Neighborhood Embedding and Re-Ranking of Disease Genes with ADAGIO

Mert Erden Megan Gelement Sarrah Hakimjee Kyla Levin
Mary-Joy Sidhom Kapil Devkota Lenore J. Cowen[*]
Department of Computer Science
Tufts University
177 College Ave
Medford, MA 02155
USA

ABSTRACT

We present ADAGIO, a new method for network-based disease gene prioritization that balances network interconnection structure with an embedding measure of network similarity. We show ADAGIO performs better than previous methods for recovering known disease genes in a recent benchmark set encompassing disease-associated genes for 22 polygenic diseases. We find ADAGIO discovers some interesting new disease gene candidates in both Alzheimer's and Parkinson's diseases.
Code, ranked lists of disease genes, and supplementary figures and tables appear at https://github.com/merterden98/ADAGIO.

ACM Reference Format:
Mert Erden Megan Gelement Sarrah Hakimjee Kyla Levin , Mary-Joy Sidhom Kapil Devkota Lenore J. Cowen[*]. 2022. Neighborhood Embedding and Re-Ranking of Disease Genes with ADAGIO. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3535508.3545542

1 INTRODUCTION

Network-based disease gene prioritization is one of the best studied classical inference problems on biological networks [7]. Given a protein-protein association network, a list of known disease genes, and a list of candidate disease genes (either from a linkage interval or genome-wide), the output is a ranked list of the candidate genes. Based on the underlying graph structure of the association network, predicted genes are ranked in order of how strongly they are related to the set of known disease genes. While many computational methods have been proposed to leverage network information to create these ranked lists [9, 14, 18, 25], until recently, it was difficult to come up with a fair and unbiased way to measure performance of methods. Fortunately, recently, several high-quality benchmarks have been constructed for exactly this purpose [19, 27]. In this work, we concentrate on the benchmark sets from [19], constructed using the Open Targets gene lists [10].

This benchmark curates two types of sets of disease genes for 22 different polygenic diseases. In practice, when looking for disease genes in a chromosomal region, the set of candidate disease genes is restricted to only a subset of the genes. In a benchmarking context, however, it is more typical as in [19] to compare whole genome ranked lists. In addition to the global measures of AUROC and AUPRC, disease-gene prioritization ranked lists are compared using measures that look only at what is placed in the top k portion of the list (see Section 3.3.1).

Recently, for the different but related problem of link prediction, we introduced GLIDE [8], which considers gene similarity using different methods for genes in the highly connected core than in the periphery of the network. GLIDE combines a simple score based on common neighbors in the dense core, with a diffusion-based embedding that encapsulates the network structure in the periphery. For the disease-gene prioritization problem, we find that throwing out the network and ranking disease genes in order of their GLIDE scores is not competitive with even the simplest RWR (Random Walk with Restart) method from Kohler's original paper [14]. However, we find that retaining the network but reweighting edges according to their GLIDE scores and then running RWR on the reweighted network produces near state-of-the art performance on the benchmark of [19]. Finally, we improve performance further by augmenting the network with new high-confidence edges predicted by GLIDE We call our new method ADAGIO (for Augmented Disease Associated GLIDE Index Order), and describe it more completely in Section 2.1.

In addition to measuring ADAGIO's global performance on the entire benchmark set of 22 diseases, we propose a new framework to flag high-ranking genes that competing algorithms fail to identify. We examine several of the brain-related diseases in our benchmark, and look to the literature for evidence of disease involvement for the novel genes found by ADAGIO and not by other methods. We suggest several new genes for involvement in Alzheimer's and Parkinson's Disease.

ADAGIO is available at https://github.com/merterden98/ADAGIO

2 METHODS

2.1 ADAGIO

An overview of ADAGIO appears in Figure 1. ADAGIO first transforms the underlying network using a variant of the GLIDE [8] similarity score, which we describe here next. On this new augmented network, standard random walk with restart (RWR; see Section 2.2.1) is run using a list of known candidate genes. The

[*] To whom correspondence should be addressed: lenore.cowen@tufts.edu

BCB '22, August 7–10, 2022, Northbrook, IL, USA

ACM ISBN 978-1-4503-9386-7/22/08.
https://doi.org/10.1145/3535508.3545542

output, as in all disease gene prioritization algorithms, is a ranked list of genes that are most similar to the candidate genes.

