CHEMINFORMATICS-BASED DRUG REPURPOSING USING RANDOM FOREST CLASSIFIER AND GAN MODEL

Hannah Archer
CPSC 5440
Prof. Liang
April 20, 2025

What is drug repurposing?

- Reusing approved drugs for new diseases
- Faster, cheaper than traditional drug development
- Reduces time from discovery to clinical use

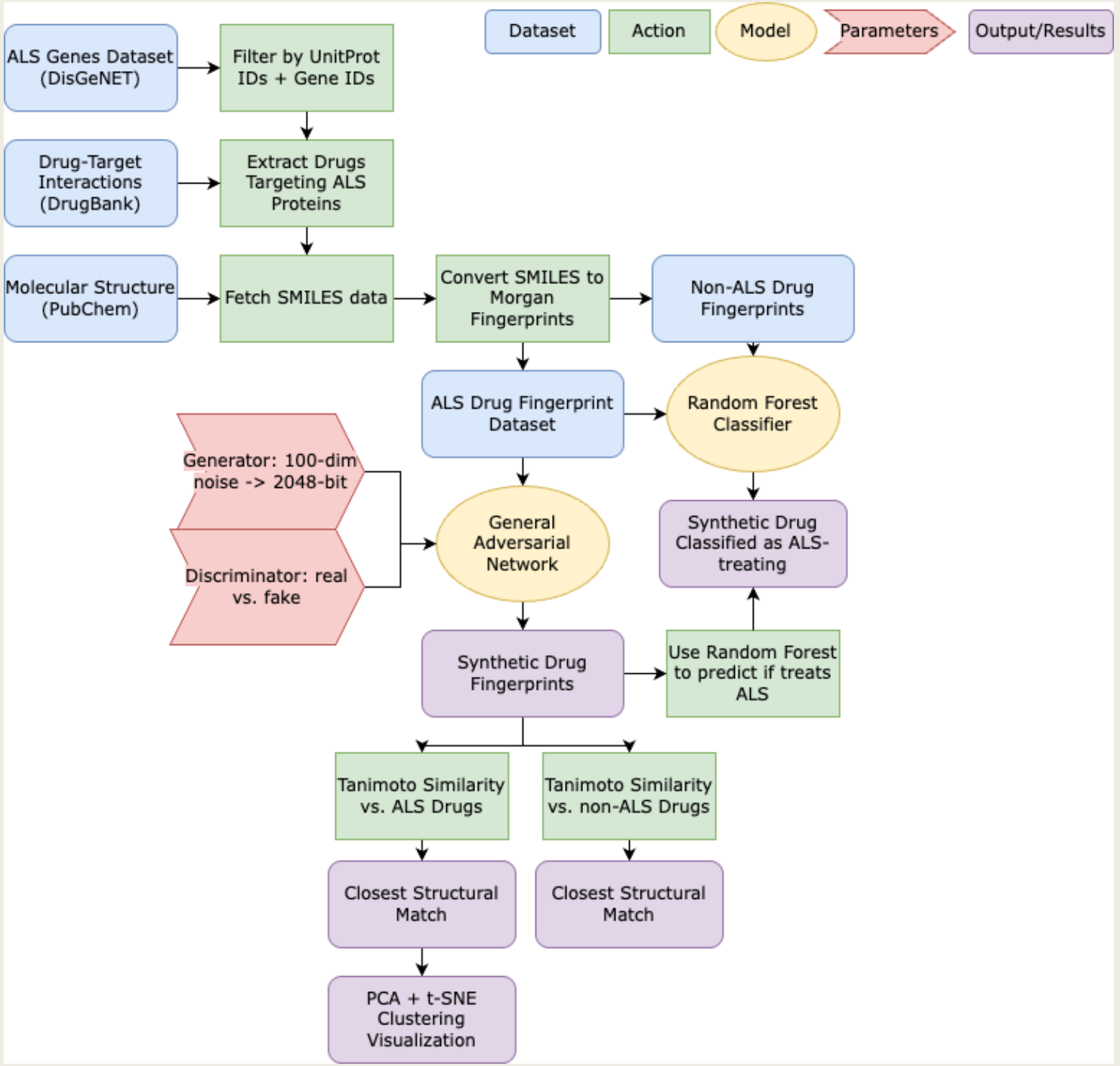
What is amyotrophic lateral sclerosis?

- ALS → a progressive neurodegenerative disease affecting motor neurons
- Symptoms→ muscle weakness, paralysis, life expectancy of 2-5 years
- No cure, few approached drugs
- Urgency for better treatment pathways

What is cheminformatics?

