VN-UK Institute for Research & Executive Education,
The University of Danang - Faculty of Information and Communication Technology

Course Title: DATA ANALYTICS FOR LIFE SCIENCE

Project Title: Applying Deep Learning (CNN-LSTM) and Explainable AI (XAI) in Predicting the Cutting Efficiency of the CRISPR-Cas9 System

Student Investigators:
Trần Xuân Cường - 22040004
Nguyễn Vân Nhi - 22040006

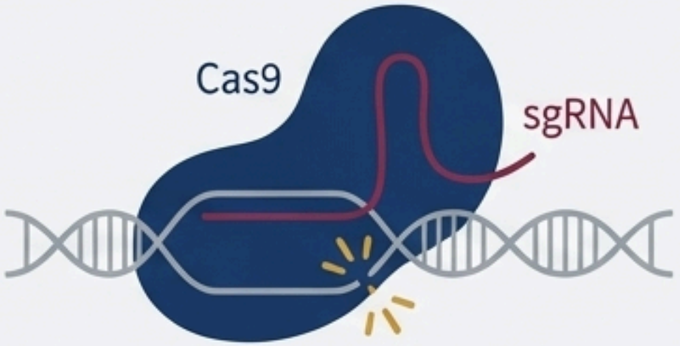
Supervisors:
MSc. Nguyễn Chí Thiện
PhD. Trần Thanh Hòa

Table of Contents

The CRISPR-Cas9 Revolution and Its Core Challenge

The Revolution: CRISPR-Cas9

A groundbreaking gene-editing technology allowing scientists to “cut” and “paste” DNA with unprecedented precision.



The Greatest Challenge: Designing the Guide RNA (sgRNA)

The effectiveness of CRISPR depends entirely on the design of the guide RNA (sgRNA) that directs the Cas9 enzyme to the correct location.

The Core Problem: How do we design an optimal sgRNA that maximizes on-target efficiency (cutting at the right place) while minimizing off-target effects (unwanted cuts elsewhere)? This is the central challenge we are addressing.

Our Mission: The Four Core Objectives

1. Build a High-Performance Deep Learning Model

Develop a hybrid neural network architecture (CNN-LSTM) to learn complex patterns from DNA sequences and accurately predict the sgRNA's "Efficiency Score."

2. Integrate Explainable AI (XAI)

Overcome the "black box" problem. Our system not only predicts but also reveals *which nucleotides* are decisive, using Saliency Maps to help researchers understand the underlying biology.

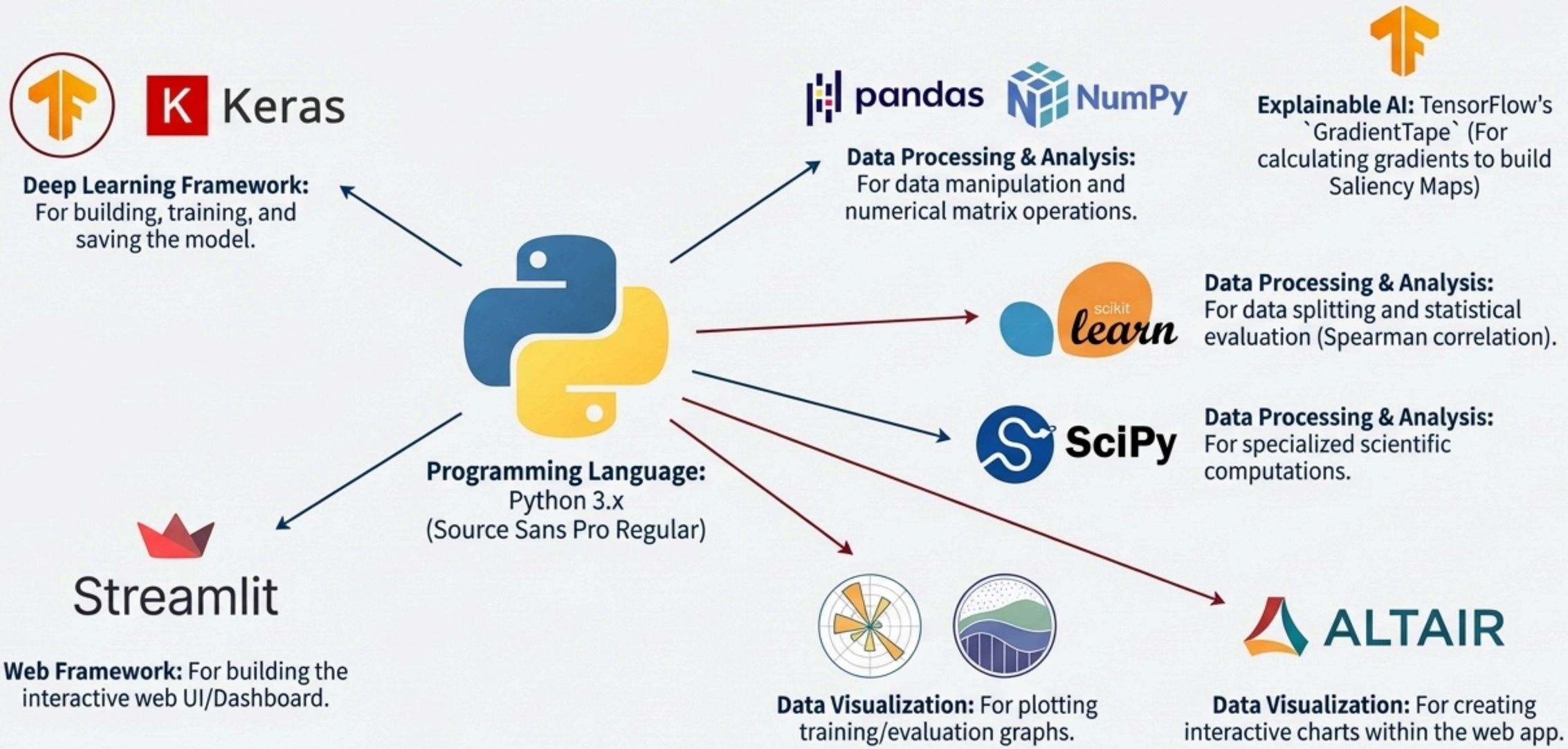
3. Develop a Practical Application (Deployment)

