HARVARD UNIVERSITY

Interactome-Driven Drug Design

Al Shodiev

APMTH 220 Poster Session, Harvard University, Cambridge, Massachusetts 02138

Harvard John A. Paulson School of Engineering and Applied Sciences

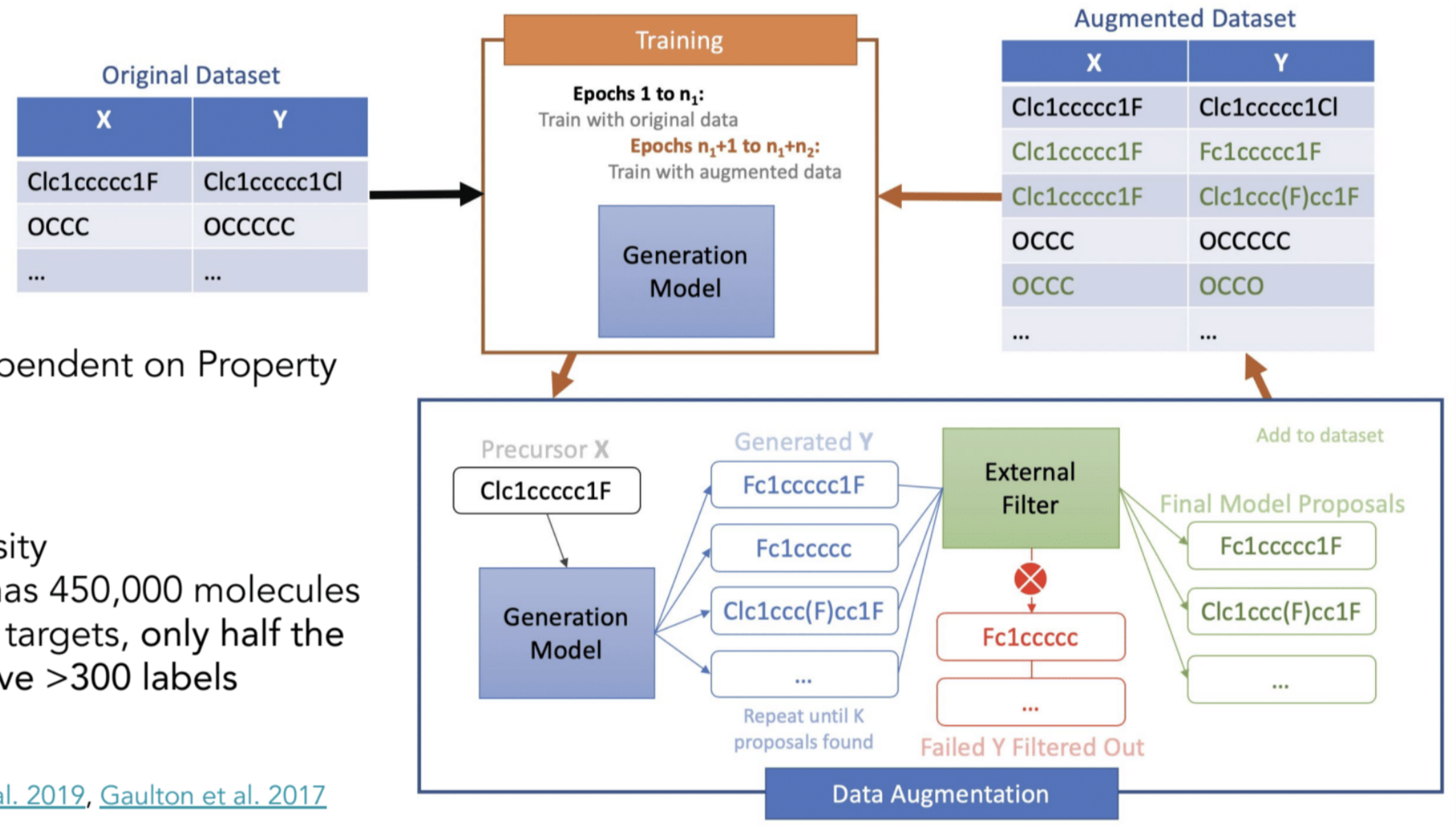
Abstract

This project addresses the bottleneck in drug development caused by data sparsity in drug-target interaction (DTI) matrices. It evaluates two data imputation strategies—matrix factorization and score propagation using an interactome dataset—validated through k-fold cross-validation and training of molecular encoder-decoder frameworks. The random walk imputation outperforms baseline matrix factorization in RMSE, and observe that semi-supervision with predictors trained on each imputed dataset for protein tyrosine phosphatase non-receptor (PTP1B) inhibition do not result in significantly different optimization success rates.

