML development investment. Better experimental data will improve accuracy and ROI.

Deployment Strategy & Risk Management

Deployment Architecture

Models are deployed via a REST API endpoint (Flask/FastAPI) accepting raw amino acid sequences and returning JSON output for secondary structure, stability, and anomaly scores. This allows seamless integration into existing LIMS or clinical genomics pipelines.

Key Considerations

- BiLSTM requires GPU for efficient inference; RF/IF run on CPU.
- Latency: 1-5s for structure, <100ms for stability/anomaly.
- Robust model versioning and periodic retraining are crucial.
- Monitoring prediction confidence helps detect data drift.

Risks & Mitigations

Overfitting	Separate test set validation, cross-validation.
Class Imbalance	Class weighting, focal loss for BiLSTM.
Engineered Stability	Validate against experimental ΔG data.
Sequence Length Limit	Sliding window or retrain with larger max_len.
Interpretability	Attention visualization, SHAP values for RF.

Ethics in Clinical AI

Managing the inherent risks of ML deployment is crucial for successful and ethical implementation in real-world scenarios.

1

Algorithmic Bias

Ensure training data is diverse. Use bias detection and human review for critical predictions.

2

Misinterpretation by End Users

Provide clear training and documentation. Include confidence scores and expert consultation options.

3

Data Quality Dependency

Establish data validation pipelines and source checks. Have backup plans for incomplete data.

Balancing ML automation with human oversight is crucial for safe and ethical application.