Package the model into an intuitive Streamlit web application, allowing users to easily input sequences, scan genes, and receive automated analysis reports.

4. Automate Biological Insights (NLG)

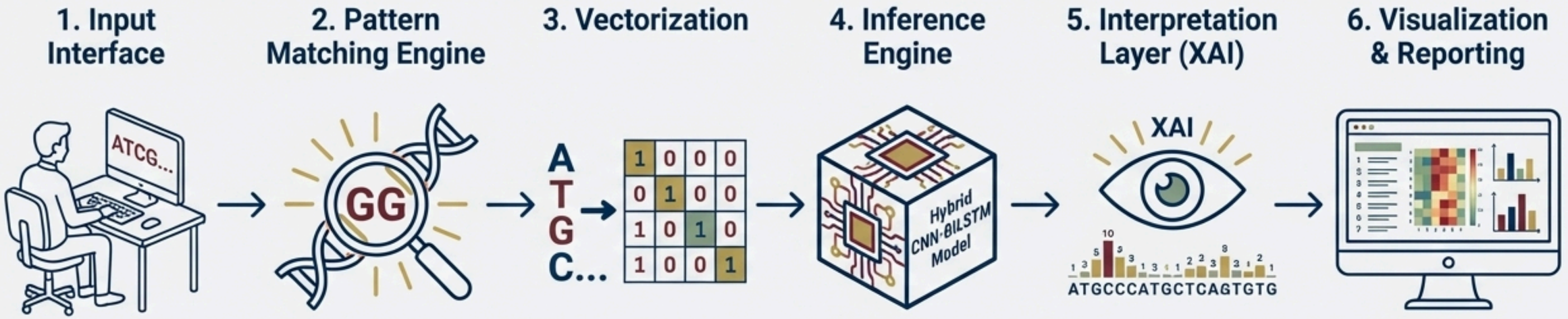
Use Natural Language Generation (NLG) techniques to translate technical model outputs into clear, understandable biological recommendations for end-users.

The Technology Stack



The End-to-End Pipeline: From Sequence to Insight

Our system is a seamless, end-to-end pipeline:



1. Input Interface

User provides the target DNA sequence.

2. Pattern Matching Engine

A sliding window algorithm scans for the "GG" PAM motif and extracts the preceding 23 nucleotides to form an sgRNA candidate.

3. Vectorization

The sequence of letters (A, T, G, C) is converted into a numerical matrix via One-Hot Encoding.

4. Inference Engine

The matrix is fed into our trained Deep Learning model to calculate the Efficiency Score.

