Interactome-Driven Drug Design

Abstract

Rapid generation of small molecule compounds targeting specific proteins remains a bottleneck for drug development even after the pathophysiology of disease has been well-characterized. Computational optimization of molecules for the task of perturbing a high-value protein target has the potential to overcome this obstacle in the drug development process. Previous efforts at using drug-target interaction (DTI) scores for de novo molecular design were hampered by data sparsity and difficult-to-navigate latent space representations of small molecule candidates. Recent appli- cation of semi-supervised learning and hierarchical graph-based molecular models [31] [17] have yielded a smoother latent space and improved property optimization accuracy, but the problem of data sparsity in DTI matrices persists. In this work, I evaluate two key strategies for data imputation on DTI matrices: (i) matrix factorization with Singular Value Decomposition; (ii) score propagation using a high-quality proteome-wide interactome dataset. Each imputation strategy is validated using (i) k-fold cross-validation on the DTI matrix; and (ii) by training state-of-the-art molecular encoder-decoder frameworks (with SMILES [20] and junction tree [17] representations) using fixed feed-forward neural net property predictors [27] trained on the augmented datasets, and then comparing molecular optimization success for the protein tyrosine phosphatase non-receptor (PTP1B) inhibition task. I find that the random walk imputation outperforms baseline matrix factorization in RMSE, and observe that semi-supervision with predictors trained on each imputed dataset for PTP1B inhibition do not result in significantly different optimization success rates. Future directions and method improvements are later discussed

Introduction

Recent advances in deep learning provide powerful tools to aid in the process of drug develop- ment. They can roughly be divided into two main use cases: property prediction and molecular optimization. In the prior case, feed-forward neural networks are trained to predict a property value (such as solubility, toxicity, or inhibition of a particular protein target) using a molecular representation (such as a SMILES string [20] or multi-resolutional 2D molecular graph [17]) as input. The problem of molecular optimization is situated in a general class of translation problems for item property optimization, from drug design [18] to deep neural network optimization [16]. State-of-the-art approaches in drug design imitate natural language processing over text, but on a tougher domain; they encode rigid, combinatorial 3D molecules in a latent space of low-dimensional vectors. Previous works in de novo drug generation used strategies including gradient ascent in the latent space towards a molecule optimized for a property [17], and reinforcement learning in which a property predictor yields a reward to a generator as the molecule is progressively constructed [5] [7].

Though de novo generation efforts are improving, they face two major problems. First, fixed-length vector embeddings of molecules do not produce smooth latent spaces. This hampers property predictor gradient ascent in that space, and thus coherent generation of a molecule optimized for some property. Second, there is a general lack of training data in drug-target interaction, especially when compared to the massive corpora of the natural language community. For example, the popular ChEMBL dataset considers interaction of 450,000 molecules with 1,300 targets, but only half the targets have over 300 labels [11]. This impedes every area of the encoding-decoding task.

A recent technique [6] combats the latent smoothness problem by exploiting the fact that property prediction for some particular feature is far more accurate than learning a latent space through training on molecular reconstruction. This is intuitive – determining the “meaning” of a molecule in regards to one particular property is a far easier task than encoding all elements of the molecular structure salient for reconstruction in a small latent space. A property predictor f is trained on a set of known tuples of molecules and properties (X_i, p_i), after which its parameters are fixed. Then, it is used for semi-supervised training of an encoder-decoder. The generator is initially trained in molecular reconstruction (i.e. with

tuples (X_i, X_i)) for n_1 epochs. Next, in each epoch of "target augmentation", up to L output molecules Y_i are generated for each precursor molecule X_i in the dataset X. Output molecules Y_i that surpass a constraint c as determined by the property predictor ($f(Y_i) > c$) are added back into the generator training set for n_2 future iterations. This method can be formulated as an expectation-maximization approach, in which the generator is "pushed" towards latent domains relevant to the property of interest. It thus successfully addresses the problem of bumpy latent spaces in fixed-length vector molecular embeddings, but does not address the general lack of training data in drug-target interaction.

Understanding emergent properties from interactions between components of biological systems represents one of the fundamental challenges of modern computational and systems biology. In particular, disruptions in protein-protein interactions form the basis of both the pathophysiology and therapeutic approaches for many diseases. Systematic examination of missense mutations from Mendelian disorders has revealed that disrupted protein-protein interactions are more frequent than changes in protein stability [28]. Organization of variants from genome-wide association studies into network-based pathways have enabled mapping of causal disease mutations across a wide variety of systems, from immune to neurodevelopmental disease [9], [2], [22].