GLIDE combines a simple local score that captures relationships in the dense core with a diffusion-based embedding that encapsulates the network structure in the periphery. For ADAGIO, we pair a local score based on common neighbors with global score UDSED$^\gamma$, a variant of DSED$^\gamma$ from the original GLIDE paper [8].

Definition 2.1. DSE$^\gamma$ Embedding (from [8]) Let P be a Markov transition matrix computed from a graph G with a unique stationary distribution π and let D be the diagonal degree matrix representing the weighted degree of all the nodes in the network. Then the DSE$^\gamma$ embedding is:

$$DSE^\gamma = I + \sum_{t=1}^{\infty} \gamma^t (P - W)^t, \tag{1}$$

where W is a constant matrix whose rows are copies of the stationary distribution π, and γ is a parameter satisfying $0 < \gamma \leq 1$, which is used to control the contribution of larger time-steps in the computation of the embedding. We set $\gamma = 1$ in all our experiments, as suggested in [8].

Definition 2.2. Global Score: UDSED$^\gamma$ Distance.

If DSE$^\gamma(p)$ and DSE$^\gamma(q)$ represent the DSE$^\gamma$ embeddings for the nodes p and q respectively, we consider the (un-normalized) L2 distance between their DSE$^\gamma$ embeddings. Formally, this can be written as

$$\textbf{UDSED}^\gamma(p,q)\& = \sqrt{\sum_k (\text{DSE}^\gamma(p)_k - \text{DSE}^\gamma(q)_k)^2}$$

Definition 2.3. Local Score: Common Weighted Normalized

Given nodes $p, q \in G$, the Common Weighted Normalized (CWN) score is

$$\text{CWN}(p,q) = \frac{\sum_{r \in \mathcal{N}_p \cap \mathcal{N}_q} (w_{p,r} + w_{q,r})}{\sqrt{k(p)k(q)}}$$

where for any node $x \in G$, $\mathcal{N}_x$ is the neighbor set of x, $w_{x,y}$ is the weight of the edge (x,y), and $k(x)$ represents the weighted degree of x. Note that this is slightly different from the CW metric described in [8], because of the square roots in the denominator.

GLIDE score. Just as in [8] we define the following score between each pair of nodes:

$$\begin{aligned} GLIDE(p,q) = \exp\left(\frac{\alpha \cdot global(x_i, x_j)}{global(x_i, x_j) + \beta}\right) local(x_i, x_j) \\ + global(x_i, x_j), \end{aligned}$$

where GLIDE chooses $local(p,q) = CWN(p,q)$ and $global(p,q) = 1/\text{UDSED}^\gamma(p,q)$. We choose the default values of α and β as suggested by [8] ($\alpha = 0.1, \beta = 1000$), where these choices for α and β makes the local embedding dominant for ranking, while the global embedding is used to break ties and order nodes with the same strong local score. For the CWN local score, if nodes have no common neighbors, the first term is 0 and only the global score is used. It was shown in [8] that a variant of GLIDE improved link prediction algorithms.

2.1.1 Glide Based Reweighting. We re-weight the original PPI network by substituting the existing edge weights with their computed GLIDE scores. For example, consider a graph $G = (V, E, W)$. Then the new weight, or $W'(p,q)$, for the edge (p,q) of the re-weighted graph $G' = (V, E, W')$ is $W'(p,q) = GLIDE(p,q)$.

The prime intuition behind this network re-weighting operation is that the distribution of edge-weights in a PPI network is often impervious to the presence of high degree hub nodes, present in the network's dense core. Randomly walking on this degree-agnostic network could then result in a distribution where hub proteins dominate over every other node, limiting the discovery of novel and more surprising associations facilitated by the peripheral nodes in the network. Our assumption is that the use of GLIDE scores can restrain this excessive influence of hubs, while making use of the powerful GLIDE global embedding component to find the interesting parts of the network that are not dominated by hubs in the dense core. In our experiments, below, we also compare GLIDE to another method that explicitly seeks to control the influence of hub nodes: DADA. [9].

