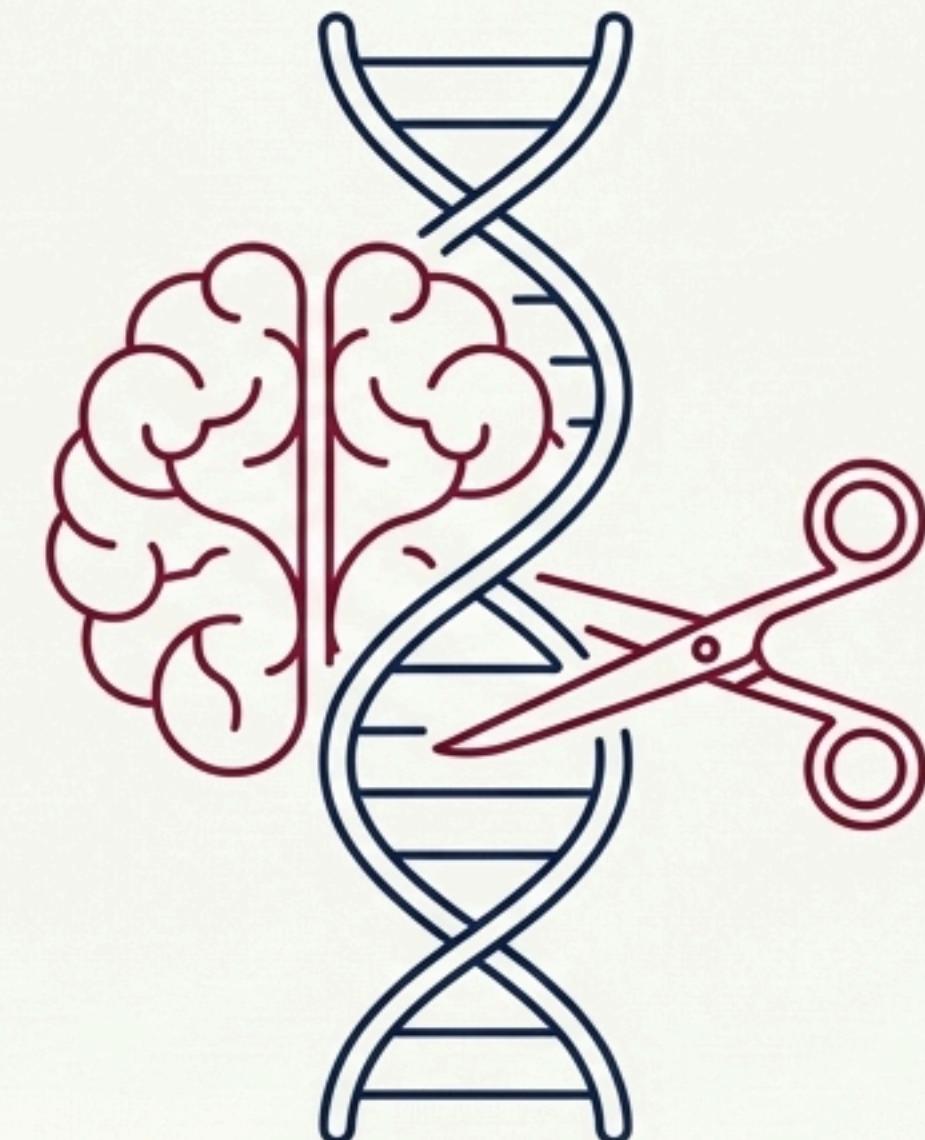


VIỆN NGHIÊN CỨU VÀ ĐÀO TẠO VIỆT - ANH, ĐẠI HỌC ĐÀ NẴNG
KHOA CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

**Ứng dụng Deep Learning (CNN-LSTM) và Explainable AI (XAI)
trong dự đoán hiệu quả cắt của hệ thống CRISPR-Cas9**



Thông tin sinh viên:

Trần Xuân Cường - 22040004
Nguyễn Văn Nhi - 22040006

Thông tin giáo viên hướng dẫn:

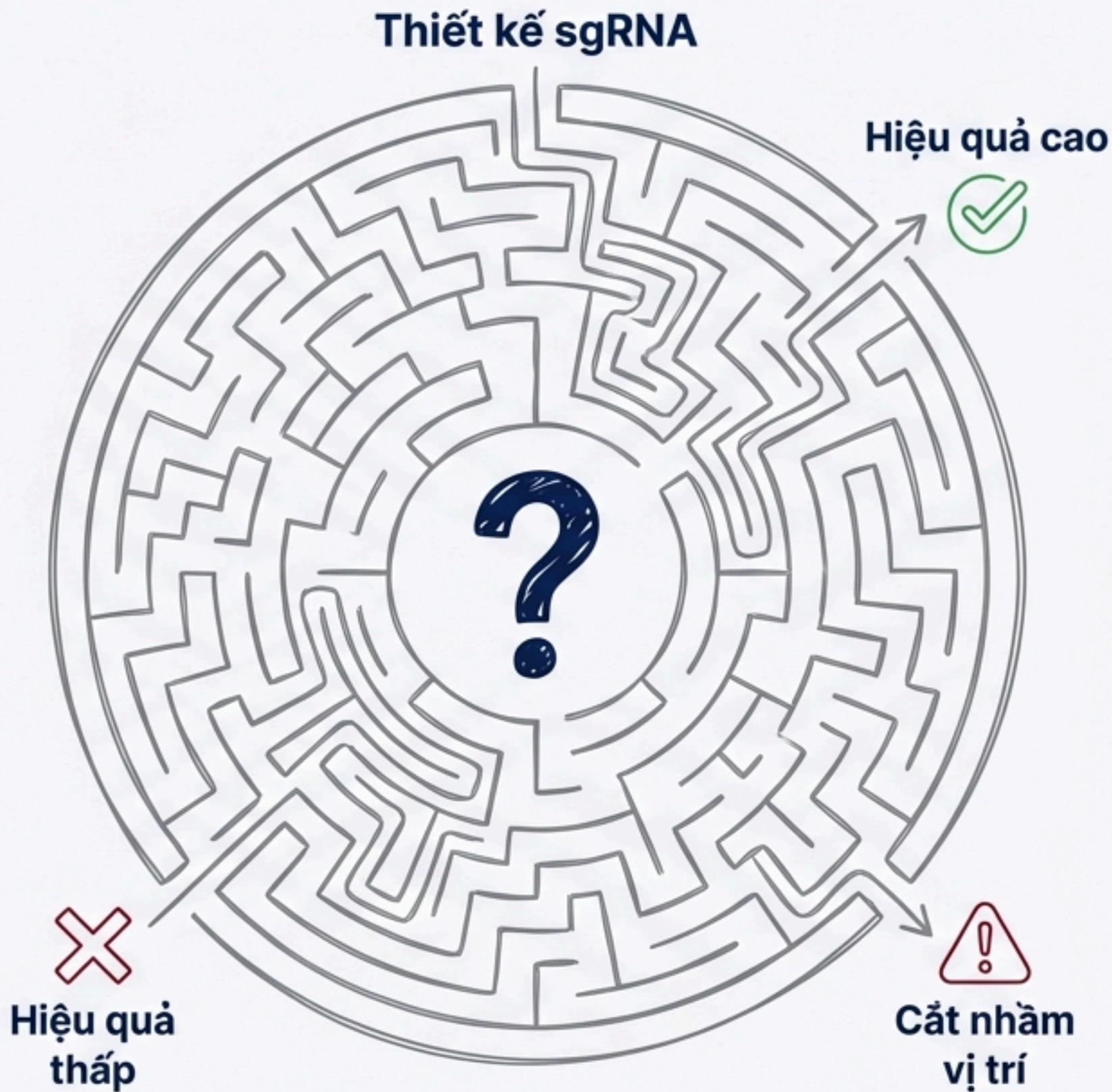
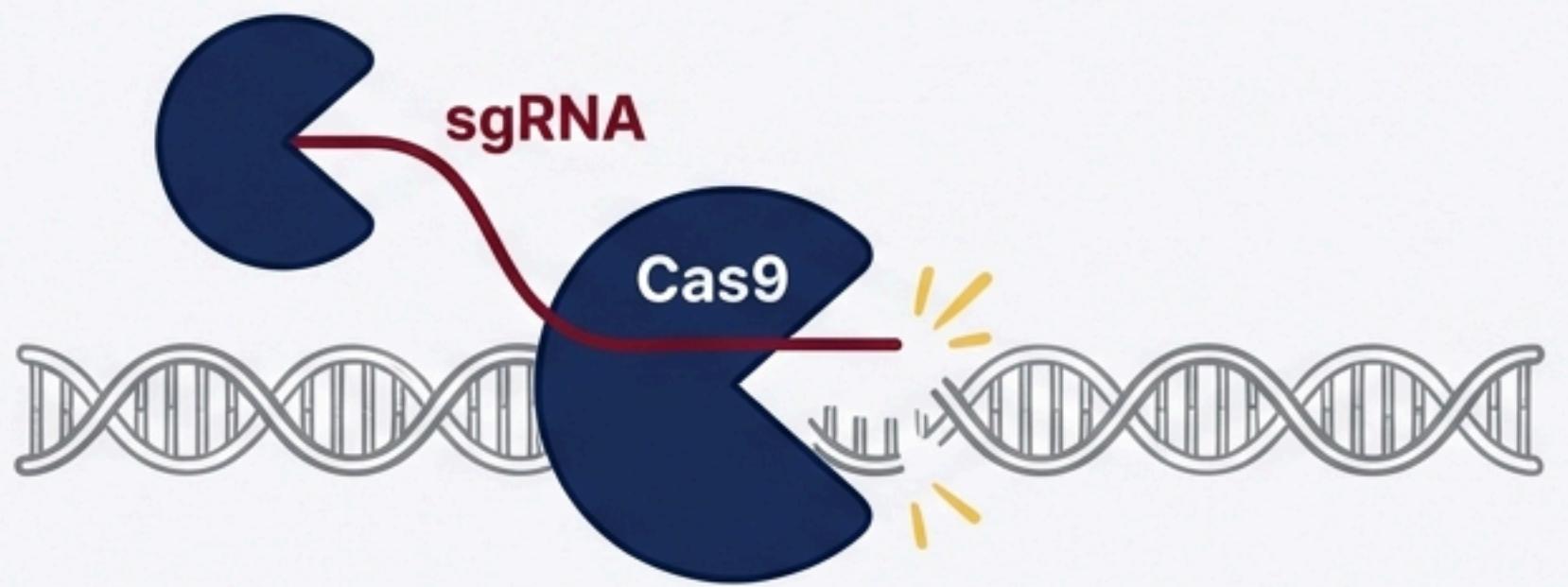
ThS. Nguyễn Chí Thiện
TS. Trần Thanh Hòa

