Drug-Disease Association Prediction for Drug Repositioning

GROUP 5:
陳冠仲、張啟廣、吳欣瑜、李知祐

MOTIVATION :

Why Drug Repositioning?

- Developing a new drug costs 1–2 billion USD
- Drug development often takes more than 10 years
- Repositioning finds new uses for existing drugs
- Much faster, cheaper, and safer

Classic Example: Minoxidil (Rogaine)

- Original Use: Treated Hypertension (高血壓)
- Side Effect: Doctors observed unexpected hair growth
- New Use: Repositioned to treat Androgenic Alopecia (雄性禿)

Goal of this project:

- Build predictive models (ML + DNN) to identify potential drug–disease associations

Problem Definition

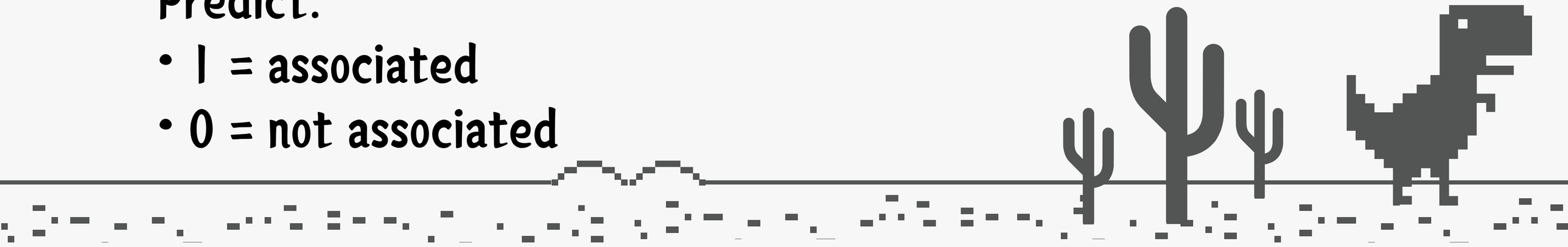
We formulate drug repositioning as a binary classification problem:

Given:

- A drug (chemical fingerprint)
- A disease (ID representation)

Predict:

- 1 = associated
- 0 = not associated



DATASET OVERVIEW

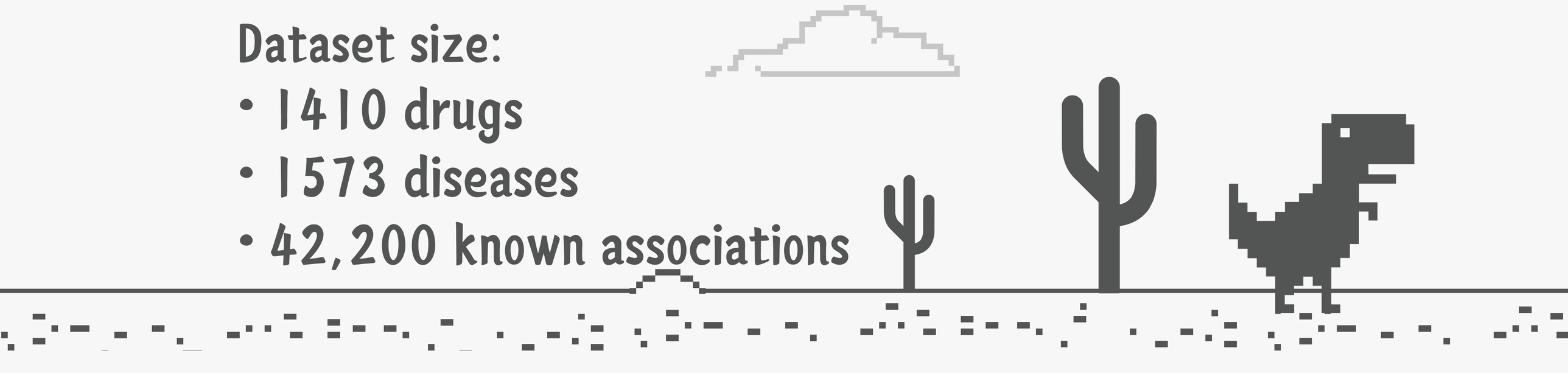
We use a Kaggle dataset containing:

