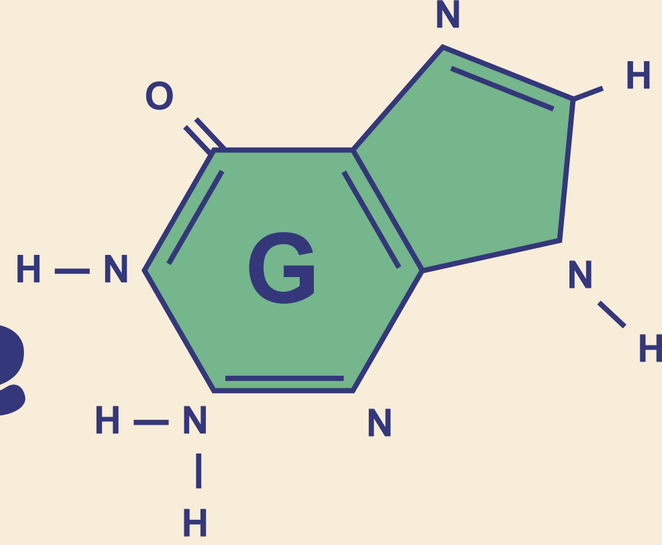
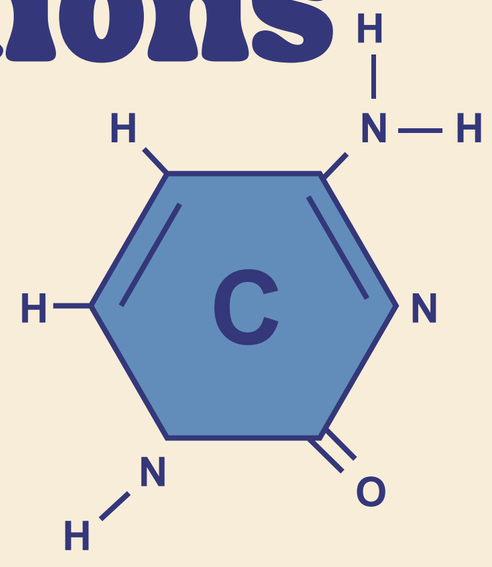
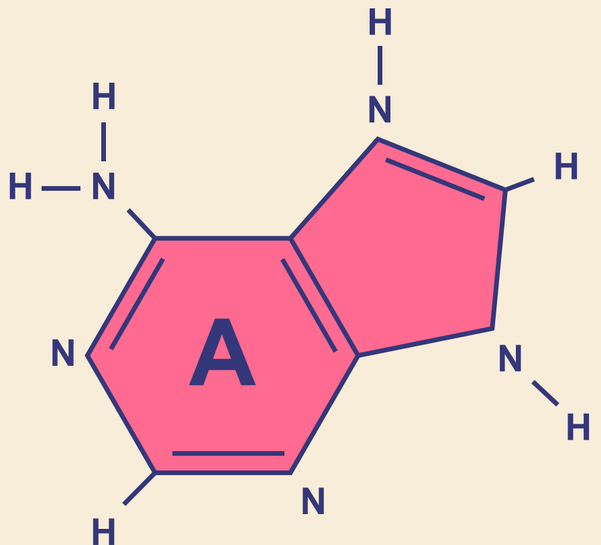
Optimizing Sequence Alignment with Machine Learning for Improved Protein-Compound Interactions

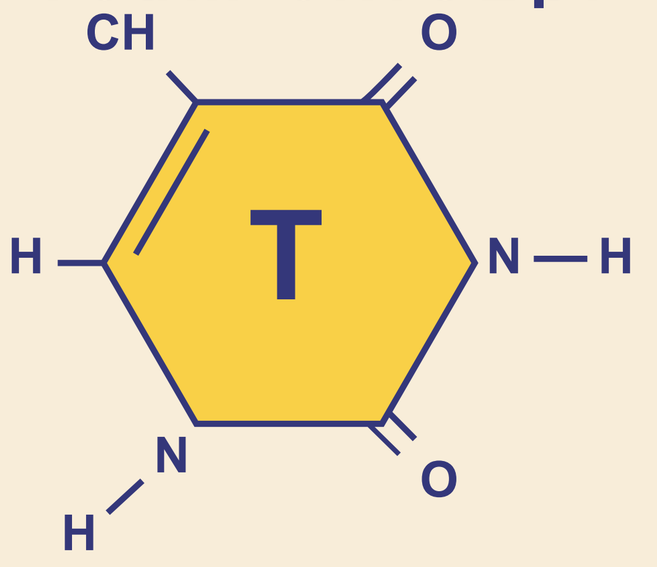


Pharmahack 2025 / Team cRICEpr

Mingee Jung
Han Lee
Ze Yu Huang
Brandon Lee Felix
Kevin Xu







Multiple Sequence Alignment for Antimicrobial Resistance

AMR occurs when microorganisms become resistant to drugs.