Cuộc Cách Mạng CRISPR-Cas9 và Thách Thức Cốt Lõi

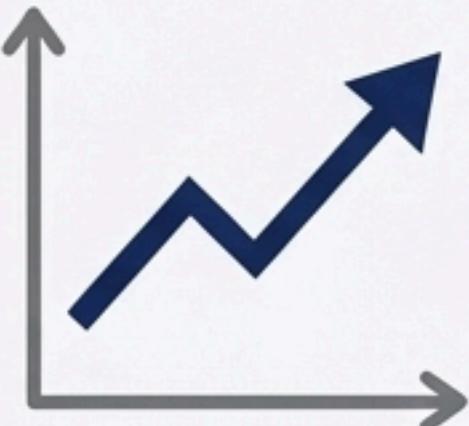
CRISPR-Cas9: Một công nghệ chỉnh sửa gen đột phá, cho phép các nhà khoa học "cắt" và "sửa" DNA với độ chính xác chưa từng có.

Thách thức lớn nhất: Hiệu quả của CRISPR phụ thuộc hoàn toàn vào việc thiết kế các đoạn RNA hướng dẫn (sgRNA).

Vấn đề: Làm thế nào để thiết kế sgRNA tối ưu, vừa tối đa hóa hiệu suất cắt tại đích (on-target efficiency), vừa giảm thiểu tác động ngoài mong muốn? Đây là bài toán cốt lõi.



Bốn Mục Tiêu Cốt Lõi Của Dự Án



Xây dựng mô hình Deep Learning hiệu suất cao

Phát triển một kiến trúc mạng nơ-ron lai ghép (Hybrid) kết hợp CNN và LSTM để dự đoán chính xác điểm hiệu quả (Efficiency Score) của sgRNA.



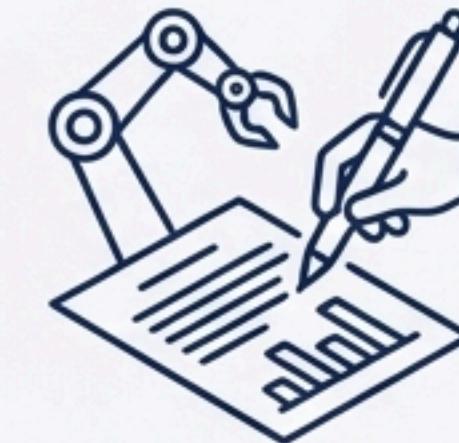
Phát triển ứng dụng thực tế (Deployment)

Đóng gói mô hình vào một ứng dụng web (Streamlit App) với giao diện trực quan, cho phép người dùng nhập liệu, quét trình tự gen và nhận báo cáo.



Tích hợp khả năng giải thích (Explainable AI - XAI)

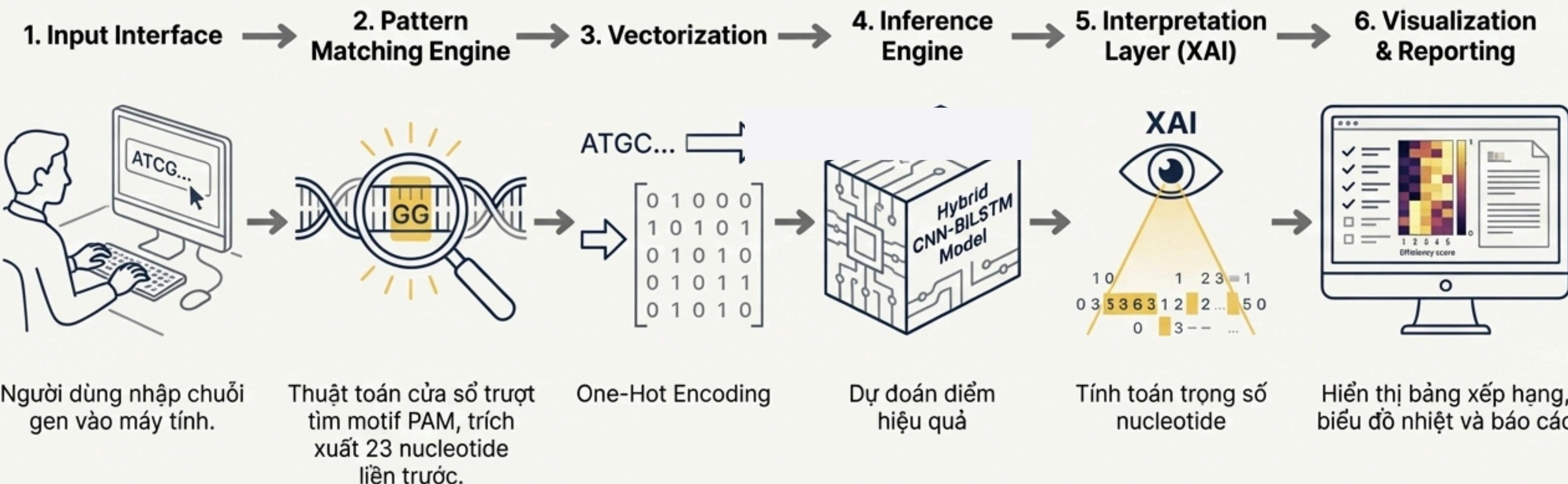
Khắc phục nhược điểm "hộp đen". Hệ thống không chỉ dự đoán mà còn chỉ rõ nucleotide nào đóng vai trò quyết định, giúp nhà nghiên cứu hiểu cơ chế sinh học đằng sau.