5. Interpretation Layer (XAI)

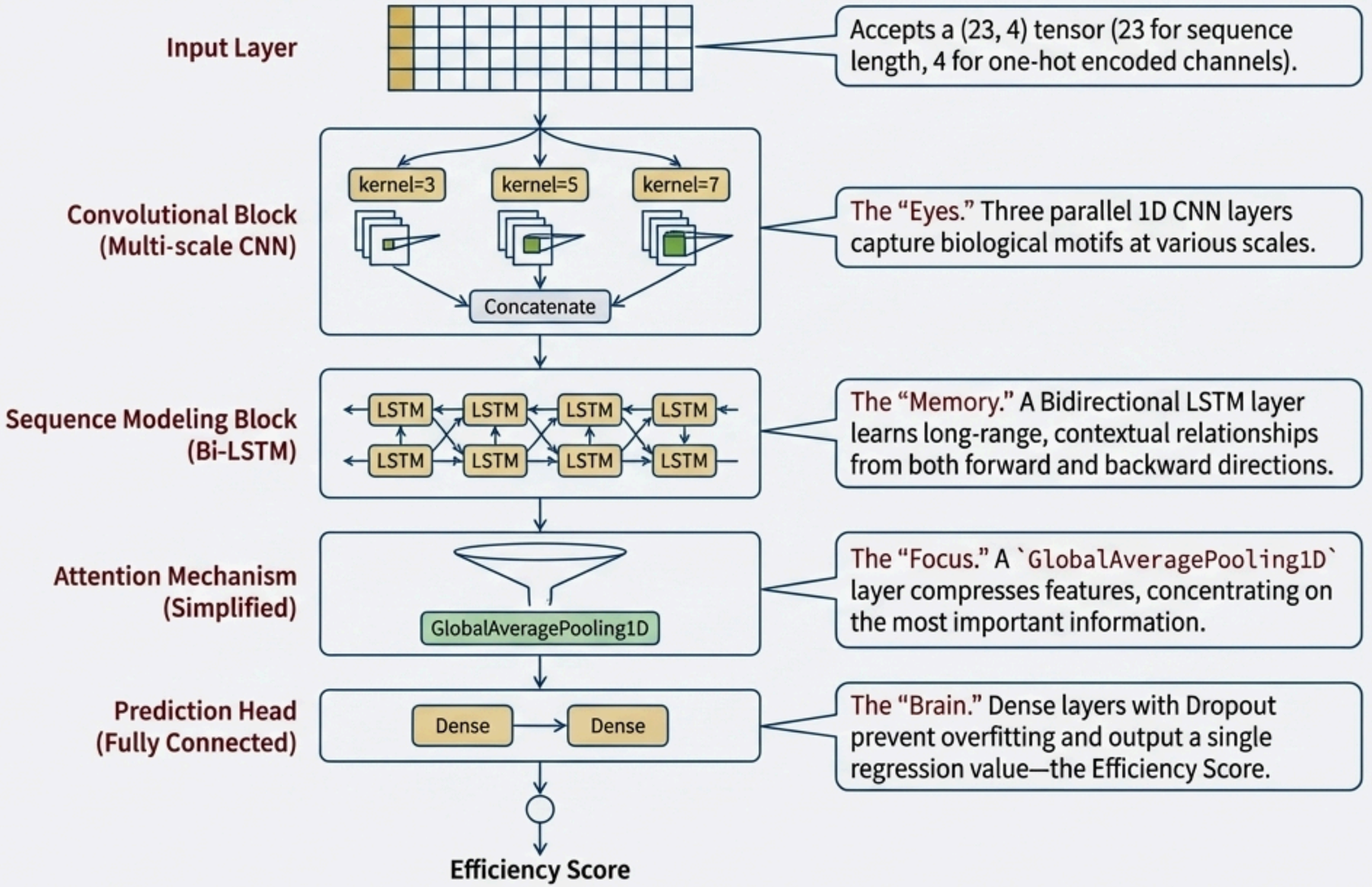
The XAI layer is activated to calculate the importance score of each nucleotide.

6. Visualization & Reporting

Results are compiled into ranked tables, heatmaps, and textual reports on the Streamlit dashboard.

The Heart of the System: Hybrid CNN-BiLSTM-Attention Model

The core of our system is a hybrid model designed for biological sequence data.



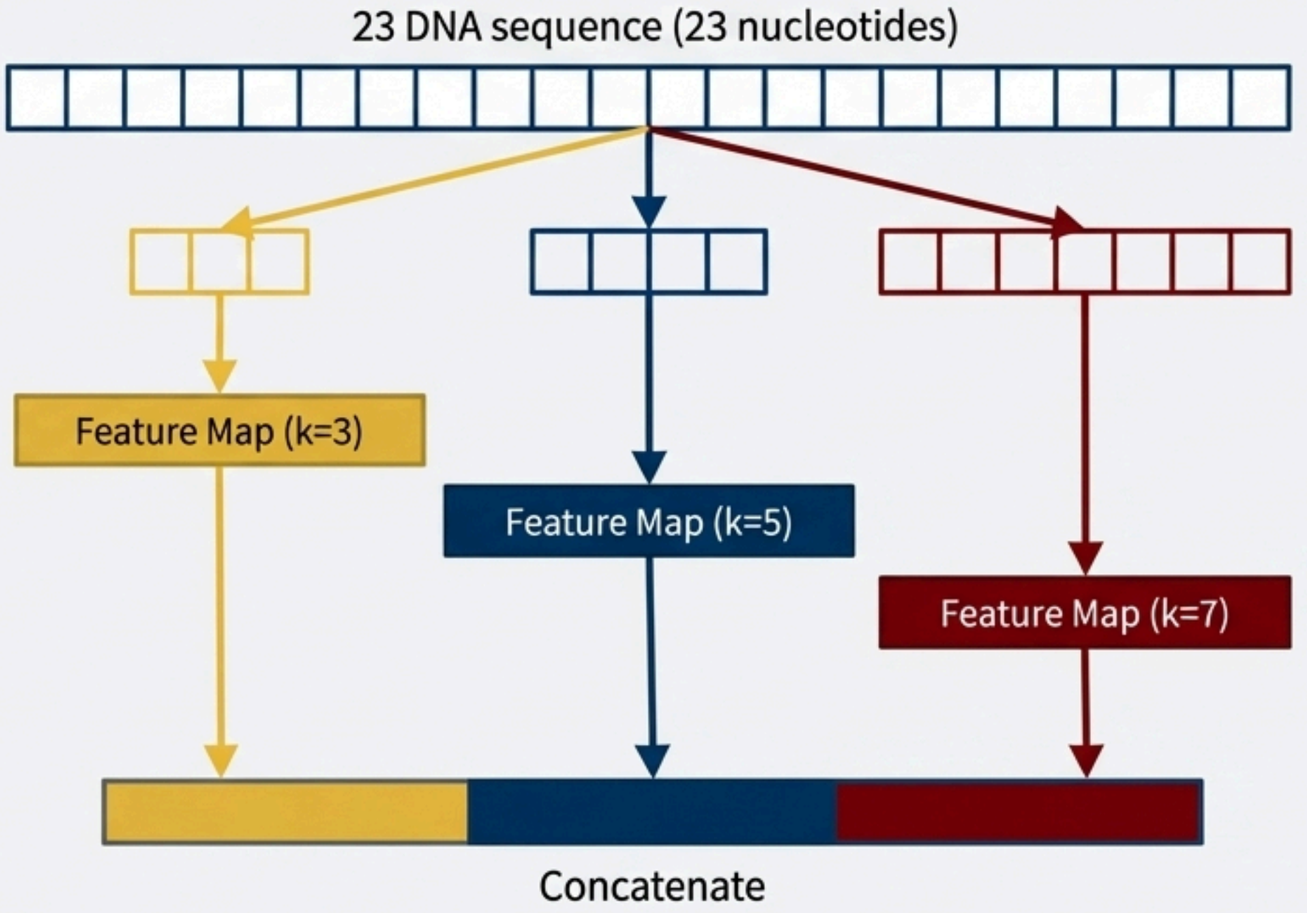
[Atomic Slide] The CNN Block: Learning Multi-Scale Features