In this project, I aim to compare the success of two approaches to drug-target interaction data augmentation, towards addressing severe data sparsity. The strategies considered are (i) matrix factorization (which uses only the drug-target interaction matrix in imputation) and (ii) weighted random walk score propagation (which uses a protein-protein interaction graph in imputation). I leverage advances in computational methods for protein-protein interaction analysis to construct a high-quality human protein-protein interactome dataset. I use this completed graph along with small molecule-protein binding data for my (ii) strategy. Validation of each strategy is performed at two depths: drug-target interaction matrix reconstruction, and downstream molecular optimization success. In the first case, I perform k-fold cross-validation for each approach and compare. For the second case, I train a property predictor using the fully imputed drug-target interaction scores of each approach, to identify small molecule compounds predicted to well-target a disease-relevant protein of interest. I consider in the downstream optimization task the inhibition of protein tyrosine phosphatase non-receptor 1, and use my baseline predictor, which is trained on a dataset with no imputation, for validation. This follows approaches from [17], [27].

I take this approach to improve property prediction. Recent works have used machine learning on low-level drug and protein representations, along with molecular docking simulation data to predict drug-target interactions [10]. However, to my knowledge, no work has used both protein interactome and drug-target interaction information to infer how drug effects are propagated throughout the protein network, therefore improving drug-target interaction matrix density [33] [29] [21].

Background and Notation

Drug-Target Interaction (DTI) Drug-Target Interaction (DTI) refers to the binding affinity between small molecule compounds (drugs) and specific protein targets. This interaction is crucial in the process of drug development, as the ability of a drug to target specific proteins determines its efficacy in treating diseases. The IC50 value is often used to quantify the interaction strength, representing the concentration of a drug required to inhibit a biological process by 50%. Protein-Protein Interaction Network (Interactome) The interactome is a network that maps the interactions between proteins within a biological system. In this project, the interactome is represented as an undirected, weighted graph, where nodes correspond to proteins and edges denote interactions. The weights of the edges reflect the strength of the interactions based on experimental assays, allowing for analysis of how protein disruptions contribute to disease pathology and therapeutic strategies. Molecular Representations Two main representations of small molecule compounds used in this study are: SMILES (Simplified Molecular Input Line Entry System): A text representation that encodes the structure of molecules in a linear string format. Junction Tree: A hierarchical graph-based representation that decomposes molecules into substructures, capturing structural motifs and features in a tree format. Matrix Factorization Matrix factorization is a collaborative filtering

technique used to impute missing DTI scores. By decomposing the sparse DTI matrix into a product of two lower-dimensional matrices, it approximates the unknown interaction scores, effectively filling in the gaps within the matrix. Random Walk Score Propagation The random walk algorithm propagates DTI scores across the interactome graph. Starting with known drug-target interaction scores as seeds, scores are iteratively updated based on protein connectivity within the graph. The transition matrix is constructed from the adjacency matrix of the graph, and the algorithm continues until convergence. Molecular Optimization Molecular optimization aims to find novel compounds with optimized properties. This project uses an encoder-decoder framework to translate existing molecules into optimized compounds with desired properties. The neural network predictor is trained on augmented datasets, where target augmentation is used to iteratively refine the training set with new molecules predicted to meet certain property constraints.

Notation
G: Interactome graph representing the protein-protein interaction network.
N: Set of proteins/nodes in the interactome graph.
E: Set of edges in the interactome graph, representing interactions between proteins.
M: Drug-target interaction matrix containing IC50 values for known drug-protein pairs.
S: Seed set of proteins with known interaction scores for a given drug.
Xi, Yi: Input molecule and generated molecule in the molecular optimization framework, respectively.
f: Property predictor, a neural network trained to predict the IC50 values based on molecular representations.
β: Hyperparameter that balances local vs. global consistency during random walk score propagation.
ε: Early stopping threshold for the random walk algorithm.
C, K, Z: Constants controlling the number of generated molecules, distinct translations, and samples in molecular optimization.

Proposed Approach

Interactome Graph and Drug-Target Interaction Matrix Generation

The protein-protein interaction graph (interactome G) was generated with the Human Reference Interactome dataset (HuRI) [24]. HuRI is a matrix of interactions between all pairwise proteins from annotated human open reading frames. The data was generated using a yeast two-hybrid system and repeated over 12 separate assays. Previous analyses demonstrated that the number of rounds of assays in which interactions between two proteins were recovered is proportional to the length and strength of interactions between them. For example, protein pairs found in protein complexes together were recovered in more assays. I generate the undirected interactome graph $G = (N, E)$ from the HuRI matrix, using the NetworkX package. Each edge $(x_i, x_j) \in E$ is weighted by the proportion of assays in which interactions were recovered between x_i and x_j in HuRI.