Files & Descriptions:

- drugsInfo.csv: Drug Chemical Information (SMILES, targets, categories)
- diseasesInfo.csv: Disease Metadata and IDs
- mapping.csv: Known drug–disease associations (positive pairs)

Dataset size:

- 1410 drugs
- 1573 diseases
- 42,200 known associations



DATASET CONSTRUCTION

Positive Samples:

- All pairs present in mapping.csv (Label = 1)

Negative Sampling Strategy:

- Problem: Dataset only contains positive links.
- Solution: Randomly sample 'unknown' drug-disease pairs as negatives.
- Ratio: 1:1 Balanced Sampling.
- Ensures the model doesn't just predict '1' for everything.

Result: A balanced, fully supervised training set.

FEATURE ENGINEERING OVERVIEW

1. Drug Features

- Morgan Fingerprints (1024 bits):
  - Encodes local chemical substructures from SMILES.
- Target Multi-Hot Encoding:
  - Captures biological protein targets.

2. Disease Features

- One-hot Encoding (for ML Baselines)
- Learned Embedding Layer (for DNN)

Final Vector: Concatenation of Drug & Disease features.

MACHINE LEARNING MODELS

We compared three models:

1. Logistic Regression (Baseline)

- Strong for high-dimensional sparse features

2. Random Forest

- Captures nonlinear interactions

3. XGBoost

- Gradient boosting on decision trees

Evaluation metric:

- ROC-AUC (most important), Accuracy, F1-score

METHODOLOGY: THE MM-DNN

The Limitation of Traditional ML:

- Relies heavily on Morgan Fingerprints (摩根指紋).
- Only looks at the Chemical Structure (化學結構).

Our Deep Learning Approach (Multi-Modal):

- Drug Targets (標靶蛋白): We parsed biological targets (e.g., proteins the drug interacts with).
  This adds Pharmacological Mechanism (藥理機制) info.
- Drug Categories: High-level functional classification.
- Disease Embeddings Layer: Instead of simple IDs, we learn dense vectors for diseases to capture Semantic Similarity (語意相似度).

PROPOSED METHOD: MM-DNN

Multi-Modal Deep Neural Network Architecture:

1. Drug Tower (Dense Layers)
- Processes 1024-bit Fingerprints + Targets.
- Layers: Dense(256) → BN → ReLU → Dropout.

2. Disease Tower (Embedding)
- Learns a dense vector representation for each Disease ID.

3. Fusion & Prediction
- Concatenates towers → Dense Layers → Sigmoid Output.
- Trained with Binary Cross Entropy Loss.

MODEL PERFORMANCE SUMMARY

Model	AUC	Accuracy	F1-score
Logistic Regression	0.8396	0.7669	0.7688
Random Forest	0.8273	0.7534	0.7176
XGBoost	0.735	0.7031	0.6111
MM-DNN (Ours)	**0.8508**	**0.7739**	**0.7784**

Conclusion:

Our MM-DNN model outperforms all baselines, achieving the highest AUC (0.8508) and F1-score,
proving the effectiveness of learned embeddings.