Purpose: CNNs learn local 'motifs'—meaningful, recurring patterns of nucleotides.

Three-Branch Parallel Architecture:

- **Kernel size = 3:** A **small magnifying glass**, detecting very local, fine-grained motifs.
- **Kernel size = 5:** A **medium lens**, detecting interactions between nearby nucleotides.
- **Kernel size = 7:** A **wide-angle lens**, capturing broader contextual patterns.

Concatenation & Stabilization:
The outputs (feature maps) from all three branches are concatenated (merged), giving the model a comprehensive, multi-scale view. Each branch includes `BatchNormalization` to stabilize learning and `ReLU` activation for non-linearity.



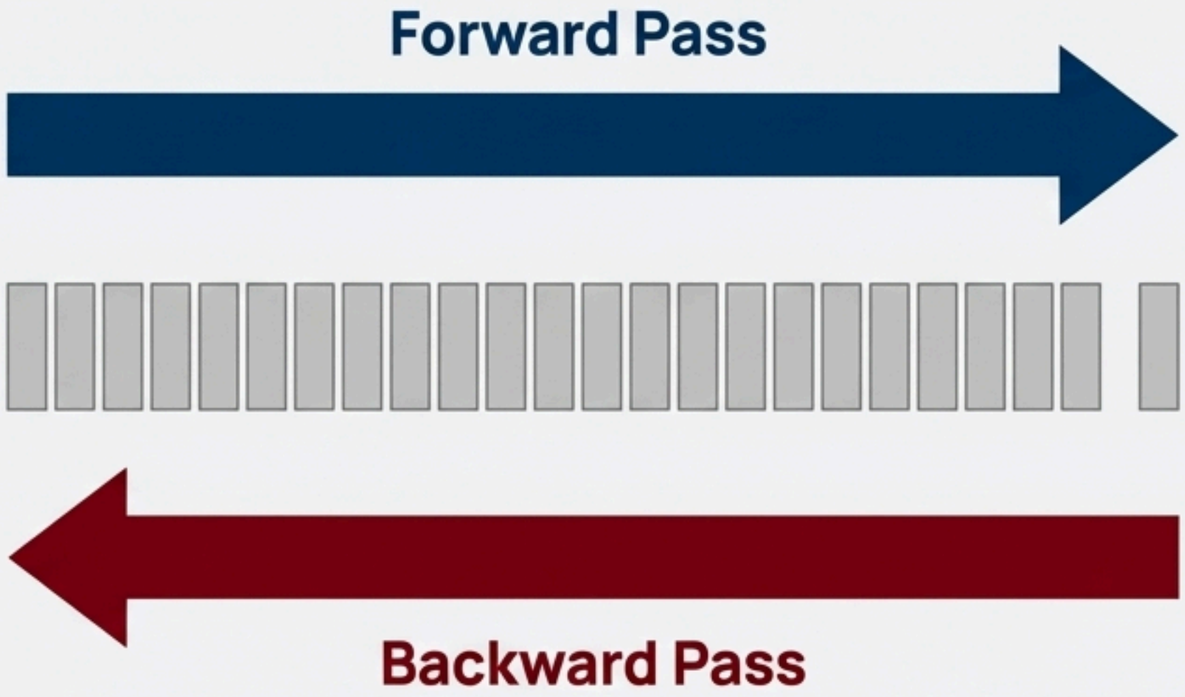
[Atomic Slide] The Bi-LSTM Block: Understanding Bidirectional Context

The Limitation of CNNs: CNNs see local patterns but don't inherently understand the overall order or long-range dependencies in the sequence.

The Role of Bi-LSTM: To learn these long-range dependencies and the full bidirectional context of the DNA sequence.

The Power of Bidirectionality: The model 'reads' the DNA from left-to-right (Forward Pass) and from right-to-left (Backward Pass). This is critical, as biological interactions can occur in both directions, allowing the model to understand the relationship between nucleotides at the start and end of the sequence.