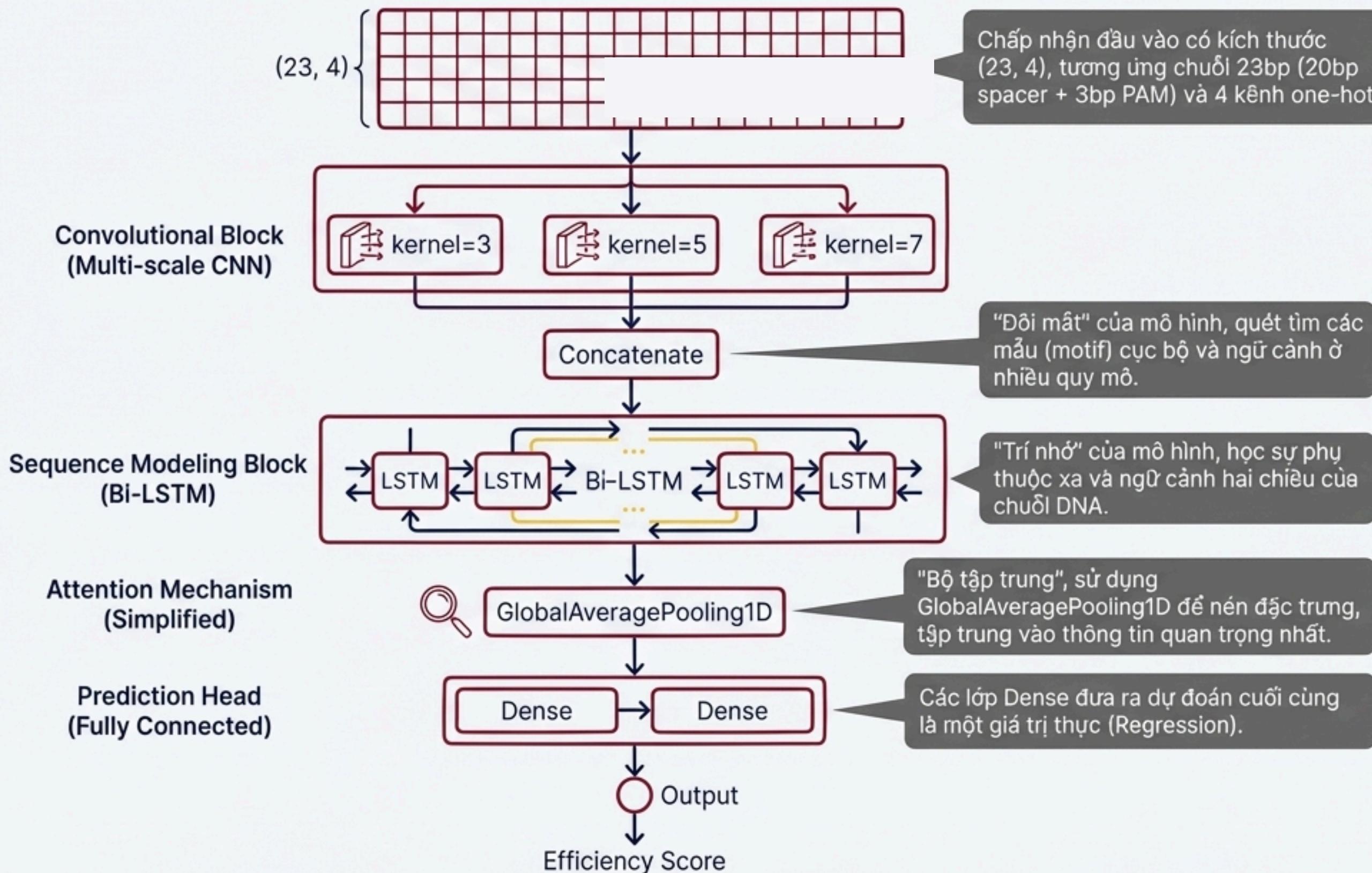
Tự động hóa phân tích (Automated Insight)

Sử dụng kỹ thuật sinh ngôn ngữ tự nhiên (NLG) để chuyển đổi các tham số kỹ thuật thành các khuyến nghị sinh học dễ hiểu.

Sơ Đồ Kiến Trúc Tổng Thể: Luồng Xử Lý Khép Kín (End-to-End Pipeline)



Trái Tim Hệ Thống: Kiến Trúc Model Hybrid CNN-BiLSTM-Attention



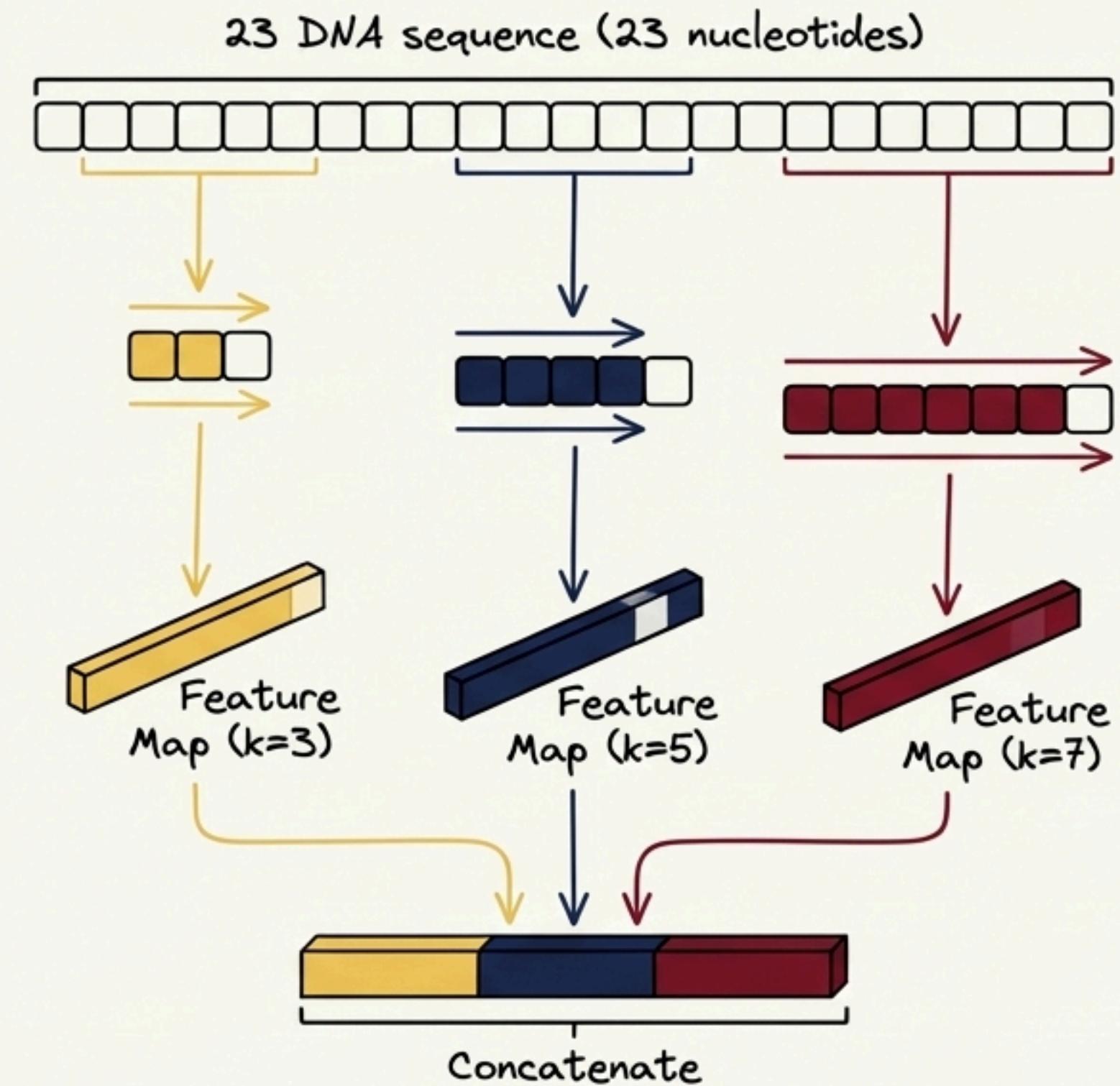
[Atomic Slide] Mô Xe Khối CNN: Học Đặc Trưng Đa Tỷ Lệ

Mục đích: CNN có thể học các "motif" sinh học (các mẫu nucleotide lặp lại có ý nghĩa).