I use the ChEMBL database [11] in generation of the drug-target interaction matrix M. ChEMBL consists of small molecule interaction data obtained from published articles and manually curated to ensure that assay representations are accurate. In particular, I consider only drug-protein interactions in ChEMBL that were curated with an intermediate or expert level of confidence, meaning that the assays were manually verified after auto-curation. I use IC50 as my binding metric, because only IC50 (as opposed to KD) was annotated consistently by high-confidence cura- tion. This initial filtering led to an interaction matrix of 125,458 drugs, 1,778 targets, and 220,861 recorded drug-target interactions.

Namespace Mapping

The protein-small molecule IC50 assay pairings were represented by CHEMBL IDs and the HuRI- derived interactome node set was annotated using ENSEMBL IDs [25]. In order to assure a fair com- parison

between the matrix factorization and interactome imputation approaches, target ChEMBL IDs were mapped to ENSEMBL IDs contained within the interactome node set N . Three routes mapping from the ChEMBL to ENSEMBL namespace were considered. First, the pybiomart pack- age [4] was used to directly map ChEMBL IDs to ENSEMBL IDs. For molecules for which mapping was not found through this direct path, I used two additional methods using CHEMBL Web Re- source Client mappings [11]. I used this client to map from CHEMBL to HGNC ID and Gene Name. pybiomart was used to map from HGNC ID and Gene Name to ENSEMBL. The combination of these methods yielded 1,493 total protein targets. I constrain the matrix M to only those targets to yield 111,827 interacting drugs and a total of 197,280 recorded interactions.

Matrix Factorization on Drug-Target Interaction Matrix

I choose matrix factorization as a benchmark among many options (ex. deep learning [10]) for drug-target interaction matrix imputation because it requires only known interaction values and still serves as the fundamental algorithm of state-of-the-art approaches [8]. To better normalize the IC50 values for both matrix factorization and random walk imputation, I clip values over 100 nM (weak binders) to 100. The values are therefore on the range [0, 100]. I perform Singular Value Decomposition on our sparse matrix M to impute scores. This formulates the matrix of drug-target interactions as a recommendation system, without reliance on the interactome G. I evaluate the learning algorithm with 10-fold cross-validation.

Random Walk on Protein Interactome Graph

I use the protein-protein interaction graph G for imputation of missing values in matrix M using a weighted random walk algorithm. Previous approaches have found success in propagation of sentiment scores across word embedding graphs learned from domain-specific lexicons [13] [15] [3].

Following [13], the adjacency matrix of the interactome E (weighted by HuRI assay results) is used to construct a symmetric transition probability matrix $\mathbf{T} = \mathbf{D}^{-1/2} E \mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a matrix with the column sums of E in the diagonal. I am able to assume E is symmetric due to the symmetric nature of protein-protein interactions. This same transition matrix $\mathbf{T}$ is used for all drugs, and describes the probability of traversal across G.

For each drug, the "seed set" S contains all of its recorded interaction scores from matrix M . I construct a score vector $\mathbf{p} \in \mathbf{R}^{|N|}$, in which each entry is 0 if it corresponds to a protein not in S, and otherwise is proportional to the recorded score for that protein. Drug-target interaction scores were normalized from the range [0, 100] to [0, 1], such that a higher value signals stronger interaction. $\mathbf{p}$ is normalized to sum to 1. Iterative random walks are executed for each drug over the interactome, using its respective seed set and initial score vector $\mathbf{p}^{(0)}$. In particular, I use $\mathbf{T}$ to iteratively update $\mathbf{p}$ until numerical convergence ($|\mathbf{p}^{(t)} - \mathbf{p}^{(t+1)}| < E$):

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

where $\mathbf{s}$ is a vector with values set to 1 -in entries corresponding to the seed set S, and zero elsewhere. The β hyperparameter determines the extent to which the algorithm favors local con- sistency (similar labels for neighbors) versus global consistency (correct scores for seed proteins). Lower β values emphasize the latter. As seen above, E is an "early stopping" hyperparameter. I chose $\beta = 0.9$ and $E = 1e - 6$ given best practices from previous works employing propagation on graphs of comparable size and density [3] [13] [15]. In the future, I will employ grid search and cross-validation for these two parameters. However, this was not possible given current hardware constraints (10-fold cross validation multiprocessed over 32 vCPUs with 250 GB main memory took over 48 hours).

For each drug, protein nodes $\in N$ reachable through random walk from the drug's seed set S are assigned a score proportional to their probability of being hit during a random walk iteration (see Figure 1). This can be interpreted biologically as a predicted probability of the drug interacting with each protein, given all its known interactions and relevant interactome pathways. In my implementation, if a node was never reached via random walk, it was assigned a default score equal to the global average of the drug-target interaction matrix M , to avoid applying an extreme prior 0 for missing data. This default imputation occurred only for drugs which had no drug-target interaction values other than measured, maximally weak interaction (IC50 = 100 in our clipped case). The high connectivity of the graph meant that any non-zero drug-target interaction score for a particular drug would result in propagation to virtually all nodes.