Parameters: We use a Bidirectional LSTM layer with 128 hidden units.



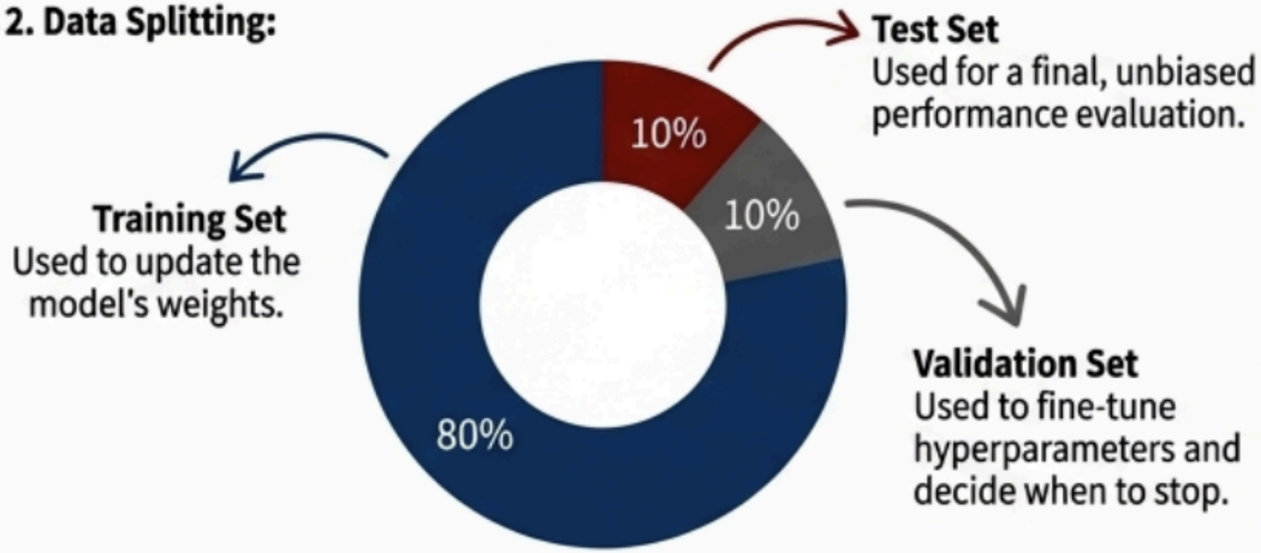
Caption: Understands the interaction between nucleotides at the beginning and end of the sequence.

The Rigorous Model Training Process

1. Data Preparation:

- Input data of 30bp sequences are sliced to the critical region (index 4 to 27) to create a standard 23bp input.
- **One-Hot Encoding:** Each nucleotide is converted into a 4-dimensional vector (e.g., A=[1,0,0,0]).

2. Data Splitting:



3. Training Setup:

- **Optimizer**: Adam with a learning rate of 0.0003.
- **Loss Function**: Mean Squared Error (MSE) for our regression task.

4. Callbacks Strategy:

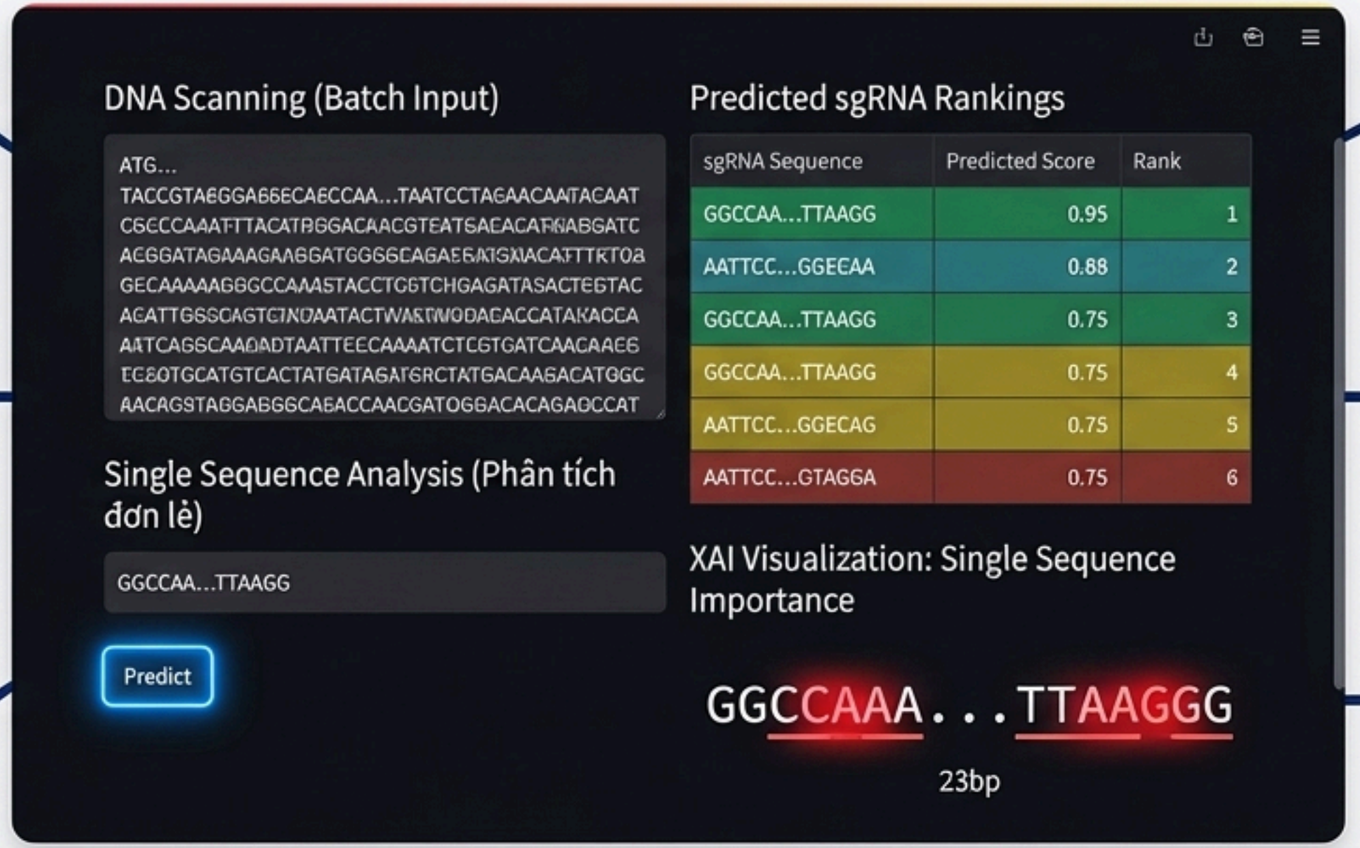
- `ModelCheckpoint`: Saves the best model based on the lowest validation loss.
- `EarlyStopping´: Halts training if validation loss doesn't improve for 15 consecutive epochs, preventing overfitting.