2.1.2 GLIDE-Based Edge Addition. The addition of new edges uses GLIDE scores to find the most likely links for each node in the network. For each node p, we select the top k highest scoring new pairs (containing p as one of the endpoints) and add it to the original network. We experimented with different choices of k and settled on setting k equal to half the average network degree. (i.e. k is set to $\lfloor \frac{|E|}{|V|} \rfloor$). Exploration of performance for different values of k appear in the supplement.

2.2 Existing Methods

We test ADAGIO against the 5 best performing methods tested in the benchmark paper (see [19]). These are Kohler's RWR [14] (which they call Personalized Pagerank), EGAD [4], a GeneMANIA [18] based disease gene prioritization method, a Monte Carlo method, and an SVM method. These methods are all implemented exactly as in [21]. In addition, we tested the popular DADA algorithm from [9], since, like ADAGIO, it is an RWR-related method that seeks to minimize contribution of hub nodes. A brief description of all these methods appears next.

2.2.1 Random Walk With Restart. Random Walk With Restart (RWR) or Personalized Pagerank (PP) is a widely-used diffusion-based approach, popular in other computational settings (social networks, web search, etc). This method uses global network diffusion properties, realized through the usage of lazy random walks, to access/rank the similarities between nodes. Given a set of "starting nodes", Pagerank computes the steady-state distribution of the lazy random walks originating from the starting nodes, and uses this distribution to rank the relatedness of the remaining nodes.

Let $\boldsymbol{W}$ be the Markov Transition matrix of a connected graph $G = (V, E, W)$ and let $S \subset V$ be the set of starting nodes. Since the starting probabilities (denoted by $\boldsymbol{p_0}$) are uniformly distributed among the elements of S, it can be written as $\boldsymbol{p_0} = \mathbf{1}_S/|S|$, provided $\mathbf{1}_S$ is a vector which has value 1 for elements in S and 0 otherwise. Then, the distribution $\boldsymbol{p_t}$ at timestep t can be described as

$$\boldsymbol{p_t} = \frac{17}{20}\boldsymbol{W}\boldsymbol{p_{t-1}} + \frac{3}{20}\boldsymbol{p_0}$$

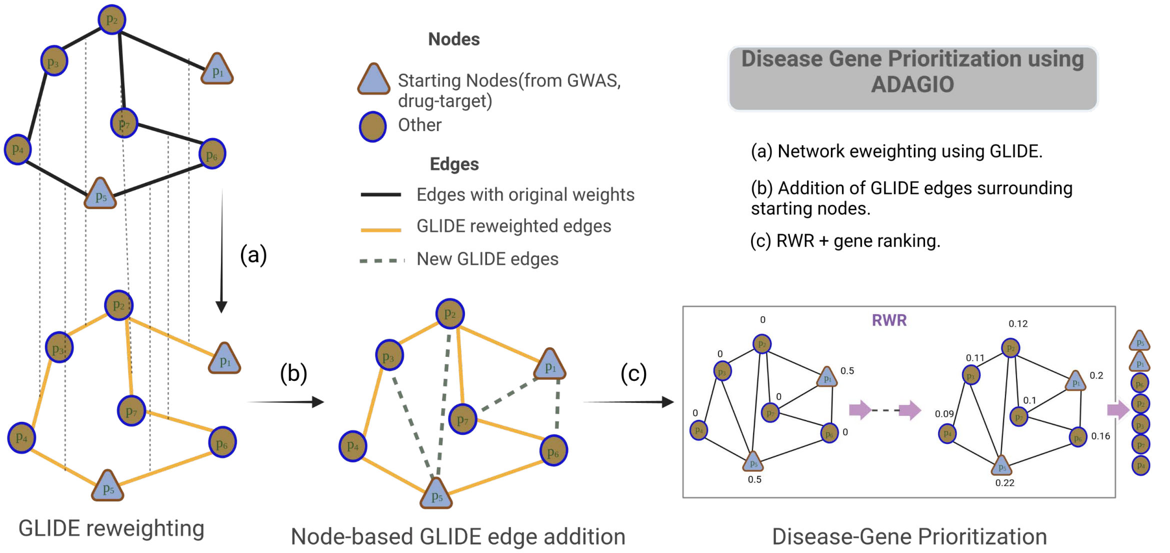

Nodes
Starting Nodes(from GWAS, drug-target)
Other
Edges
Edges with original weights
GLIDE reweighted edges
New GLIDE edges
Disease Gene Prioritization using ADAGIO
(a) Network eweighting using GLIDE.
(b) Addition of GLIDE edges surrounding starting nodes.
(c) RWR + gene ranking.
(a)
(b)
(c)
RWR
GLIDE reweighting
Node-based GLIDE edge addition
Disease-Gene Prioritization