Kiến trúc 3 nhánh song song:

- **Kernel size = 3:** Giống như một kính lúp nhỏ, phát hiện các motif rất cục bộ.
- **Kernel size = 5:** Kính lúp vừa, phát hiện các tương tác ở khoảng cách gần.
- **Kernel size = 7:** Kính lúp lớn, nhìn được ngũ cảnh rộng hơn trên chuỗi.

Gộp kết quả: Đầu ra từ 3 nhánh được gộp lại (Concatenate), cho mô hình một cái nhìn toàn diện về các đặc trưng. Mỗi nhánh đi kèm lớp BatchNormalization để ổn định quá trình học và hàm kích hoạt ReLU.



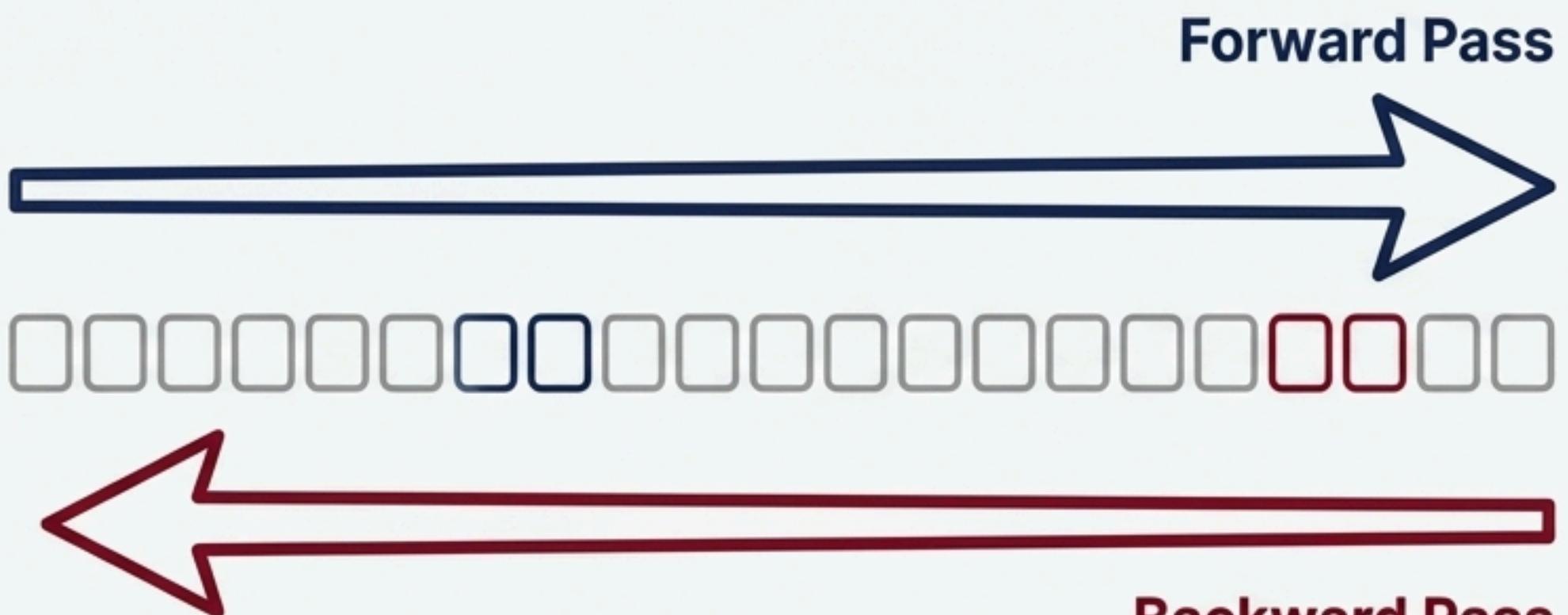
[Atomic Slide] Mở Xe Khối Bi-LSTM: Thấu Hiểu Ngữ Cảnh Hai Chiều

Vấn đề với CNN: CNN chỉ nhìn thấy các mẫu cục bộ, vì, nó không hiểu được thứ tự và sự phụ thuộc xa (ví dụ: nucleotide ở đầu ảnh hưởng đến nucleotide ở cuối chuỗi).

Vai trò của Bi-LSTM: Học sự phụ thuộc xa và ngữ cảnh hai chiều.

Sức mạnh của Bidirectional: Mô hình không chỉ "đọc" chuỗi DNA từ trái sang phải mà còn "đọc" từ phải sang trái. Điều này cực kỳ quan trọng vì các tương tác sinh học có thể xảy ra theo cả hai hướng.

Thông số: Sử dụng lớp Bidirectional LSTM với 128 đơn vị ẩn (units).



Chú thích: Hiểu được sự tương tác giữa các nucleotide ở đầu và cuối chuỗi.

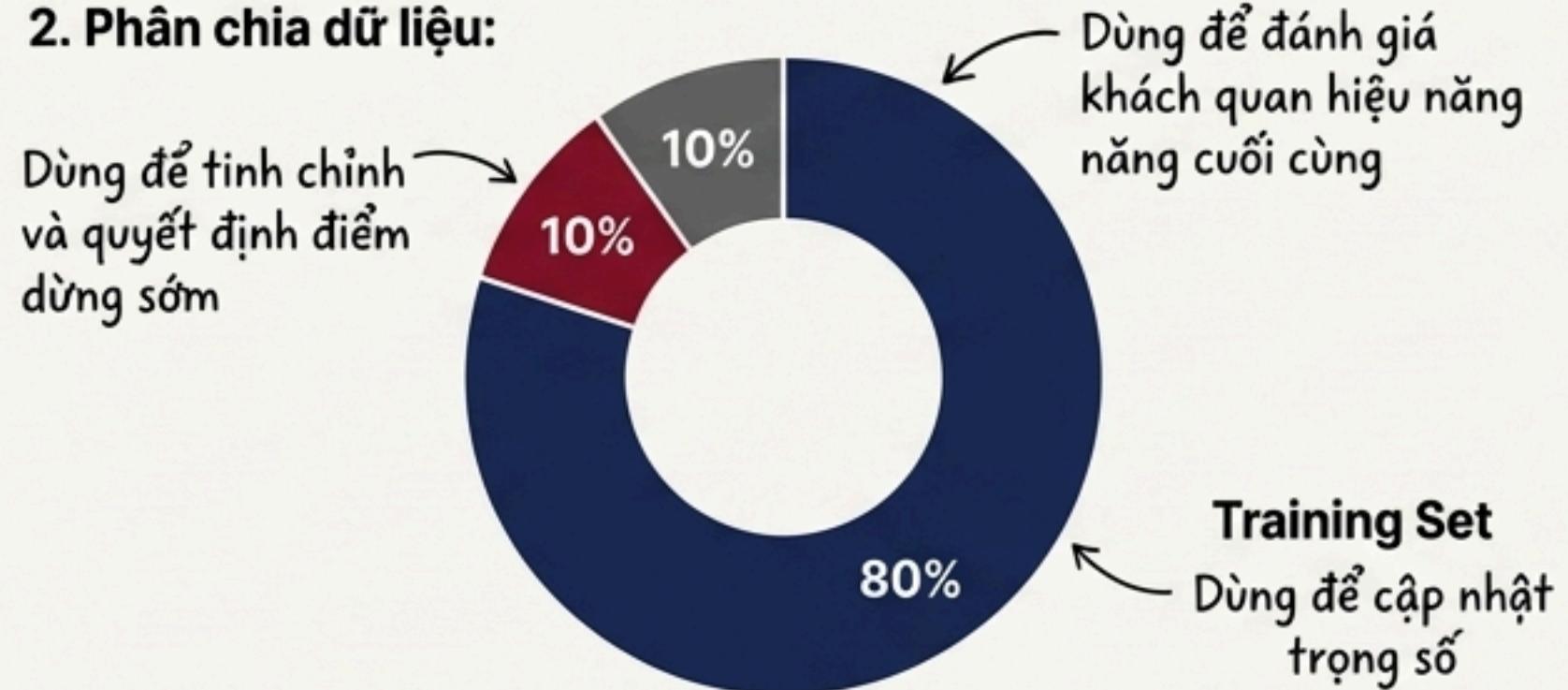
Quy Trình Huấn Luyện Nghiêm Ngặt

Data & Setup

1. Chuẩn bị dữ liệu:

- Cắt chuỗi **30bp** (đầu vào) thành chuỗi **23bp** (lấy từ index 4 đến 27).
- **Mã hóa One-Hot Encoding**: Chuyển đổi mỗi nucleotide thành vector 4 chiều.