The Streamlit App: Putting AI Power in Your Hands

Platform: Built with Streamlit for rapid deployment.

Design: A custom 'Dark Biotech" theme for a professional, modern feel.

Backend: Python environment with TensorFlow caching for fast response times.



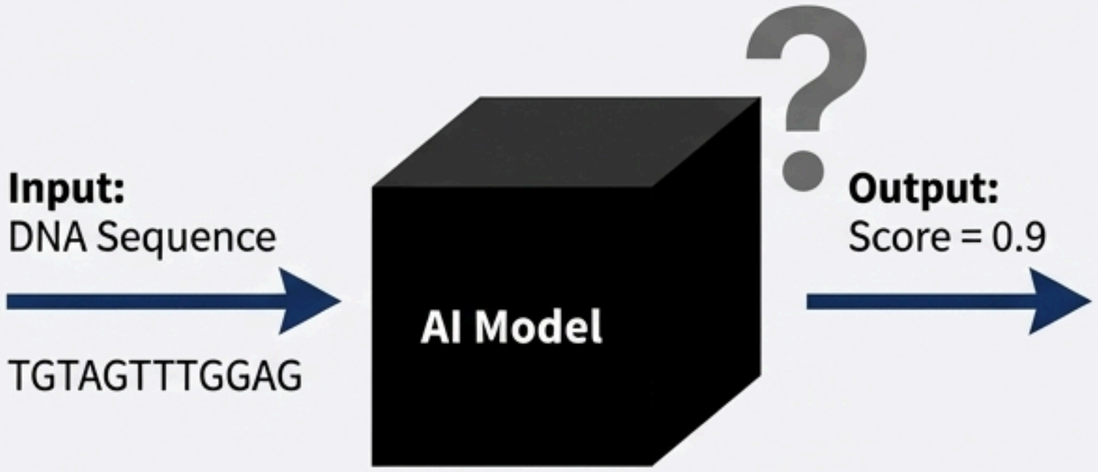
Input a long gene sequence for batch scanning.

Results are ranked by predicted efficiency score.

Instant XAI visualization shows which nucleotides are most important.

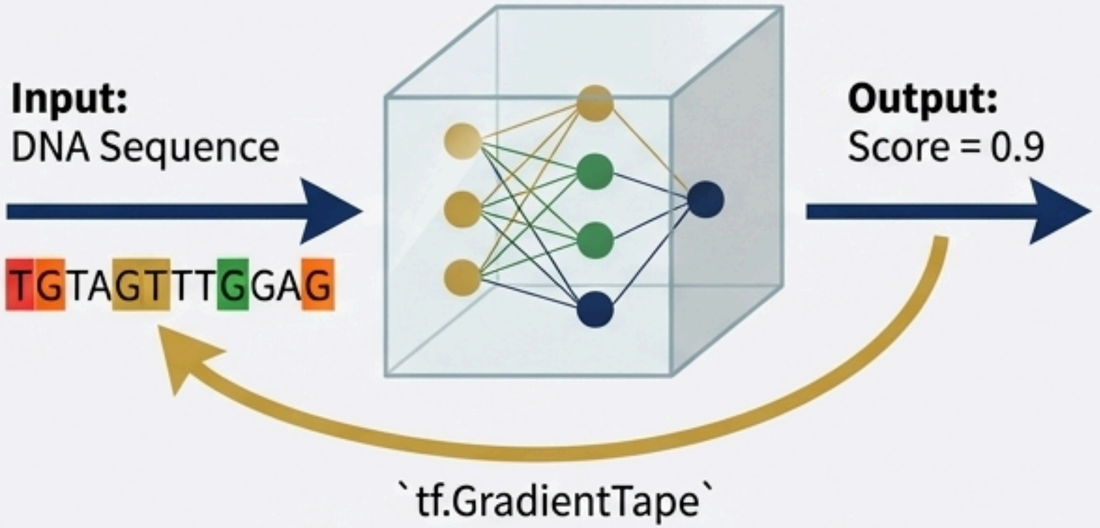
Unlocking the Black Box: How Explainable AI (XAI) Works