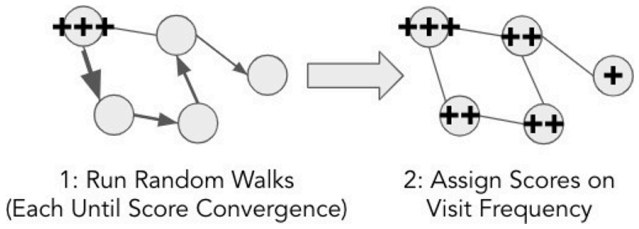

Figure 1: Graphical representation of random walk procedure.

Downstream Validation

The two data imputation approaches outlined above were evaluated in their downstream impact on property values of the best produced de novo molecules. These methods are general and can be applied to any protein target in the namespace-resolved drug-target matrix M. I choose the protein tyrosine phosphatase non-receptor type 1 (PTP1B) as our target for validation because: (i) the target is contained in the HuRI interactome; (ii) of such targets, the protein has the most observed drug-target interactions; (iii) the baseline matrix factorization yields high heterogeneity (over 9,000 unique score predictions for 111,827 drugs) and low cross-validation error for the target. I compare three datasets, each composed of tuples of (SMILES string, PTP1B IC50 Score): (i) base measured scores from ChEMBL; (ii) matrix factorization imputed scores; (iii) random walk imputed scores. Dataset (i) has 1,011 entries. The raw dataset (ii) has 111,827 entries and strong negative bias, because the majority of imputations assign maximum IC50 = 100. Dataset (iii) has 12,805 entries, pertaining to all 12,805 drugs that traversed PTP1B during a random walk propagation, and is fairly heterogenous. I normalize for negative bias in (ii) by randomly selecting a subset of drugs with prediction IC50 = 100 to report, such that the total size of normalized dataset (ii) is also 12,805 entries (approximately 1/4 of the dataset is then IC50 = 100).

In our validation, the manner in which data imputation affects drug generation is through semi- supervision using a pre-trained and fixed feed-forward neural network predictor. I use the Chemprop framework [32] to train a predictor f for each of our three datasets. The architecture takes as input a SMILES string, begins with a layer conducting message-passing over a junction tree molecular representation as described in [31], has two hidden layers of size 300, and uses ReLU activation.The encoder-decoder architecture is adapted from the Junction Tree Variational Autoencoder (JT- VAE) [17]. I use the Zinc dataset [19] to create a library of chemically valid molecular subgraphs, termed atomic clusters. This vocabulary is of size 780. I split the dataset into a training set of 250,000, validation set of 25,000, and test set of 5,000 molecules. Informed by [17], the en- coder and decoder of the JTVAE each have depth 1 and 450 hidden units and the latent vector has size 56 (total of 5,110,000 parameters). I use the following strategy to train our JTVAE in the molecular reconstruction task, prior to any target augmentation [6]. The initial dataset D consists of tuples (X_i, X_i) of SMILES strings from Zinc, in which the target of the generator is to reconstruct the input molecule. I pretrain on D for 3 epochs by learning an autoencoder similar to the final VAE, but lacking a regularizing β parameter. Then, I train with said parameter for 7 epochs. All parameters, both variational and generative, are initialized by random sampling from $Gaussian(0,\ 0.01)$, and are jointly and stochastically optimized using maximum a posteriori estimation with the validation set. Stepsizes are

adapted with Adagrad [12]; the Adagrad global stepsize parameters were chosen from 0.01, 0.02, 0.1 based on performance on the training set in the first few iterations. $\beta = 0.005$ based on the approach of [17].

For each property predictor f, I conduct independent target augmentation. In each epoch of target augmentation, I take an *Augmentation* and a *Training* step. In *Augmentation*, I construct the generator training dataset $D^{(t+1)}$ by feeding each input $X_i \in D$ (the original training set, not D_t) into the model up to C times to sample C candidate translations $Y^1 ... Y^C$. I add the first K distinct translations satisfying the constraint $f(Y_i) < c$ to $D^{(t+1)}$. In *Training*, I train the JTVAE model over the new training set $D^{(t+1)}$ for one epoch. Target augmentation is run for 3 epochs per property predictor. In evaluation between the predictors, I translate each molecule in the test set Z times, and measure success rate by the fraction of molecules X for which any of the corresponding outputs $Y_1 ... Y_Z$ meet the constraint c. Per [6] I set C = 200, K = 4, and Z = 20.

Results

Completed Interactome Graph and Drug-Target Interaction Matrix O generated an undirected, weighted protein-protein interactome graph G with 8,275 proteins and 52,569 interac- tions using data from HuRI. I also generated a drug-protein target interaction matrix, weighted by drug impact on protein IC50 value, using the ChEMBL database (see Methods for details of graph and matrix generation).