2. Phân chia dữ liệu:



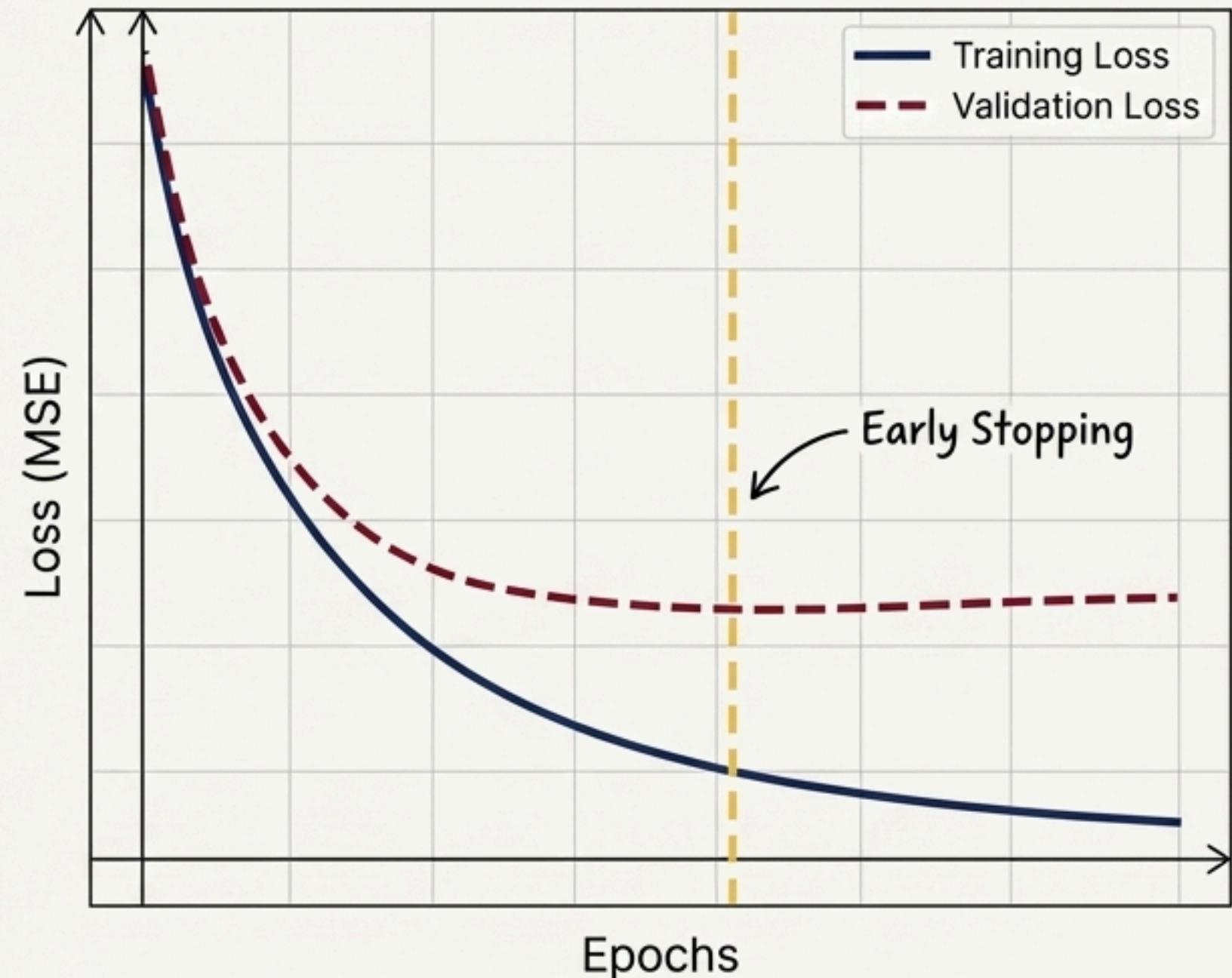
3. Thiết lập Training:

- **Optimizer**: Adam với learning rate = 0.0003.
- **Loss Function**: Mean Squared Error (MSE) cho bài toán hồi quy.

4. Chiến lược Callbacks:

- **ModelCheckpoint**: Tự động lưu lại bộ trọng số có val_loss thấp nhất.
- **EarlyStopping**: Dừng huấn luyện nếu val_loss không giảm sau 15 epochs để tránh Overfitting.

Visualization



Giao Diện Người Dùng: Sức Mạnh AI Trong Tầm Tay

Nhập một đoạn gen dài để quét và xếp hạng tất cả sgRNA tiềm năng.

Phân tích sâu một chuỗi 23bp cụ thể.

Nền tảng: Xây dựng bằng Streamlit.

Thiết kế: Giao diện 'Dark Biotech' chuyên nghiệp với Custom CSS.

Backend: Python và TensorFlow caching để tăng tốc độ phản hồi.

DNA Scanning (Xử lý hàng loạt)

ATG...TACCGTA6GGAGGGCA6CCAA
TAAGAGAGA6CCATTAAATACAGGAC
CCCC6ATC6SE6AC6CGEETTC6AECAC
TEA6AAAGA6GT66AC6SETAA6TTAA
TCCC6BAAA6TEETCATTAGATAT6AT
CAC...TAC

Single Sequence Analysis (Phân tích đơn lẻ)

GGCAA...TTAAGG

Predict

Predicted sgRNA Rankings

sgRNA Sequence	Predicted Score	Rank
GGCAA...TTAAGG	0.95	1
AATTCC...66ECAA	0.88	2
CC66AA...TTAACCC	0.75	3
AATTCC...6GGCCC	0.90	4
CC66AA...TTAACCC	0.85	5
AATTCC...TTAGGG	0.75	6
AATTCC...6GECAA	0.75	7