The Problem: The “Black Box”



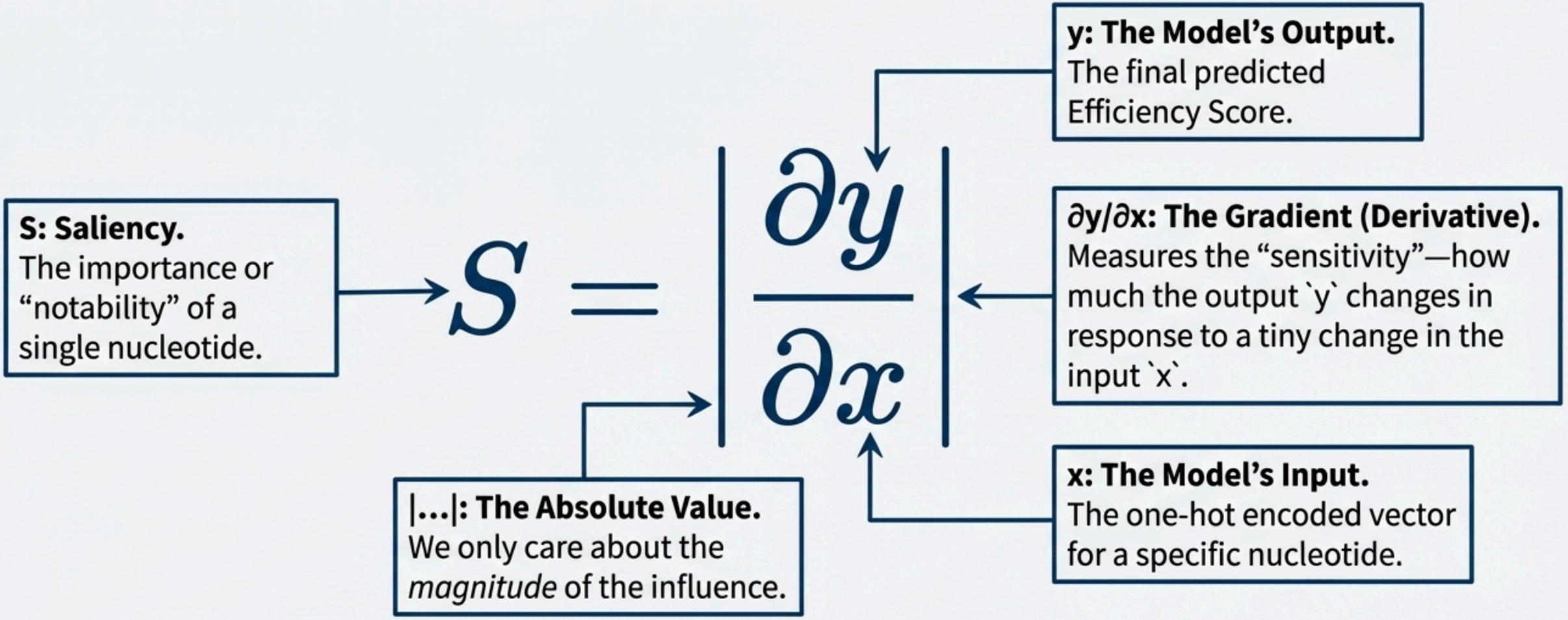
Deep Learning models are often opaque. They give an answer but don’t explain *why*.

The Solution: The Transparent “Glass Box” with Saliency Maps



We use a gradient-based Saliency Map technique. We ask the model: ‘If you could change one input nucleotide, how much would your final prediction change?’ The result is a ‘heatmap’ showing the importance of each nucleotide. Hot colors mean high importance; cool colors mean low importance.

[Atomic Slide] The Saliency Formula Explained



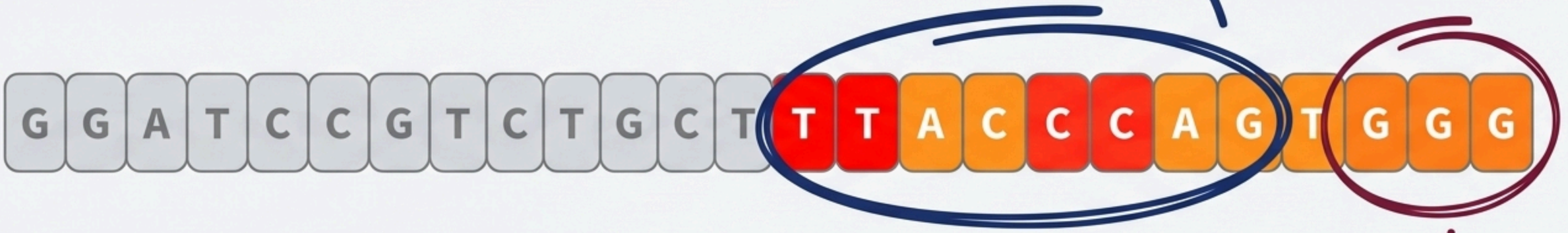
In Plain English: "The larger the gradient, the more powerfully that nucleotide influences the model's final decision."