Drug development is expensive and time-consuming.

MSA helps identify resistance patterns in microbial genomes.

Our Approach

- Sequence alignment optimization is crucial for tasks like antimicrobial resistance detection.
- Challenges: Structural similarity among sequences during primary screening.
- Approach: Used Optimized Gearbox Score with LSTM and RL models.
- Improved alignment with higher scores using Reinforcement Learning and sequence modeling.
- Result: Decreased testing cost and faster screening.

Inspiration

Content:

- Problem: Traditional sequence alignment methods are time-consuming and costly.
- Goal: Speed up the alignment process in drug discovery and biological research.
- Impact: More efficient pipelines for biological studies and better drug-target interactions.
- Vision: Accelerate medical advancements and improve quality of life using optimized computational models.

What it Does / How We Built It

What is does?

- Predicts the optimal sequence alignment moves to maximize alignment scores.
- Focus on protein-compound interactions using the Optimized Gearbox Score.
- Features:Sequence position / Gap density / Mutation rate
- Goal: Accurate sequence alignment predictions to minimize experimental testing.

How we built it?

- Data: Used bioassay data from PubChem for training.
- Model: Combined LSTM-based sequence prediction with Reinforcement Learning (RL).
- Scoring: Optimized Gearbox Score to measure sequence alignments.
- Outcome: Model predicts the best alignment moves and provides a higher final score.

Overcoming Challenges

Feature Selection:

Difficult to identify which features (mutation rate, gap density) would significantly improve predictions.

Model Overfitting:

Ensuring the model doesn't overfit to small sample sizes was a challenge.

What's Next for Challenge

Feature Expansion:

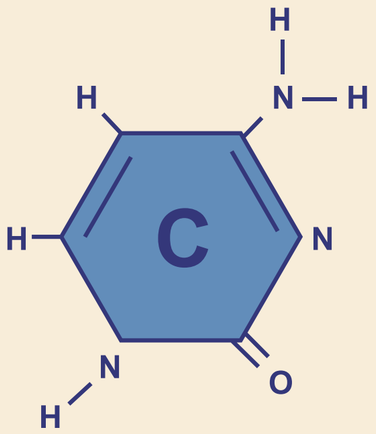
Use evolutionary motifs and substitution matrices to improve model performance.

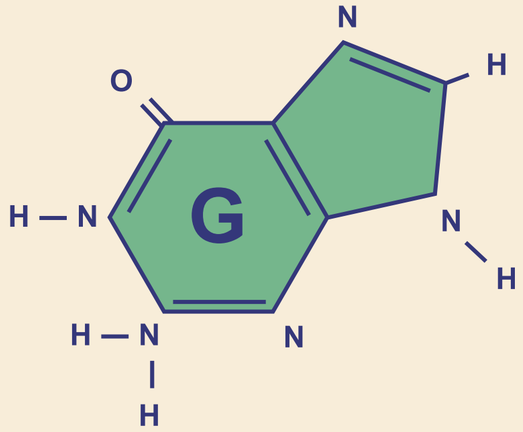
Model Tuning:

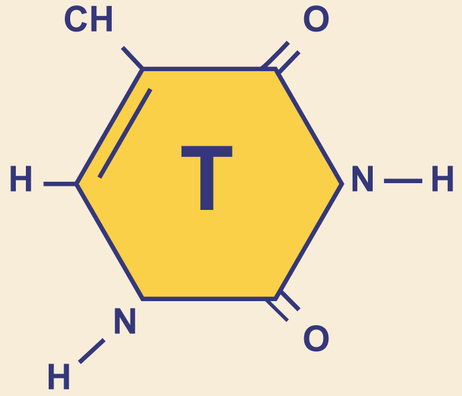
Explore genetic algorithms and reinforcement learning improvements for better convergence.

Scalability:

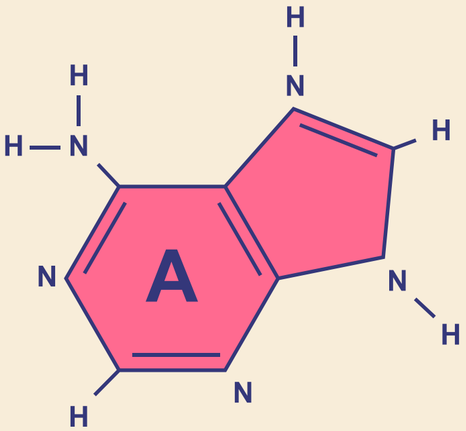
Test on larger real-world datasets to optimize model for production use.







Thanks!



Do you have any questions?