Figure 1: Overview of Network Augmentation and Re-weighting with ADAGIO. (a) Original edge weights are replaced with GLIDE weights. (b) For each disease gene, the network is augmented with the k missing heaviest GLIDE-scoring edges with that endpoint, where k is the average degree of the network. (c) Kohler's classical Random Walk with Restart is run on the modified network to produce the ranked gene list.

where $\frac{3}{20}$ represents the probability. In theory this is a changeable parameter; however, to stay consistent with the benchmark [19] we fix it to be $\frac{3}{20}$. At equilibrium, the steady-state value $\boldsymbol{p}$ can be computed, but is often much slower to compute compared to an iterative solution due to matrix inversion. This $\boldsymbol{p}$ is very useful in accessing the degree of association between the starting nodes in S and the rest of the network. In fact, [2] showed that the ordering of vertices obtained from the steady-state Pagerank vector produced localized clusters with good intra-cluster association. This, equipped with a sufficiently strong restart probability (which, in our instance is 0.15), would then result in the ordering of nodes where the top nodes have a high degree of relatedness with the starting nodes. In ADAGIO the matrix $\boldsymbol{W}$ corresponds to the reweighted network with a new set of links as described in Section 2.1.2.

2.2.2 *Degree-Aware Disease Gene Prioritization Algorithm (DADA).*

The skewed degree distribution of PPI networks can make the RWR-based prioritization algorithms assign disproportionate importance to high-degree hub genes. DADA [9] addresses this hub effect by employing a suite of statistical adjustment strategies to detect loosely connected disease genes that are missed by the existing RWR-based approaches.

This suite is composed of three reference models, each of which can be independently applied to the results obtained from RWR. The degree-adjusted RWR vector, after the application of these models, can then be used for the discovery of new disease gene associations.

2.2.3 *Diffusion with GeneMANIA-based weights* (gm).

The label propagation mechanism in GeneMANIA, originally used for predicting gene function [18], can also be employed for predicting new disease genes. The reworked version of GeneMANIA is available in the diffusStats [21] package.

This new adaptation uses the graph Laplacian as the coefficient matrix for the propagation of disease genes. Additionally, the vector of node-level biases, namely $\boldsymbol{y_{gm}}$, which incorporates the strength of association between the disease and genes, is set in the following way

$$\boldsymbol{y_{gm}}(n) = \begin{cases} -1 & , n \text{ is a known negative} \\ 1 & , n \text{ is a known positive} \\ \frac{N^+ - N^-}{N^+ + N^- + N^u} & , n \text{ has unknown association} \end{cases} \tag{2}$$

where N^+, N^-, and N^u denote the numbers of known disease genes, known negatives, and genes with unknown association, respectively.

The output disease associations $\boldsymbol{f}$ is now obtained by linearly transforming $\boldsymbol{y_{gm}}$ using the kernel matrix $\boldsymbol{K}$. Mathematically, this can be represented as $\boldsymbol{f} = \boldsymbol{K}\, \boldsymbol{y_{gm}}$.

So, the gm method uses this transformed vector output $\boldsymbol{f}$ to compare and rank the disease associations of genes.

2.2.4 *Extending 'Guilt-by-Association' by Degree (EGAD).*

EGAD [4], originally developed for predicting functional relationships, ranks disease-gene associations based on the assumption that genes with common functions preferentially have a localized network relationship. Given a node with unknown disease association, EGAD uses a neighborhood voting algorithm to calculate the ratio of neighbors with known disease association. This ratio is then used for comparing disease associations between genes.

2.2.5 *Monte Carlo Normalized Scores.*

Monte Carlo normalized scores (mc) is a diffusion-based technique that uses statistical normalization to assign scores to genes. As described in [20], input genes are permuted multiple times to observe the number of times a randomly permuted set of disease genes leads to higher diffusion scores compared to just using a kernel matrix alongside an indicator vector of disease genes.