Introduction

Recent developments in deep learning have advanced drug development, but challenges remain in achieving smooth latent spaces and overcoming data scarcity in drug-target interactions. Previous solutions were limited by rigid vector embeddings and insufficient training data. In response, the project integrates matrix factorization with weighted random walk score propagation, utilizing a detailed human protein-protein interactome. This innovative approach specifically targets these issues by enhancing the data density of drug-target matrices and refining molecular optimization processes for better predictive accuracy.

Background and Notation



Highly Dependent on Property Prediction

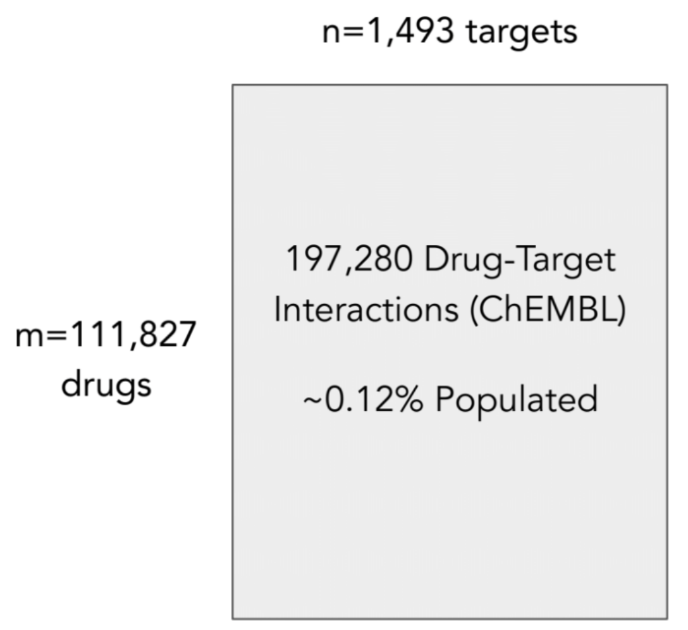
Problem:
Data Sparsity
ChEMBL has 450,000 molecules and 1,300 targets, only half the targets have >300 labels

Cites: Yang et al. 2019, Gaulton et al. 2017

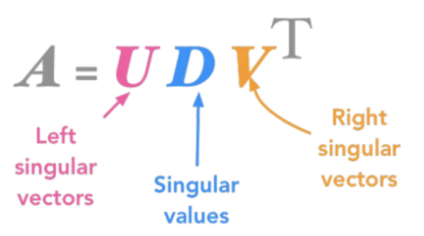
Approach

- **Approach:** Compare imputation strategies by training neural net property predictors used for molecular optimization
  1. No target augmentation (control)
  2. Target augmentation using matrix factorization imputation
  3. Target augmentation using random-walk imputed interactome

Naive Solution: Matrix Factorization on DTI Matrix



Singular Value Decomposition



Parameters
- Do not consider auto-curated DTI pairs
- Upper-bound IC50 to 100 nM

Results (averaged, 10-fold cross validation)
- 48.8 RMSE
- 29.0 MAE
- Expected given sparsity of matrix

Cite: Gaulton et al. 2017

Protein Interactome Construction and Analysis

Build PPI Graph
- Interactions: HuRI
  - Weight edges by sum of binary interactions across 9 screens, 3 assays
- Filtering (Optional): GTEx
  - Ex. remove all nodes with low/no expression in all brain tissues

Result
- |N|=8,275 proteins, |E|=52,569 interactions
  - Toggle inclusion of promiscuous nodes as determined by centrality measures (degree, eigenvector, closeness, betweenness)
- Highly connected nodes make biological sense and provide heuristic validation of approach

Centrality Metric	Top 5 Protein Nodes	
Degree	PICK1	0.0475
	LNX1	0.0423
	WDYHV1	0.0390
	TLE5	0.0331
	GOLGA2	0.0325
Eigenvector	TLE5	0.1671
	PICK1	0.1586
	MTUS2	0.1486
	GOLGA2	0.1476
	LNX1	0.1466
Closeness	LNX1	0.3660
	TLE5	0.3529
	WDYHV1	0.3518
	PICK1	0.3501
	KIFC3	0.3501
Betweenness	LNX1	0.0627
	PICK1	0.0482
	WDYHV1	0.0438
	SDCBP	0.0351
	UBQLN2	0.0330

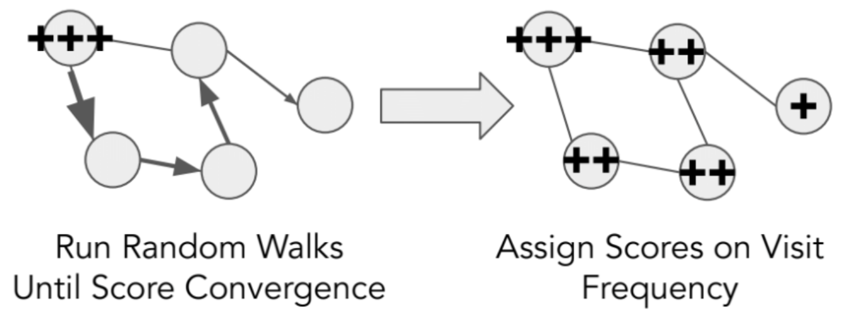
Table 1:
Top connected protein nodes in protein-protein interaction graph using filtering of genes expressed in at least one brain tissue. Different methods of centrality are shown.