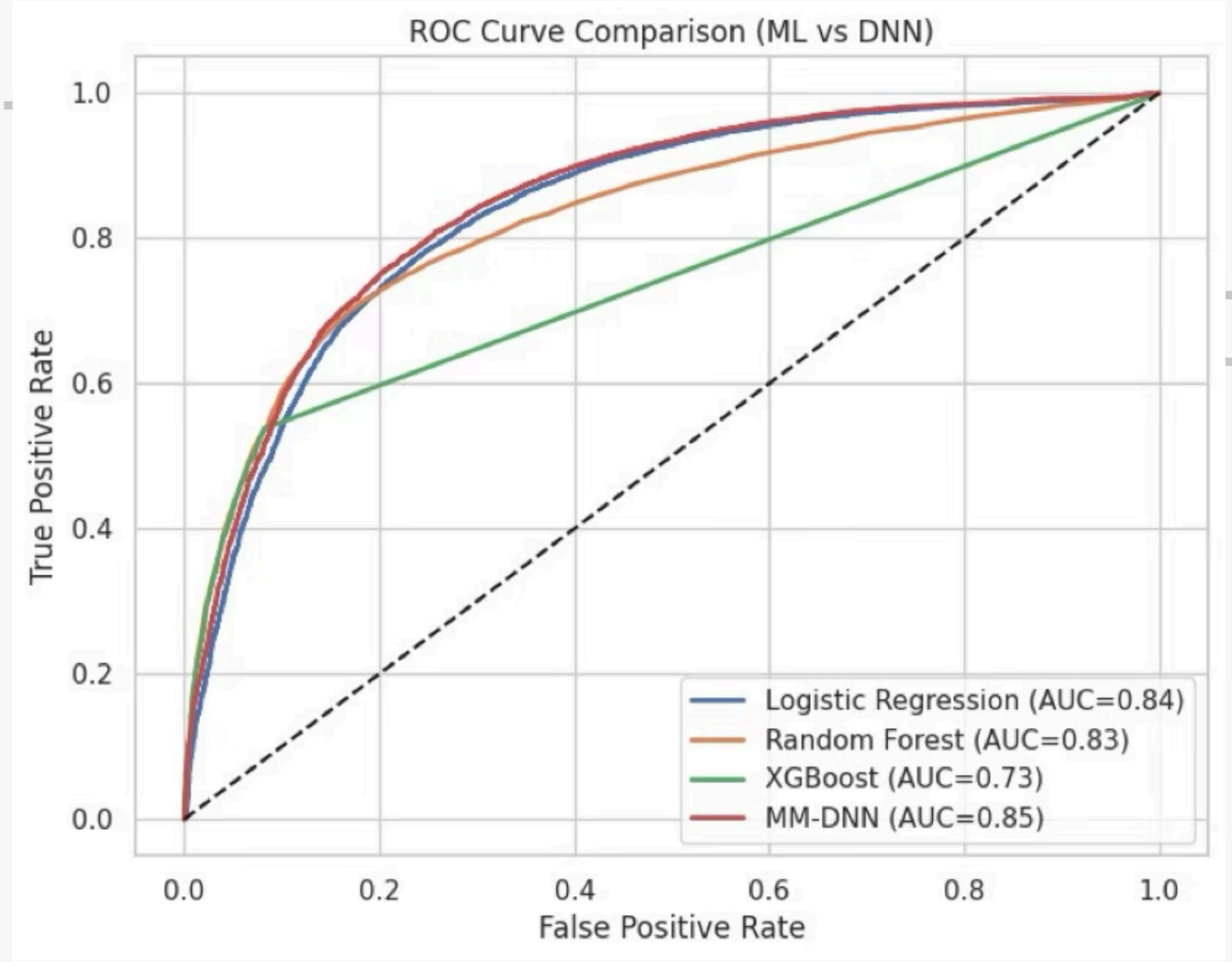
ROC CURVE COMPARISON

Performance Visualization:

- Red Line (MM-DNN):
  Shows the largest area under curve (AUC=0.85).

- Comparison:
  DNN > Logistic Regression > Random Forest > XGBoost.

- Insight:
  The smooth curve of DNN indicates stable ranking capabilities.