	Centrality Metric	Protein Node	Value
0	Degree	PICK1	0.0475
1	Degree	LNX1	0.0423
2	Degree	WDYHV1	0.0390
3	Degree	TLE5	0.0331
4	Degree	GOLGA2	0.0325
5	Eigenvector	TLE5	0.1671
6	Eigenvector	PICK1	0.1586
7	Eigenvector	MTUS2	0.1486
8	Eigenvector	GOLGA2	0.1476
9	Eigenvector	LNX1	0.1466
10	Closeness	LNX1	0.3660
11	Closeness	TLE5	0.3529
12	Closeness	WDYHV1	0.3518
13	Closeness	PICK1	0.3501
14	Closeness	KIFC3	0.3501
15	Betweenness	PICK1	0.0482
16	Betweenness	WDYHV1	0.0438
17	Betweenness	SDCBP	0.0351
18	Betweenness	UBQLN2	0.0330

In evaluating the generated interactome, I wanted to verify data integrity, establish sufficient connectedness for successful score propagation even from a small seed set S, and use known infor- mation on certain disease pathways to sanity check any found well-connected nodes. First, I found that the largest connected component of G is 8,152 proteins, meaning that propagation from each node will reach the majority of the graph. Next, I filtered G to only include proteins with GTex expression level in some brain tissue over 5 [23]. This would allow us to more conclusively locate nodes critical to neurological diseases, thereby providing some basis of interpretation of validity. Table 1 on the right shows '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.'

Several metrics of connectedness were analyzed, and several highly-connected nodes were identified (Table 1). Many of the same nodes appeared highly across the various metrics, asserting that such elements would facilitate propagation across most of the graph's nodes.

Several of these genes make biological sense in the context of interacting with many different proteins, thereby providing some reassurance of the validity of our interactome. For example, UBQLN2 is implicated in a number of neurodegenerative diseases and its role in the proteasomal degradation pathways makes sense in terms of its interaction with a number of other proteins in the cell [14]. Similarly, PICK1 is implicated in neurological diseases such as epilepsy and is thought to function as an adaptor protein that interacts with and organizes membrane proteins [1].

Furthermore, LNX2 and UBQLN2 are both involved in the ubiquitin-proteasome pathway, which may explain why they are so highly connected. WDYHV1, a glutamine deamidase, and GOLGA2, all proteins

involved in interactions with or modifications of many proteins, may also represent

	Fold	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Mean	Std
0	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
1	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
2	Fit time (sec)	8.5500	8.2300	8.0300	7.8800	8.3000	8.4700	7.9500	8.8100	8.9700	8.8200	8.4600	0.3500
3	Test time (sec)	0.2100	0.1100	8.6300	0.1100	0.1300	0.1100	0.1100	0.1800	0.1100	0.1400	0.1400	0.0400

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

	Fold	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Mean	Std
0	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
1	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.

important conserved protein quality control units in the proteome. In this way, even initial anal- ysis of important aspects of the proteome provides insight into highly-connected protein nodes. These could be advantageous or deleterious to drug development, depending on whether or not the presence of these proteins is important in the pathophysiology of the disease of interest. Matrix Factorization of Interactome and Drug-Target Interaction Matrices I employed an interactome-agnostic approach to drug-target interaction value imputation as a baseline to which one may compare the effectiveness of random walk imputation. In this method, I perform Singular Value Decomposition on the drug-target interaction matrix to impute scores across the matrix (see Methods section for details). The root mean squared error (RMSE) and mean absolute error (MAE) of 10-fold cross validation are shown in Table 2. The large error here is unsurprising given the sparsity of the input matrix; the matrix has a density of approximately 0.01%. Random Walk Imputation of Protein-Protein Interactome Using the interactome G, target interaction scores are imputed for each drug from a seed set of drug- target interaction scores S. I evaluate the success of the method in matrix imputation against matrix factorization with 10-fold cross validation. The default value (described in the Methods section, used for targets that are not reached during random walk) is determined by averaging across the 90% of the total drug-target interaction data points used as training data in each fold. Root mean squared error (RMSE) and mean absolute error (MAE) are shown in Table 3. Fit time and test time were not computed by default; each training fold took several hours, but testing is lookup-based. Comparison of Imputation Approaches with Downstream Validation The downstream success of these imputation strategies was evaluated by using the augmented datasets to train neural net property predictors, which were in turn used in a molecular optimization task. In particular, I employ the target augmentation approach based on [6] and outlined in the Methods section. The three approaches compared are: (i) no target augmentation (control); (ii) target augmentation using a predictor trained on matrix factorization imputed data; and (iii) target augmentation using a predictor random walk imputed data. For the latter two, I conduct target augmentation for three epochs, to produce a final model. I define the success rate as the fraction of molecules X for which any of the corresponding model outputs Y1...YZ (using the final model for translation) satisfy the constraint of the ground truth property predictor – the one that is trained on a non-augmented dataset. I find that the vast majority of generated molecules have predicted targeting of PTP1B close to 100. I vary the constraint of predicted IC50 to c = 90, and find that the success rate of all approaches is 5 in 5000 test examples (0.1%). Included in Figure 2 are 2D molecular representations of the five molecules generated (identically) from each final model, and their associated predicted IC50 score.