XAI Visualization: Single Sequence Importance

GGCAA...TTAAGGG

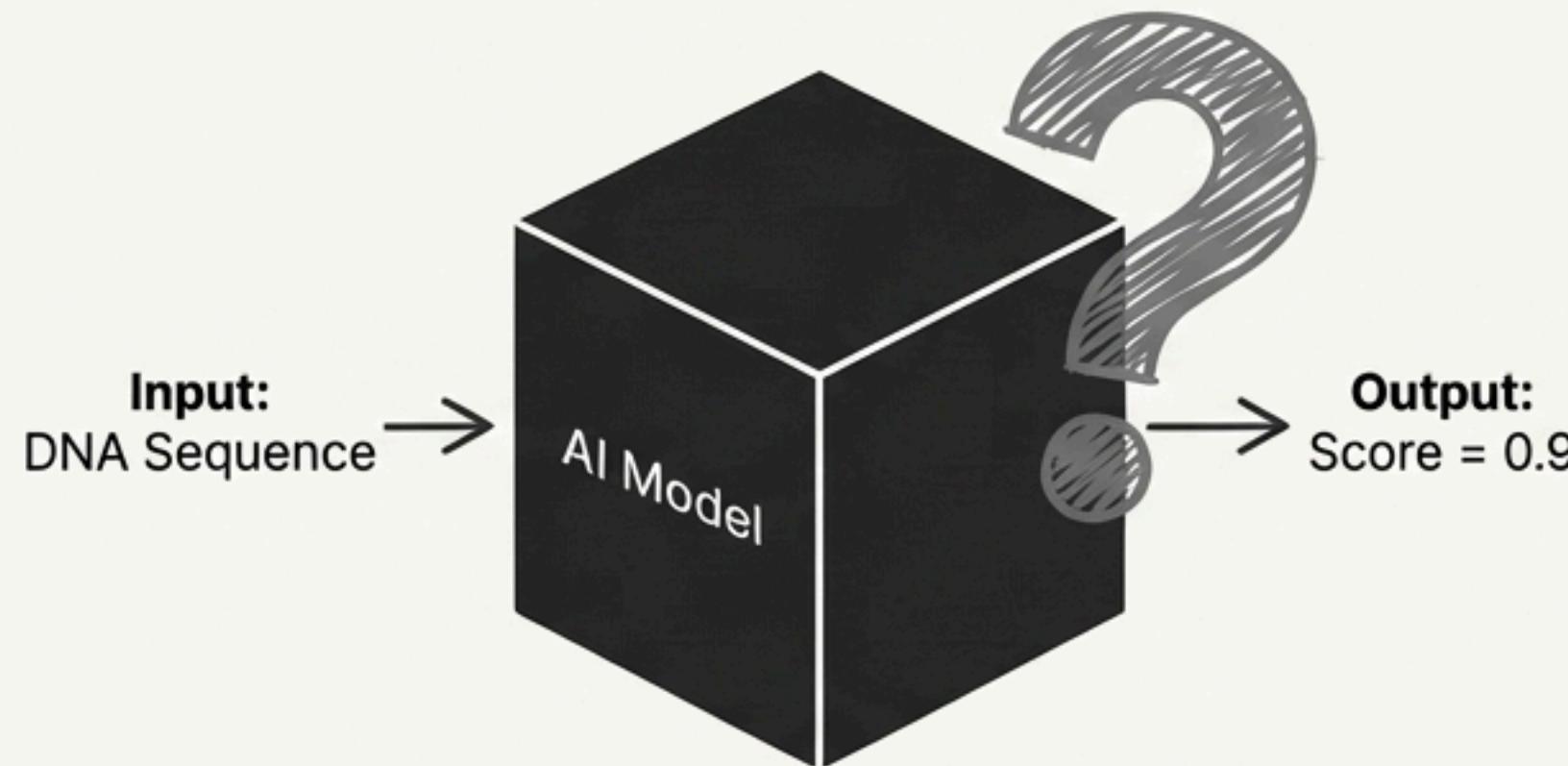
Bảng kết quả được xếp hạng theo điểm hiệu quả dự đoán.

Trực quan hóa XAI: Các nucleotide quan trọng nhất được tô màu đậm.

"Mở Hộp Đen": Explainable AI Hoạt Động Như Thế Nào?