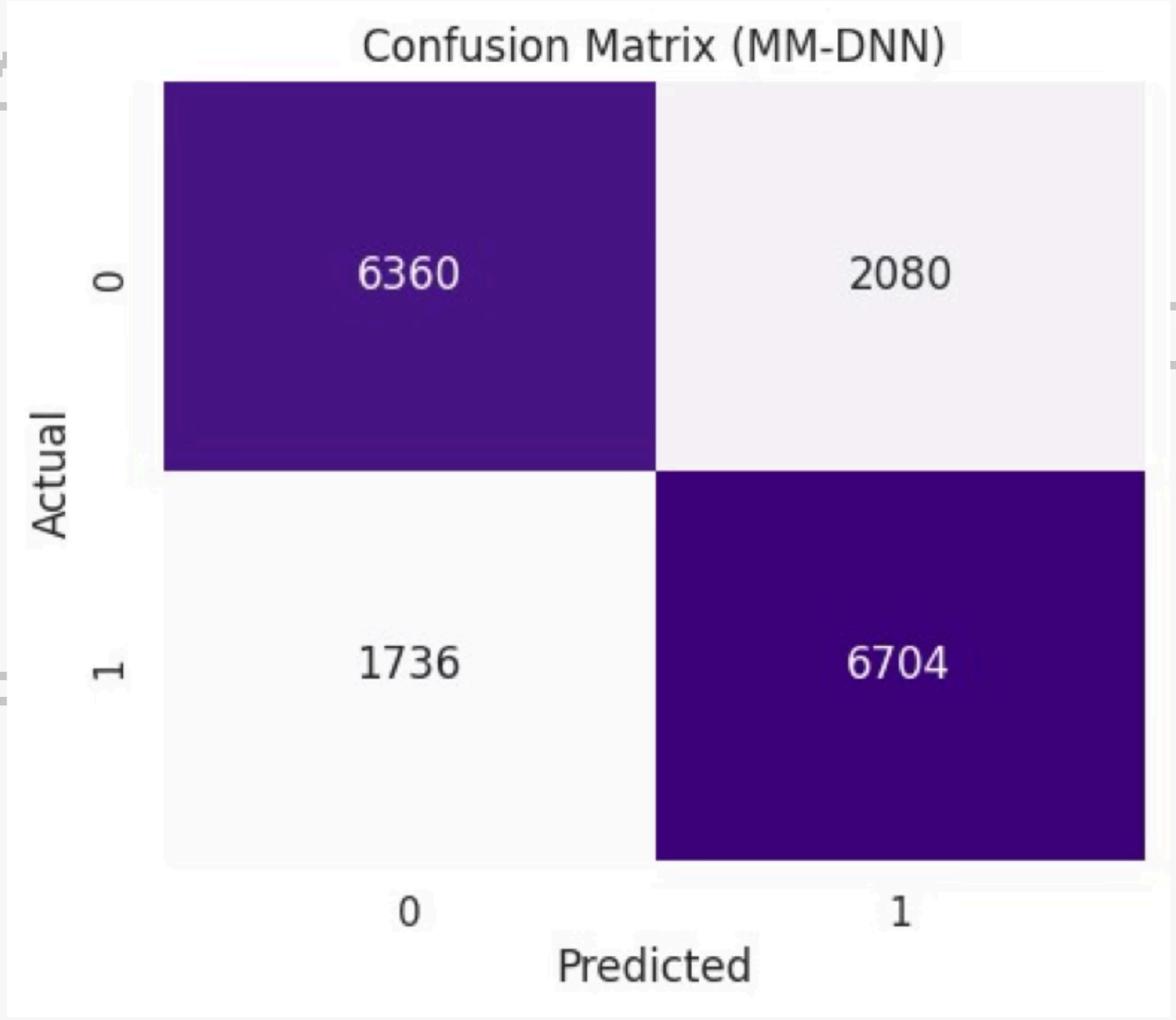
CONFUSION MATRIX (MM-DNN)

Detailed Error Analysis:

- True Positives (6704):
  Correctly identified valid drug-disease pairs.

- True Negatives (6360):
  Correctly rejected invalid pairs.

- Balance:
  The model shows balanced performance between sensitivity and specificity, avoiding bias towards one class.