Discussion and Conclusion

In this work, I develop a novel random walk propagation technique to impute drug-target interaction scores using protein interaction data. I aimed to address the extreme sparsity of drug-target datasets, resulting from high experimental collection costs, towards improving downstream molecular optimization. I obtained a matrix of high-confidence protein-drug interaction values from the ChEMBL database [11], and used a high-quality protein interaction dataset representing all binary interactions between open reading frames in the human genome [25] to generate an interactome graph. The interactome was verified to have sufficient connectivity for successful usage of the random walk algorithm, and it was filtered by brain gene expression data to assert that influential nodes implicated in known neurodegenerative pathways were surfaced. I used two drug-target interaction score imputation methods: (i) matrix factorization with Singular Value Decomposition; (ii) score propagation using iterative weighted random walks from a seed score set in the interactome. Imputation success was measured at two depths: (i) k-fold cross-validation on the drug-target interaction matrix; and (ii) molecular optimization success by use of imputed data in training state-of-the-art molecular encoder-decoder frameworks. In the first comparison of methods, we observe that the mean RMSE across all folds of the random walk imputation approach outperforms that of matrix factorization, approximately 39.6 versus 48.8. However, it performs worse in mean MAE; 35.1 to 29.0. This discrepancy is likely explained by treatment of outliers by the two approaches. RMSE punishes outliers more than MAE, and random walk tends to be more conservative in its predictions than matrix factorization. In random walk for a particular drug, scores of each protein start at maximum IC50 and are gradually lowered through propagation. Therefore, predictions will generally be less dispersed, and slightly tend towards IC50 = 100 (though this is dependent on hyperparameters). In order to test this hypothesis, I will generate synthetic edge-case interactome topologies in which many nodes with low IC50 scores (strong binders) are very well-connected, to see if this results in a larger range of predicted values and therefore wider-falling outliers. Next, I consider the downstream method comparison. We can observe uniformity in success rate across all approaches in PTP1B target optimization success rate, as computed on a test set of 5,000 held-out molecules. This could have been the case due to the base truth property predictor having a limited scope in assessing the ability of any generated drug to target PTP1B, given that it is trained on only 1,011 entries. Therefore, in the future I will consider using another model less likely to overfit on such a small dataset, such as a Support Vector Machine. Another cause for this result may be the fact that the ground truth data used to train the base predictor is contained with relatively little perturbation in the matrix factorization and random walk datasets. Those data points are likely to be the ones most informative in guiding the target augmented models to produce the same success rate. In order to alleviate these concerns, I will aim to secure a much larger dataset for a particular target (such as through a proprietary source) so that an external predictor can be used for proper validation. Ultimately, I found that previous approaches had used dependent training and validation datasets, with high intersection due to ChEMBL as a common source [6] [27]. Another potential issue resulting in the same success rate across the three approaches is ineffective target augmentation. Due to a complicated bug in the RDKit SMILES to Molecule conversion, a suboptimal proportion of the original training dataset precursor molecules could be successfully translated. Therefore, the target augmentation training datasets had overly high similarity to the original datasets; in the case of no successful translation meeting the constraint $f(Yi) < c$, the original training data tuple is just added. My next priority will be to fix this error and immediately increase the impact of target augmentation. In summary, I find that I am able to successfully build and impute both an interactome and drug-target interaction graph using matrix factorization and a novel random walk algorithm. Using both the base, un-imputed data, as well as the imputed datasets, I am able to propose potential small molecule candidates for protein targets (in particular, tested for PTP1B). However, the extent to which performance is improved by the presence of imputed data is unclear, given the current target augmentation result.

Comparison to Original Proposal:

The original proposal aimed to create a model to predict adverse drug events from chemical structures as an intermediate step, with predicting symptoms for novel drugs being the next step. However, as I was working through the project, it became clear that the extreme sparsity of drug-target interaction data limited my ability to effectively address the original aims. This led to a pivot towards addressing the challenge of data sparsity in drug-target interactions, which I identified as a critical bottleneck in molecular optimization.

In this revised project, I developed a novel random walk propagation technique to impute drug-target interaction scores using protein interaction data. The focus shifted to developing a methodology that could overcome data sparsity and improve downstream molecular optimization, leveraging high-confidence interaction values from the ChEMBL database and comprehensive protein interaction data to build an interactome graph. By employing matrix factorization and score propagation techniques, I aimed to improve the quality of drug-target interaction predictions.