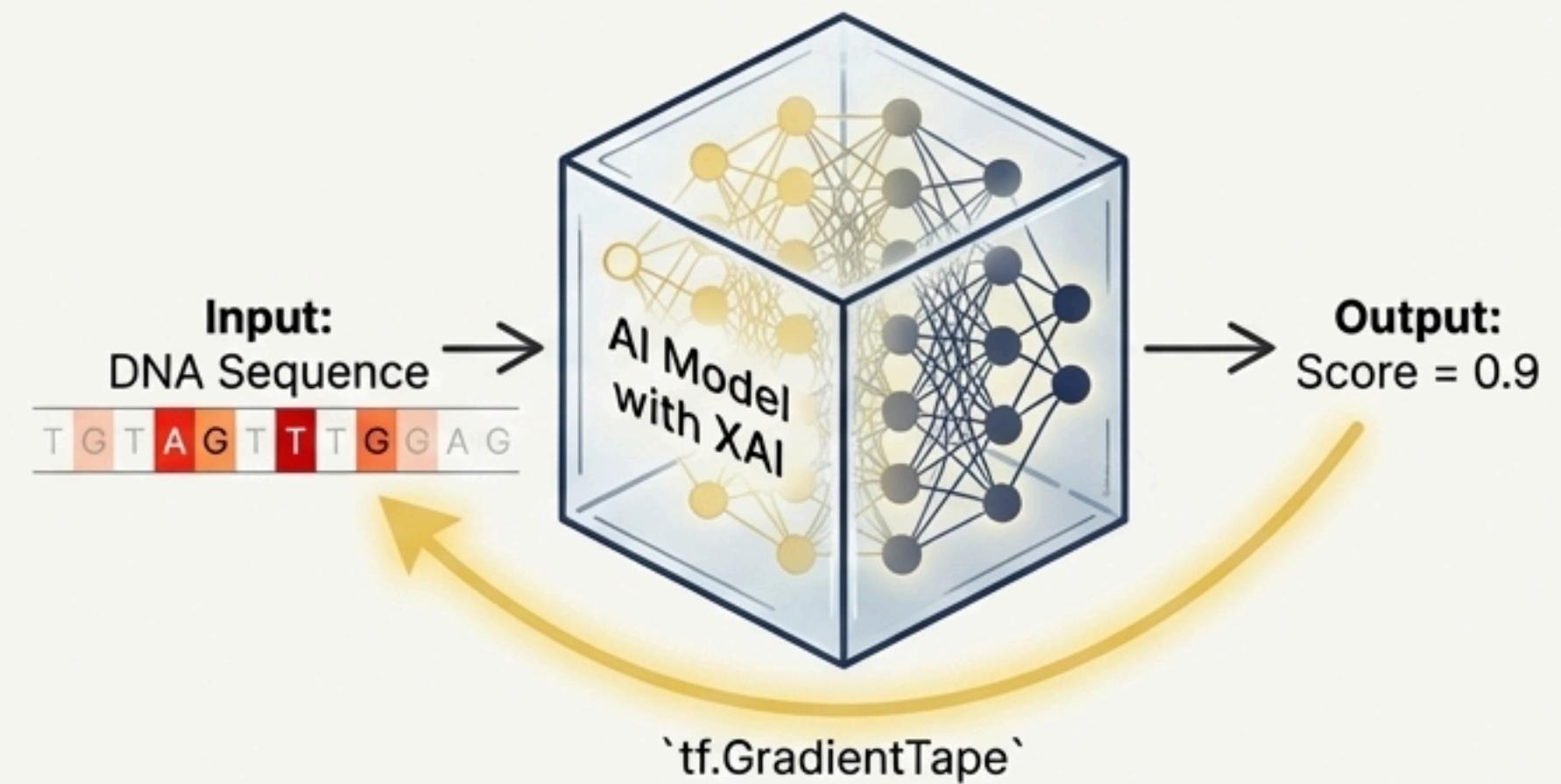
Trước

The Black Box Problem



Sau

The Transparent Solution

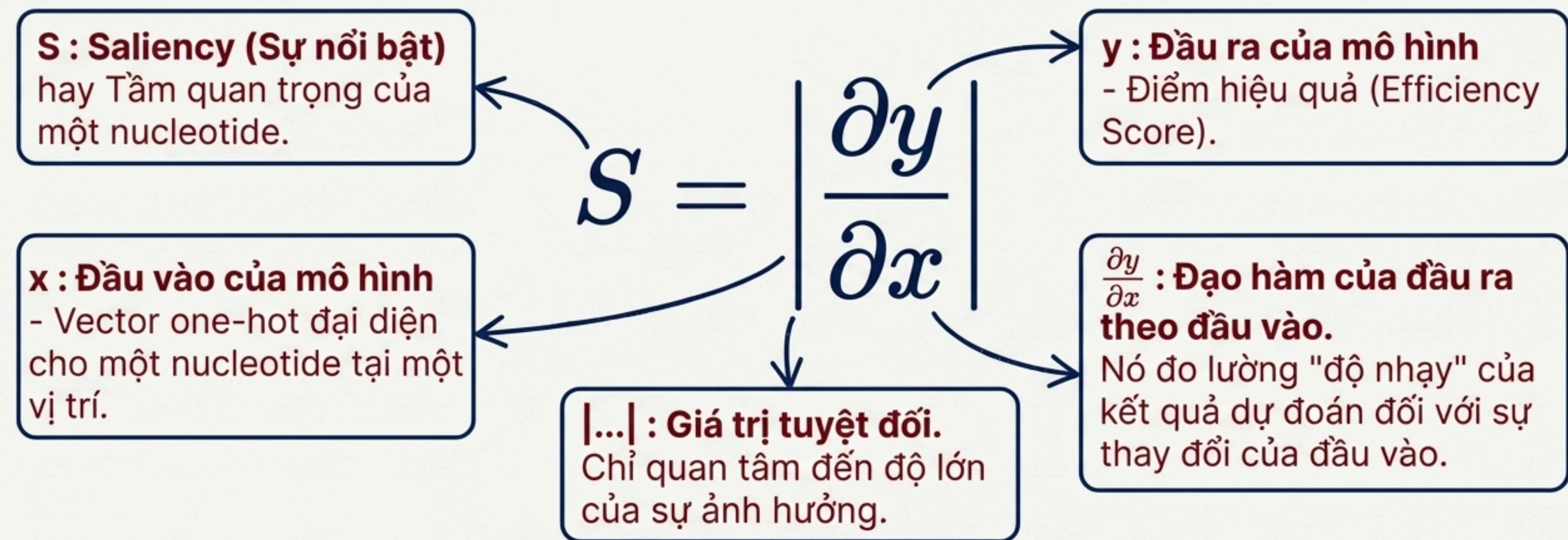


Vấn đề: Các mô hình Deep Learning thường là "hộp đen". Chúng ta không biết tại sao chúng lại đưa ra một dự đoán cụ thể.

Giải pháp: Sử dụng Saliency Maps dựa trên Gradient. Ta tính toán xem 'Sự thay đổi nhỏ tại một nucleotide đầu vào sẽ làm thay đổi kết quả dự đoán nhiều như thế nào?'

Công thức: $S = \left| \frac{\partial y}{\partial x} \right|$

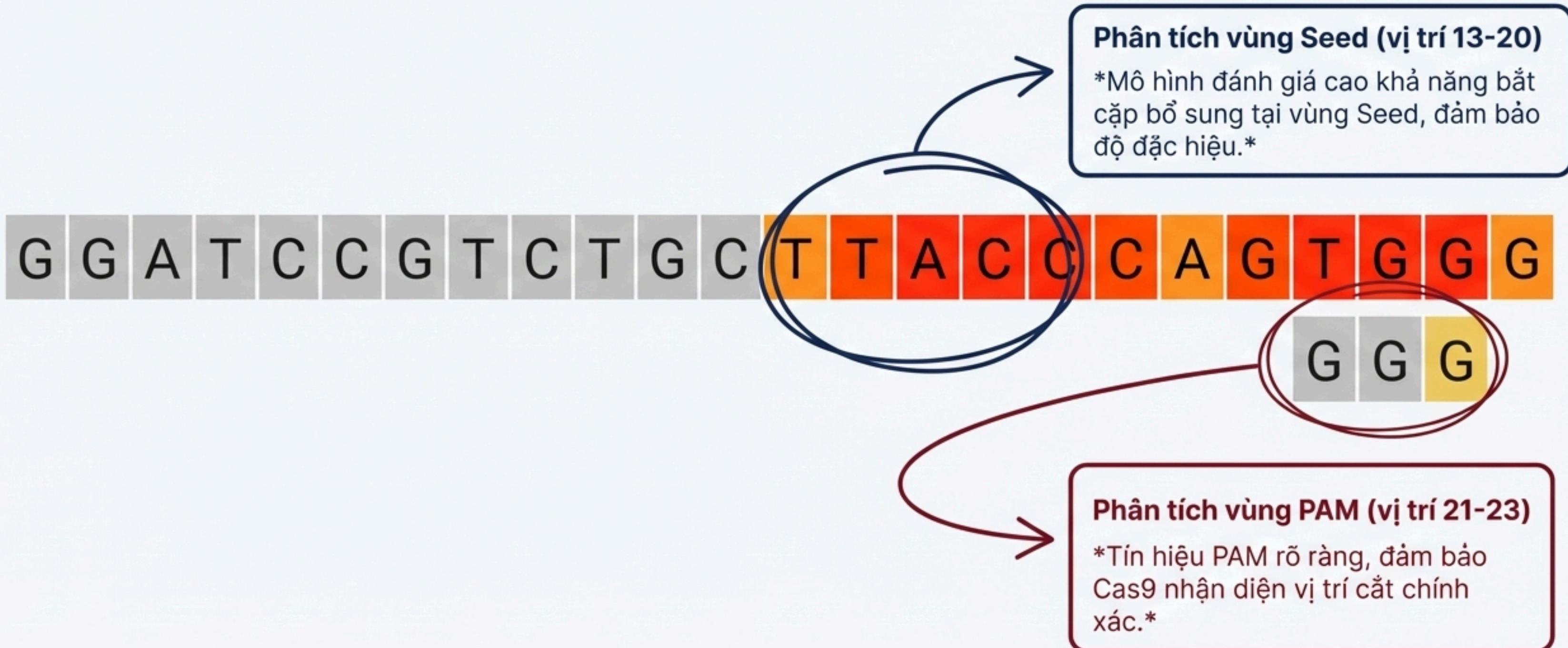
[Atomic Slide] Giải Thích Công Thức Saliency



Diễn giải bằng lời: 'Đạo hàm càng lớn, chứng tỏ nucleotide đó có ảnh hưởng càng mạnh mẽ đến quyết định cuối cùng của mô hình.'

Từ Bản Đồ Nhiệt Đến Báo Cáo Sinh Học Tự Động (NLG)

Hệ thống phân tích sự phân bố trọng số để tạo ra nhận xét tự động.