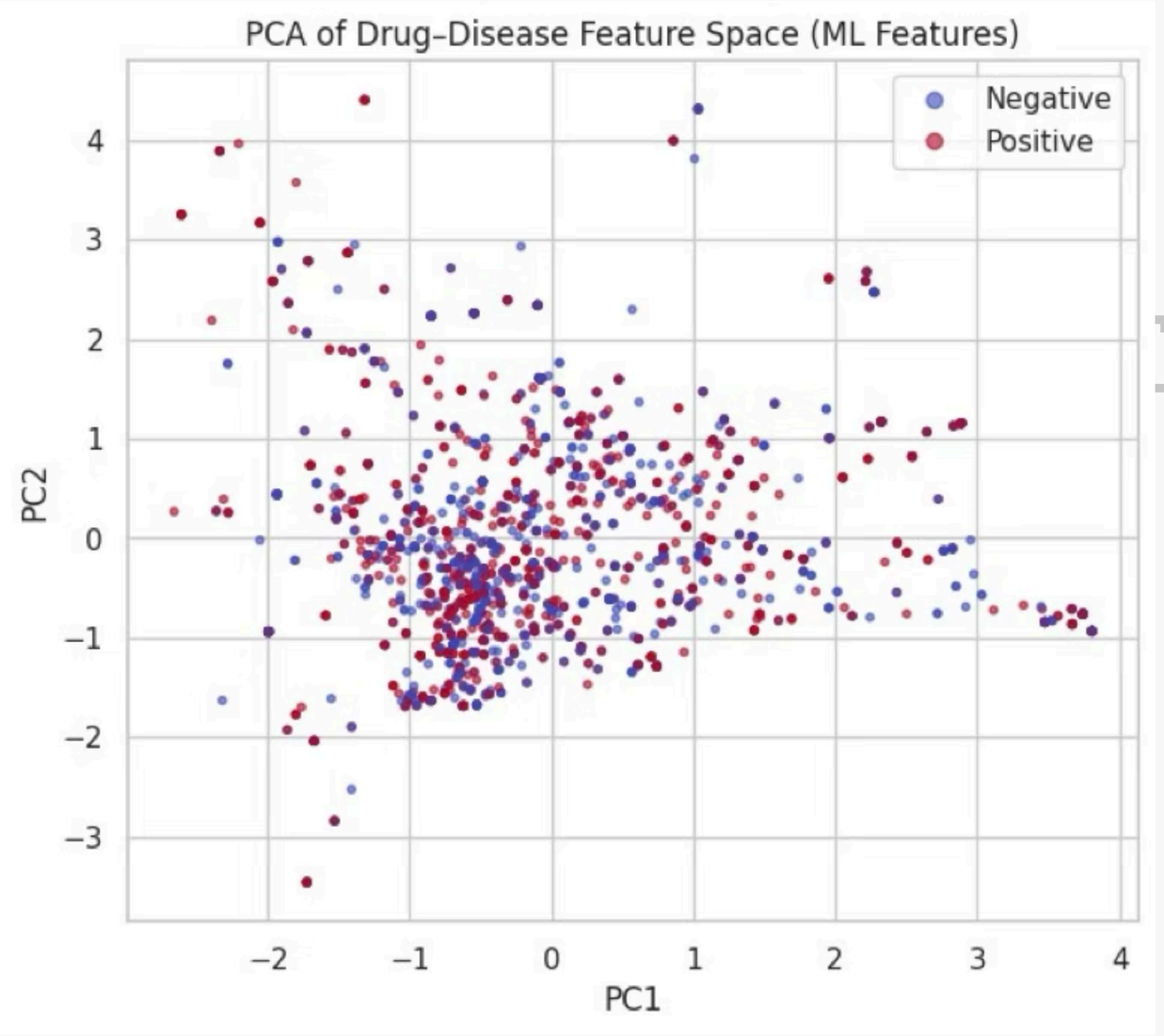
PCA Feature Space Visualization

Data Complexity:

- High Dimensionality:
  Reduced 1000+ features to 2 principal components.

- Observation:
  Positive (Red) and Negative (Blue) classes overlap significantly.

- Why DNN wins:
  Linear models struggle with this overlap, but DNNs can learn non-linear boundaries to separate them.