- Applying computer science to chemical data
- SMILES strings → molecular fingerprints (e.g., Morgan)
- Enables structure-based drug comparison

System Overview



Data Sources

- DisGeNET: ALS-associated genes
- DrugBank: drug-target mapping
- PubChem: SMILES strings for structures

Fingerprint Encoding

- Morgan fingerprints (2048-bit binary)
- Use RDKit to convert SMILES
- Input for ML models

Random Forest Classifier

- Trained to distinguish ALS vs. non-ALS drugs
- Positive class: known ALS-targeting drugs
- Negative class: random PubChem compounds

GAN Architecture

- Generator: 100-dim noise→ 2048-bit
- Discriminator: real vs. synthetic fingerprints
- Trained for 200 epochs on ALS fingerprints

Evaluation Metrics

- Random forest performance: confusion matrix, F-1 score
- Feature importance chart (top 20 fingerprint bits)
- Tanimoto similarity to fingerprint comparison

Random Forest Classifier Results

Class	*Precision*	*Recall*	*F1-Score*	*Support*
0 (Non-ALS)	0.9100	1.0000	0.9529	91
1 (ALS)	1.0000	0.1818	0.3077	11
Accuracy				**0.9118**
Macro Avg	0.9550	0.5909	0.6303	102
Weighted Avg	0.9197	0.9118	0.8833	103

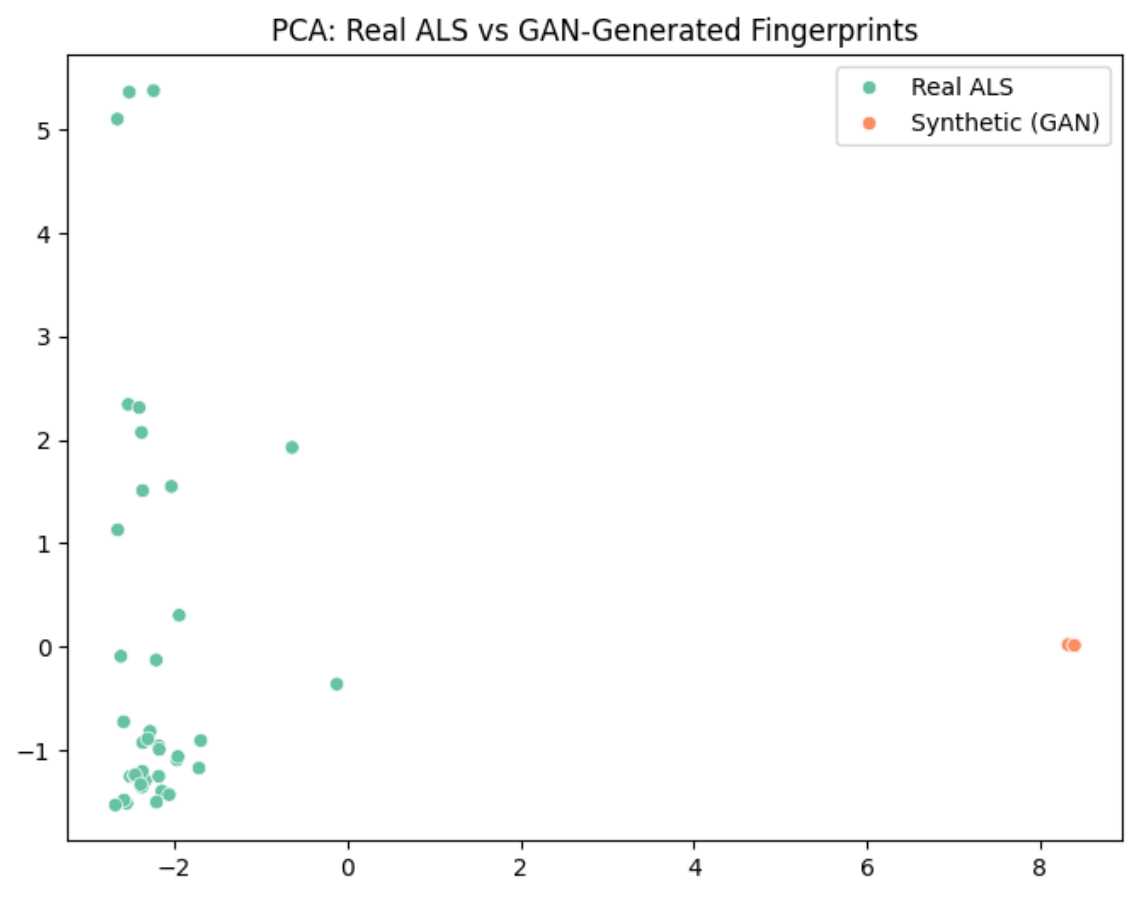
GAN Fingerprints→ Classifier Predictions