Although the project deviated from the original proposal, my work demonstrates a proof-of-concept for utilizing interactome data to impute drug-target interaction scores, providing a general approach that can be adapted for various drug-target prediction tasks. This shift in focus was guided by the need to address fundamental data challenges that limited the project's original scope.

References

1. Federica Bertaso, Chuansheng Zhang, Astrid Scheschonka, Frédéric De Bock, Pierre Fontanaud, Philippe Marin, Richard L Huganir, Heinrich Betz, Joël Bockaert, Laurent Fagni, et al. Pick1 uncoupling from mglur7a causes absence-like seizures. *Nature Neuroscience*, 11(8):940, 2008.
2. Siwei Chen, Robert Fragoza, Lambertus Klei, Yuan Liu, Jiebiao Wang, Kathryn Roeder, Bernie Devlin, and Haiyuan Yu. An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nature Genetics*, 50(7):1032, 2018.
3. Thomas Navin Lal Jason Weston Dengyong Zhou, Olivier Bousquet and Bernhard Scholkopf. Learning with local and global consistency. 2004.
4. Julian DeRuiter. pybiomart: A simple and pythonic biomart interface for python.
5. A. Zhavoronkov et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotech*.
6. K. Yang et al. Iterative target augmentation for effective conditional generation. In Submitted to International Conference on Learning Representations.
7. M. Popova et al. Deep reinforcement learning for de novo drug design. *Science Advances*.
8. Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 14(3):646–656.
9. Hai Fang, Hans De Wolf, Bogdan Knezevic, Katie L Burnham, Julie Osgood, Anna Sanniti, Alicia Lledó Lara, Silva Kasela, Stephane De Cesco, Jörg K Wegner, et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature Genetics*, 51(7):1082.
10. Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In IJCAI.

11. A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magarinos, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach. The ChEMBL database. *Nucleic Acids Res.*, 45(D1):D945–D954.
12. H. B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. COLT, 2010.
13. William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. NIH Public Access, page 595.
14. Roland Hjerpe, John S Bett, Matthew J Keuss, Alexandra Solovyova, Thomas G McWilliams, Clare Johnson, Indrajit Sahu, Joby Varghese, Nicola Wood, Melanie Wightman, et al. Ubqln2 mediates autophagy-independent protein aggregate clearance by the proteasome. *Cell*, 166(4):935–949.
15. German Rigau, Inaki San Vicente, Rodrigo Agerri, and Donostia-San Sebastian. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages.
16. Zhihao Jia, Oded Padon, James Thomas, Todd Warszawski, Matei Zaharia, and Alex Aiken. Taso: Optimizing deep learning computation with automatic generation of graph substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, SOSP '19, pages 47–62.
17. Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation.
18. Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi S. Jaakkola. Learning multimodal graph-to-graph translation for molecular optimization. *CoRR*, abs/1812.01070.
19. John J. Irwin and Brian K. Shoichet. ZINC: A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model*.
20. Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder.
21. Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, 15(6):e1007129.
22. Taibo Li, April Kim, Joseph Rosenbluh, Heiko Horn, Liraz Greenfeld, David An, Andrew Zimmer, Arthur Liberzon, Jon Bistline, Ted Natoli, et al. Genets: A unified web platform for network-based genomic analyses. *Nature Methods*, 15(7):543.
23. J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M.

Donovan, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. Kyung, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585.

24. Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J. Campos-Laborie, Benoit Charloteaux, Dongsic Choi, Atina G. Cote, Meaghan Daley, Steven Deimling, Alice Desbuleux, Amélie Dricot, Marinella Gebbia, Madeleine F. Hardy, Nishka Kishore, Jennifer J. Knapp, István A. Kovács, Irma Lemmens, Miles W. Mee, Joseph C. Mellor, Carl Pollis, Carles Pons, Aaron D. Richardson, Sadie Schlabach, Bridget Teeking, Anupama Yadav, Mariana Babor, Dawit Balcha, Omer Basha, Suet-Feung Chin, Soon Gang Choi, Claudia Colabella, Georges Coppin, Cassandra D'Amata, David De Ridder, Steffi De Rouck, Miquel Duran-Frigola, Hanane Ennajdaoui, Florian Goebels, Anjali Gopal, Ghazal Haddad, Mohamed Helmy, Yves Jacob, Yoseph Kassa, Roujia Li, Natascha van Lieshout, Andrew MacWilliams, Dylan Markey, Joseph N. Paulson, Sudharshan Rangarajan, John Rasla, Ashyad Rayhan, Thomas Rolland, Adriana San Miguel, Yun Shen, Dayag Sheykhkarimli, Gloria M. Sheynkman, Eyal Simonovsky, Murat Tasan, Alexander Tejeda, Jean-Claude Twizere, Yang Wang, Robert Weatheritt, Jochen Weile, Yu Xia, Xinping Yang, Esti Yeger-Lotem, Quan Zhong, Patrick Aloy, Gary D. Bader, Javier De Las Rivas, Suzanne Gaudet, Tong Hao, Janusz Rak, Jan Tavernier, Vincent Tropepe, David E. Hill, Marc Vidal, Frederick P. Roth, and Michael A. Calderwood. A reference map of the human protein interactome. *bioRxiv*.
25. Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J. Campos-Laborie, Benoit Charloteaux, Dongsic Choi, Atina G. Cote, Meaghan Daley, Steven Deimling, Alice Desbuleux, Amélie Dricot, Marinella Gebbia, Madeleine F. Hardy, Nishka Kishore, Jennifer J. Knapp, István A. Kovács, Irma Lemmens, Miles W. Mee, Joseph C. Mellor, Carl Pollis, Carles Pons, Aaron D. Richardson, Sadie Schlabach, Bridget Teeking, Anupama Yadav, Mariana Babor, Dawit Balcha, Omer Basha, Christian Bowman-Colin, Suet-Feung Chin, Soon Gang Choi, Claudia Colabella, Georges Coppin, Cassandra D'Amata, David De Ridder, Steffi De Rouck, Miquel Duran-Frigola, Hanane Ennajdaoui, Florian Goebels, Liana Goehring, Anjali Gopal, Ghazal Haddad, Elodie Hatchi, Mohamed Helmy, Yves Jacob, Yoseph Kassa, Serena Landini, Roujia Li, Natascha van Lieshout, Andrew MacWilliams, Dylan Markey, Joseph N. Paulson, Sudharshan Rangarajan, John Rasla, Ashyad Rayhan, Thomas Rolland, Adriana San-Miguel, Yun Shen, Dayag Sheykhkarimli, Gloria M. Sheynkman, Eyal Simonovsky, Murat Tasan, Alexander Tejeda, Jean-Claude Twizere, Yang Wang, Robert J. Weatheritt, Jochen Weile, Yu Xia, Xinping Yang, Esti Yeger-Lotem, Quan Zhong, Patrick Aloy, Gary D. Bader, Javier De Las Rivas, Suzanne Gaudet, Tong Hao, Janusz Rak, Jan

Tavernier, Vincent Tropepe, David E. Hill, Marc Vidal, Frederick P. Roth, and Michael A. Calderwood. A reference map of the human protein interactome. *bioRxiv*.

26. Craig O Mackenzie, Jianfu Zhou, and Gevorg Grigoryan. Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences*, 113(47):E7438–E7447.
27. Olivecrona, M., Blaschke, T., Engkvist, O. et al. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*.
28. Nidhi Sahni, Song Yi, Mikko Taipale, Juan I. Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I. Karras, Yang Wang, István A. Kovács, Atanas Kamburov, Irina Krykbaeva, Mandy H. Lam, George Tucker, Vikram Khurana, Amitabh Sharma, Yang-Yu Liu, Nozomu Yachie, Quan Zhong, Yun Shen, Alexandre Palagi, Adriana San-Miguel, Changyu Fan, Dawit Balcha, Amelie Dricot, Daniel M. Jordan, Jennifer M. Walsh, Akash A. Shah, Xinping Yang, Ani K. Stoyanova, Alex Leighton, Michael A. Calderwood, Yves Jacob, Michael E. Cusick, Kourosh Salehi-Ashtiani, Luke J. Whitesell, Shamil Sunyaev, Bonnie Berger, Albert-László Barabási, Benoit Charloteaux, David E. Hill, Tong Hao, Frederick P. Roth, Yu Xia, Albertha J.M. Walhout, Susan Lindquist, and Marc Vidal. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660.
29. Yasuo Tabei, Masaaki Kotera, Ryusuke Sawada, and Yoshihiro Yamanishi. Network-based characterization of drug-protein interaction signatures with a space-efficient approach. *BMC Systems Biology*, 13(2):39.
30. Haidong Wang, Eran Segal, Asa Ben-Hur, Qian-Ru Li, Marc Vidal, and Daphne Koller. Insite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biology*, 8(9):R192.
31. Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388.
32. Yang, Kevin, Swanson, Kyle, Jin, Wengong, Coley, Connor, Eiden, Philipp, Gao, Hua, Guzman-Perez, Angel, Hopper, Timothy, Kelley, Brian, Mathea, Miriam, Palmer, Andrew, Settels, Volker, Jaakkola, Tommi, Jensen, Klavs, and Barzilay, Regina. Molecular Property Prediction. Github.
33. Hongyi Zhou, Mu Gao, and Jeffrey Skolnick. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Scientific Reports*, 5:11090.