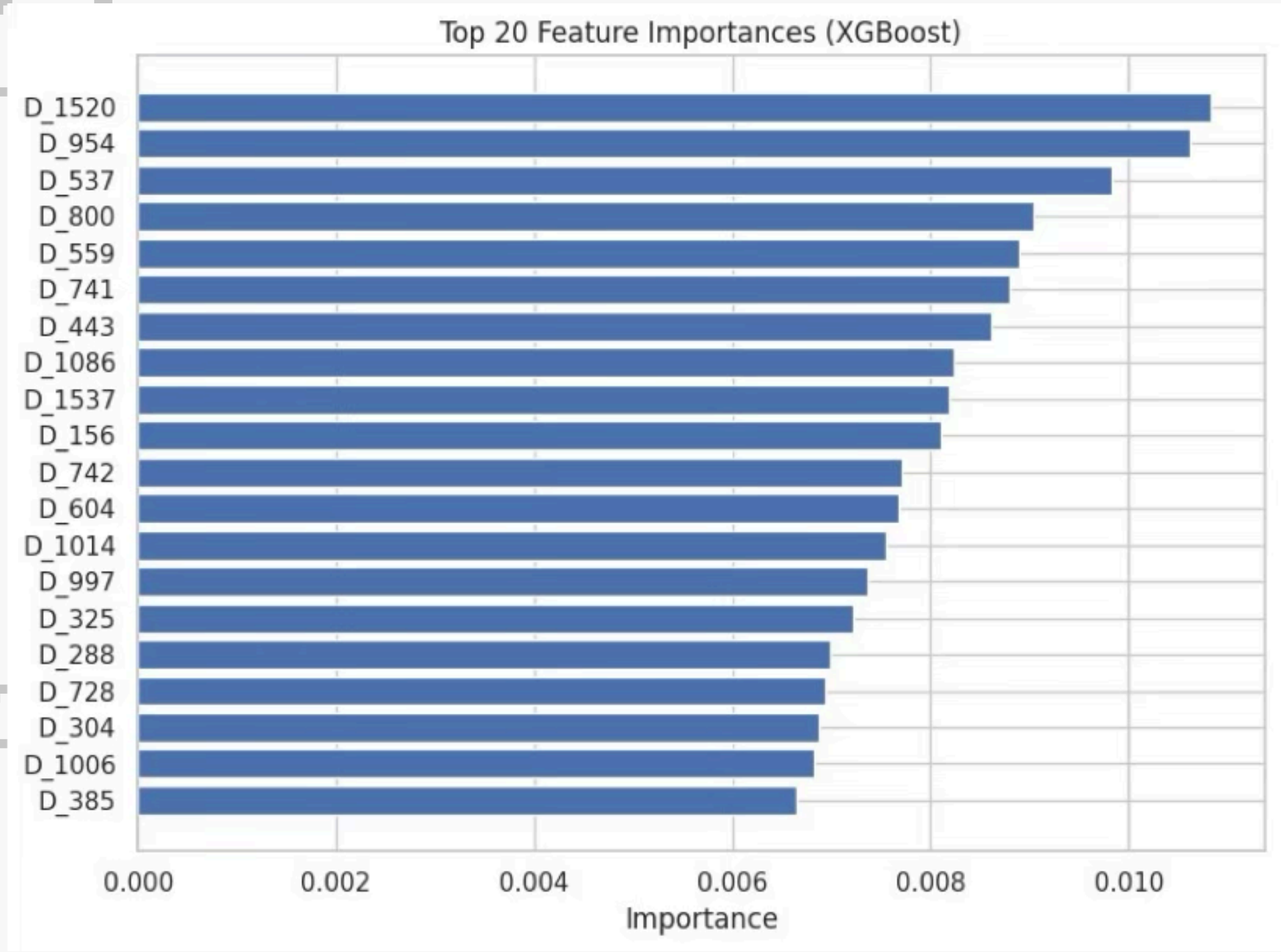
Feature Importance (XGBoost)

Model Interpretability:

- Although MM-DNN performs best, XGBoost provides interpretability.

- Key Features:
  - D_xxx features dominate the top 20.
  - No FP_xxx features appear → tree-based models rely on disease identity.