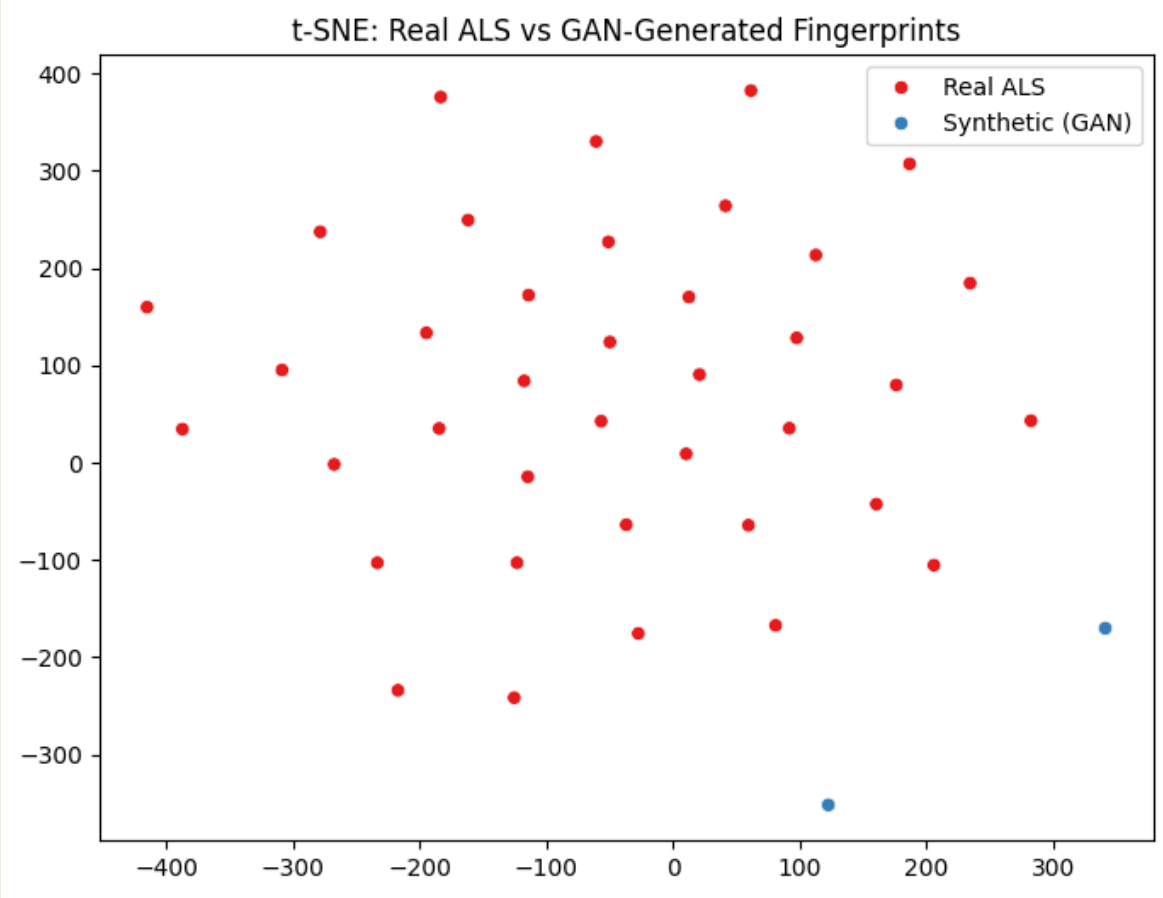
- 10/10 synthetic drugs were predicted as ALS treating by random forest
- Indicates GAN captured ALS-relevant structures

Structural Similarity to ALS Drugs

- Tanimoto scores: 0.212-0.214
- Matched to: omaveloxolone
- Some similarity to purine analgos: ~0.12

PCA and t-SNE Visualization





Conclusion

- Combined ML, cheminformatics, and public datasets
- Baseline model + GAN successfully generated ALS-like compounds
- Omaveloxolone match

Future Work

- Use graph neural networks (GNNs) for richer molecular modeling
- Improve fingerprint diversity
- Evaluate drug-likeness
- Collaborate with biochem researchers for validation