From Heatmap to Automated Biological Report (NLG)

The system analyzes the distribution of importance scores to generate automated, human-readable insights.

Analysis of the 'Seed' Region (positions 13-20):

"The model highly values the base pairing in the Seed which is crucial for ensuring binding specificity."



Analysis of the PAM Region (positions 21-23):

"A strong PAM signal was detected, ensuring the Cas9 enzyme correctly recognizes the cutting site."

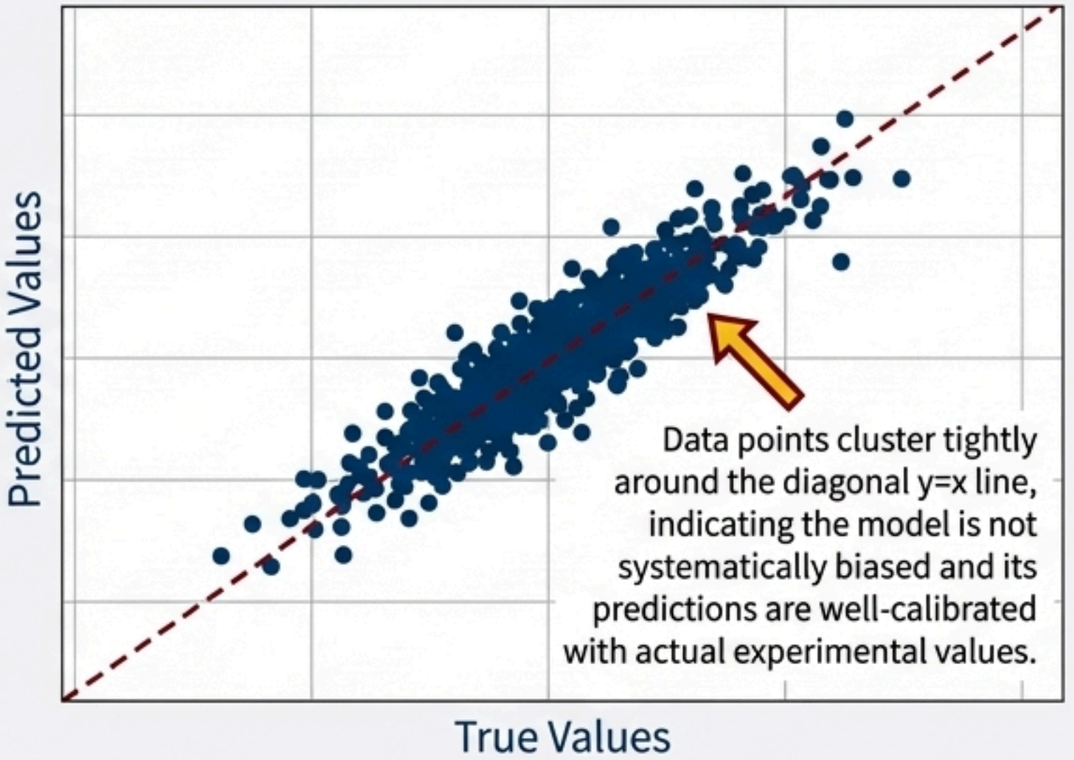
Performance Evaluation: The Numbers Speak for Themselves

The model's performance was rigorously evaluated on the unseen Test Set.

Spearman Correlation

Achieved a Spearman correlation coefficient of **ρ ≈ 0.66**. In CRISPR efficiency prediction, a coefficient > 0.6 is a strong result, indicating our model's ranking of sgRNAs closely matches real-world biological experiments.

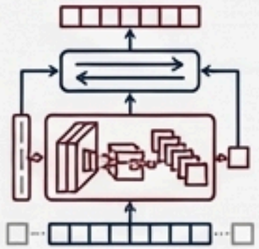
Scatter Plot Analysis



Real-World Case Study: On a real gene sequence, the system identified Top 5 cutting sites with scores > 0.8. XAI analysis confirmed these candidates had a high-importance 'Seed Region,' aligning with biological theory.

Conclusion: Key Contributions & Future Directions

Summary of Key Contributions



1. Superior Architecture:

Successfully designed and validated a Hybrid CNN-BiLSTM architecture demonstrating superior feature learning for genetic sequences.

2. Solving the Transparency Problem:

Integrated XAI to build trust, turning AI into an interpretable scientific tool.

3. Ready for Deployment:

Delivered a stable, user-friendly Streamlit application that packages the entire workflow.

Future Development Directions

Expand the Dataset

Train on more diverse datasets to enhance general accuracy.

Integrate Off-Target Prediction

Add a module to predict off-target cutting for a complete safety and efficacy profile.

Develop an API

Create an API to allow integration into larger, automated bioinformatics pipelines.