- Validation:
  - XGBoost captures disease-driven signals.
  - MM-DNN can additionally learn chemical–biological interactions.



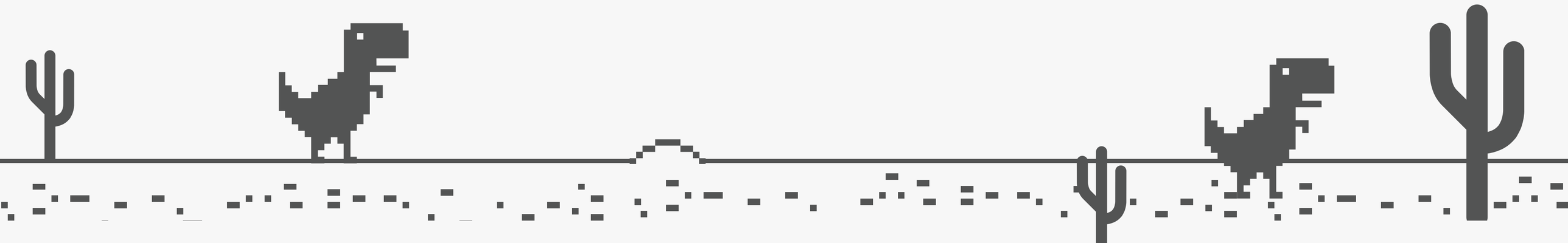
CRITICAL ANALYSIS: THE VALUE OF DEEP LEARNING

- Why is DL (0.8537) only slightly better than LR (0.8459)?
  - Strong Linear Features:
    - Morgan Fingerprints are highly explicit linear features. Simple rules (e.g., "Contains Benzene Ring -> Effective") work surprisingly well for Logistic Regression.
  - The Value of Deep Learning:
    - Generalization : LR relies on "memorizing" specific bits. DL "understands" similarity via Embeddings.
- Conclusion:

We chose DL not just for the score, but for its robustness And inference capability.

CONCLUSION

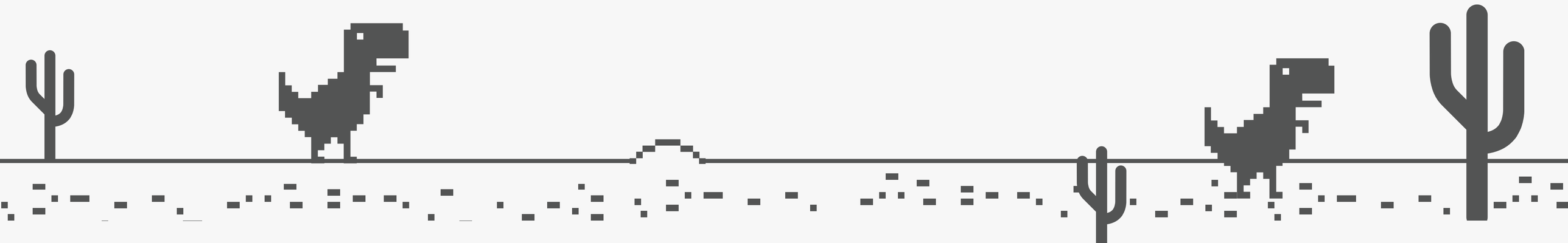
- Deep Learning (MM-DNN) successfully predicts drug repositioning candidates.
- It outperforms traditional ML baselines (AUC 0.85 vs 0.84).
- Feature Engineering (Morgan FP + Embeddings) was critical



FUTURE WORK

Future Work:

- Graph Neural Networks (GNN): Explicitly model the graph structure.
- External Data: Integrate Gene Expression profiles (L1000).
- Cold Start Problem: Test on completely new drugs not seen in training.



Q&A

Between traditional ML models and DNN, which one do you expect to perform best on high-dimensional sparse drug data?

PARTICIPATION AND PEER EVALUATION

陳冠仲:Code, Report Preparation
張啟廣:Code, Report Preparation
吳欣瑜:Oral Presentation, Report Revision
李知祐:Oral Presentation, Report Revision

THANK YOU