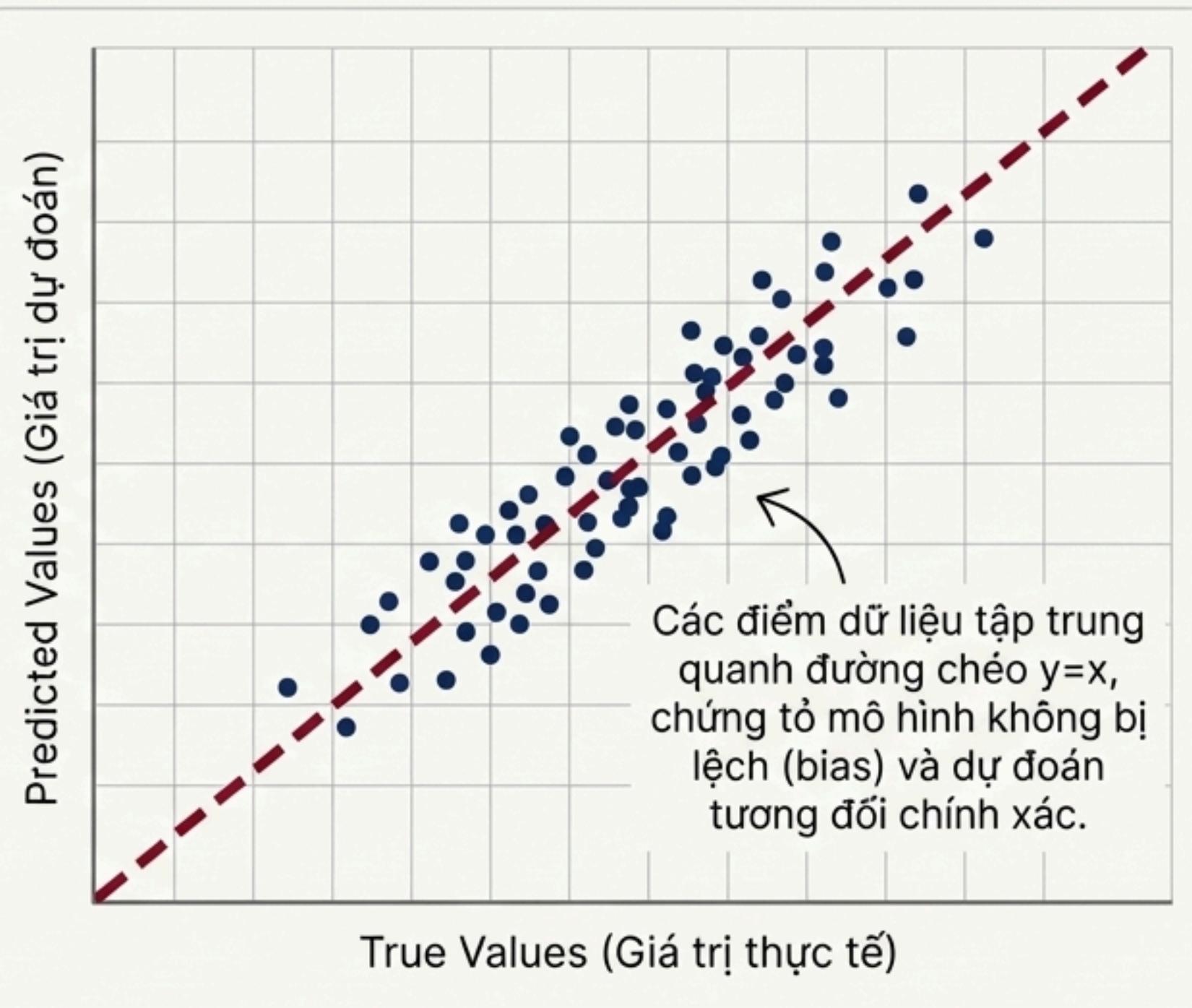


Đánh Giá Hiệu Năng: Những Con Số Biết Nói



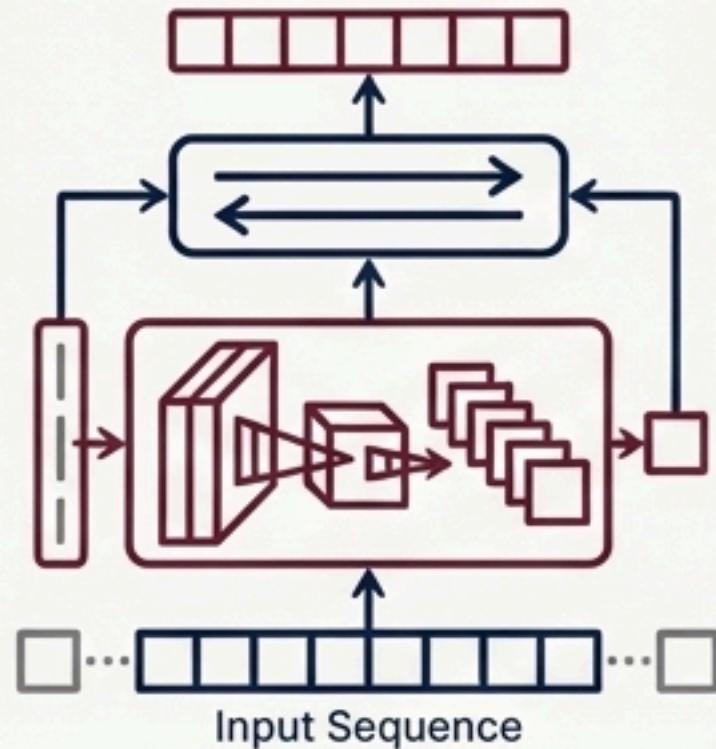
Spearman Correlation

Trong lĩnh vực này, chỉ số $\rho > 0.6$ được coi là là kết quả tốt. Điều này cho thấy mô hình xếp hạng các sgRNA rất sát với thực nghiệm sinh học.



Case Study: Thủ nghiệm trên gen đích thực tế, hệ thống đã lọc ra được Top 5 vị trí cắt có điểm > 0.8 , và phân tích XAI cho thấy vùng Seed Region rất "sáng", phù hợp lý thuyết.

Tổng Kết Các Đóng Góp Chính

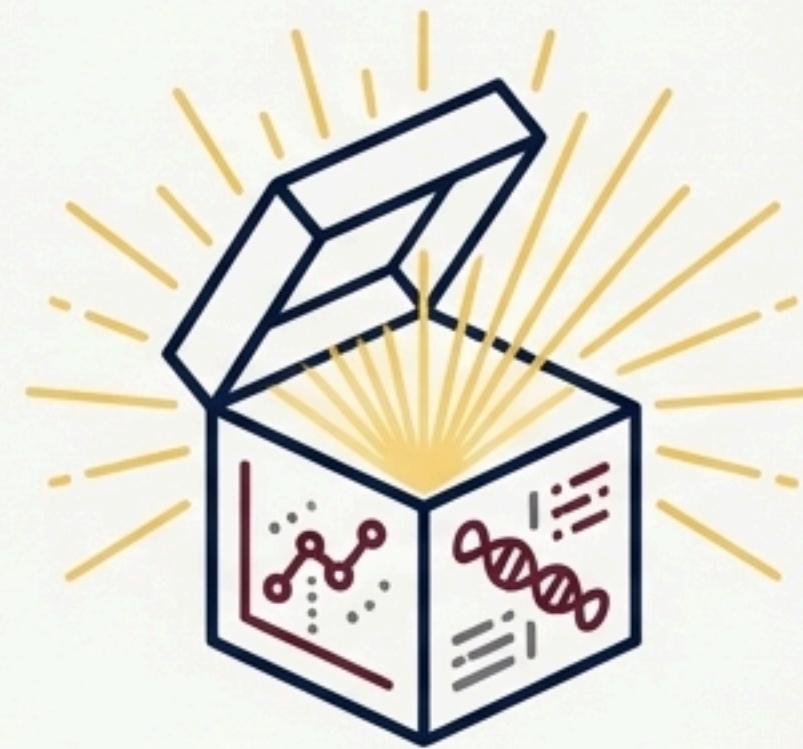


1. Kiến trúc Vượt trội

Đề xuất thành công kiến trúc mạng Hybrid CNN-BiLSTM, chứng minh khả năng học đặc trưng chuỗi gen vượt trội so với các phương pháp truyền thống.

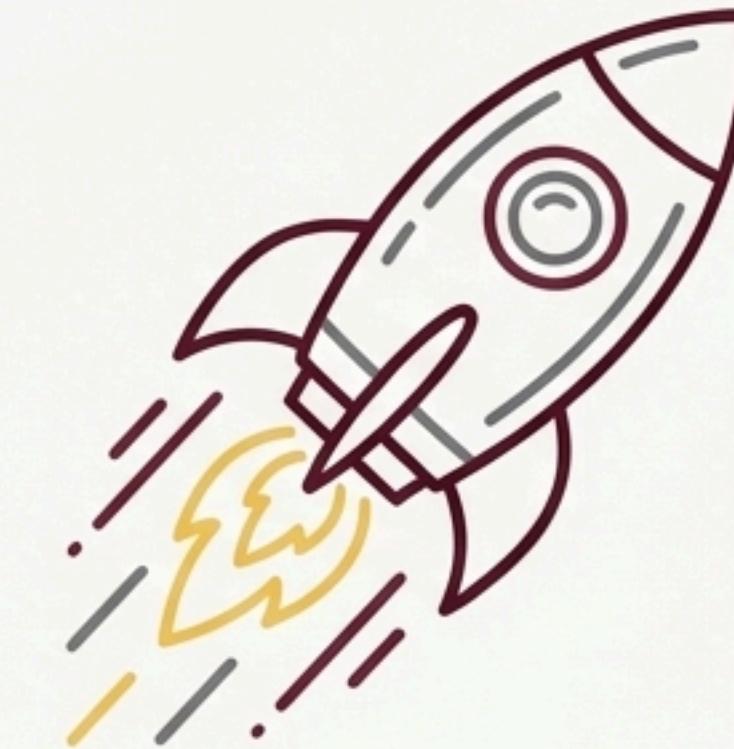
2. Giải quyết vấn đề Minh bạch

Tích hợp XAI (Saliency Maps), giúp xây dựng niềm tin cho người dùng là các nhà sinh học, biến AI thành công cụ có thể diễn giải được.



3. Sẵn sàng Triển khai

Xây dựng ứng dụng Streamlit hoạt động ổn định, giao diện thân thiện, tích hợp đầy đủ quy trình từ nhập liệu đến báo cáo.



Hướng Phát Triển Trong Tương Lai