Cites: Luck et al., Lonsdale et al.

Weighted Random Walk on Protein Interactome

Algorithm Setup
- Probabilistic Transition Matrix T
  - Build from PPI graph adjacency matrix
- Drug Seed Set S
  - For each drug, set of proteins with nonzero normalized DTI scores (ChEMBL)
- Score Vector p ∈ $R^{|N|}$
  - All non-seed set entries = 0
  - Seed set entries proportional to DTI score

Iterative Random Walks per Drug
- Construct $p^{(0)}$ from S
- Use T to iteratively update p until score convergence



Run Random Walks Until Score Convergence

Assign Scores on Visit Frequency

$$\mathbf{p}^{(t+1)} = \beta\mathbf{T}\mathbf{p}^{(t)} + (1-\beta)\mathbf{s}$$

Update Rule

β: local consistency (similar labels for neighbors) vs. global consistency (correct scores for seed proteins)

Results (averaged, 10-fold cross validation)
- 39.6 RMSE (better)
- 35.1 MAE (worse)
- Random walk more conservative in predictions → produces fewer outliers

Cites: Hamilton et al., San Vicente et al., Zhou et al.

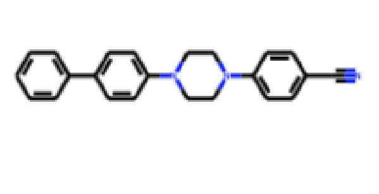
Results

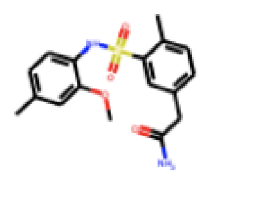
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
RMSE (testset)	49.2153	49.0103	49.2689	48.8027	48.7409	48.4714	48.6724	48.8348	48.3274	48.3795	48.7724	0.3095
MAE (testset)	29.4916	29.2606	29.5092	29.0598	28.8949	28.6085	28.7688	29.1414	28.5559	28.5206	28.9811	0.3512
Fit time (sec)	8.55	8.23	8.63	7.88	8.30	8.47	7.95	8.81	8.97	8.82	8.46	0.35
Test time (sec)	0.21	0.11	8.63	0.11	0.13	0.11	0.11	0.18	0.11	0.11	0.14	0.04

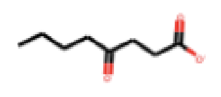
Table 2:
Singular value decomposition 10-fold cross-validation results.

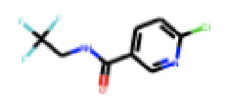
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
RMSE (testset)	39.655	39.753	39.563	39.704	39.567	39.485	39.602	39.434	39.589	39.786	39.614	0.112
MAE (testset)	35.301	35.051	35.215	35.094	34.995	35.132	34.995	35.126	35.124	34.869	35.090	0.121

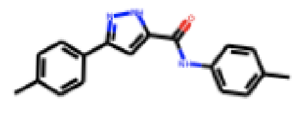
Table 3:
Weighted random walk score imputation 10-fold cross-validation results.


Predicted Score: 84.9


Predicted Score: 87.06


Predicted Score: 87.37


Predicted Score: 88.22


Predicted Score: 89.68

- Success Rate = fraction of test set for which model outputs satisfy constraints of ground truth property predictor
- For PTP1B, success rate of all approaches is comparable at 0.1 (with IC50 constraint at c = 90). Similar top molecules nominated

Discussion

Using the random walk propagation technique alongside traditional matrix factorization, I addressed the data sparsity in DTI. Random walk yielded a lower RMSE of 39.6 compared to 48.8 in matrix factorization but showed higher mean MAE at 35.1 versus 29.0, indicating a more conservative prediction distribution. Both methods demonstrated equivalent success rates in downstream molecular optimization tests on a set of 5,000 molecules, highlighting the need for larger, varied datasets to enhance model performance and validation accuracy.

Future Direction

- Explore substructure motifs specific to targets to guide molecular generators towards relevant chemical spaces.
- Develop a joint objective function that combines matrix factorization and random walk approaches.
- Extend the use of weighted random walk beyond DTI to propagate other properties across the interactome.