2.2.6 SVM. This method, described in [19], uses the results from a SVM binary classifier for gene prioritization. The gene embeddings supplied to the classifier are derived from the graph Laplacian, where each row of the matrix functions as the feature vector for a particular gene. A portion of the disease genes are separated out as positive examples for training, and the negative examples are randomly selected from the network. The resulting classifier output is then used to rank the remaining genes.

3 EXPERIMENTAL SETUP

3.1 Networks

We use exactly the same STRING-derived PPI network as in the benchmarks in Picart et al. [19] (based on STRING 10, where the edges for STRING are filtered by removing any edge that has a combined score less than 600, and that is not labeled as an "experimental" or "database" association.) The characteristics of the benchmarked network are described below:

Graph Properties	Values
# Nodes	11748
# Edges	236963
Average Degree	40.34
Average Clustering Coeff.	0.51

Table 1: Graph Properties of STRING-E

3.2 Disease Gene Lists

The benchmark of [19] provides 22 diseases, listed in Tables 6 and 7. For each disease, the benchmark provides two different sets of disease genes, one set derived using Genome-Wide Association Studies (GWAS) and the second from evidence provided from clinical drug studies. Both the GWAS and the drug-target associations for all the diseases are aggregated from Open Targets [10].

3.3 Evaluation

As recommended in the benchmark, the evaluation of all prioritization algorithms was done through 3-fold cross-validation, where the benchmark randomly split the gene-list (which is either GWAS or drug-based) into three equal-sized blocks and in each CV iteration, two blocks (i.e. ~ 66.6% of the dataset) were used for training and the remaining block (~ 33.3% of the dataset) for testing (resulting in 3 CV runs). We report both the mean and the standard deviation from the CV runs, and use the Area Under the Precision Recall Curve (AUPRC), Area Under the Receiving Operating Characteristic (AUROC) and truncated AUROC (t-AUROC). We also introduce a modified variant of Picart et al.'s [19] Top-K metric where we normalize the score by the size of the testing set. The mathematical formulation of Top-K and the Witness Analysis is provided below.

3.3.1 Top-K. Given a parameter k, the Top-K metric is proportional to the number of true-positive (tp) genes the gene-prioritization algorithm was successfully able to recall in the first k positions of its predicted ranked list. Let $tp(k)$ represent the number of true positives at the position k, and T represent the total number of true positives. Then the Top-K score for the prediction is given as

$$\text{Top-K}(k) = \frac{tp(k)}{T} \tag{3}$$

Top-K evaluates the strength of the prioritization algorithm by only observing the top section of its predicted gene-list, making it more focused and sensitive than methods like AUROC and AUPRC, which use the complete gene-list for evaluation.

We used two values of k ($k = 100, 250$) for our evaluations, and report both the obtained means and the standard deviations from the CV runs. Tables 6 and 7 report ADAGIO and competitor algorithms' under the Top-K metric where $k = 100$, which we assume is the most informative and practically relevant amongst all the metrics measured. In the supplement we also report this measurement for $k = 250$. Comparative results using the other scoring metrics show similar trends and can be found in the supplement.

We describe the other evaluation metrics next:

3.3.2 AUPRC. Let $\text{RECALL}(k)$ be the ratio of the number of true positives between positions 1 through k (i.e. $tp(k)$) to the total positives P:

$$\text{RECALL}(k) = \frac{tp(k)}{P} \tag{4}$$

Also, define the precision at position k as $\text{PRECISION}(k) = \frac{tp(k)}{k}$. Then we can construct a graph by tracing the curve $c(k) = (\text{RECALL}(k)$ $, \text{PRECISION}(k))$ for each position k. The Area Under the Precision Recall Curve (AUPRC) is then the area of $c(t)$.

3.3.3 AUROC. Describe the false positive rate $\text{FP}(k)$ as the ratio of all false positives between positions 1 through k to P:

$$\text{FP}(k) = \frac{k - tp(k)}{P} \tag{5}$$

Then, as in the precision curve described in Section 3.3.2, we can create a curve $c'(k) = (\text{FP}(k), \text{RECALL}(k))$ by varying k, which we call the Receiving Operating Characteristics (ROC) curve. We now calculate the AUROC score by simply measuring the area of this ROC curve.

3.3.4 tAUROC. Let V be the total number of genes. If we only want to look at the performance of prioritization algorithms up to a certain fraction t, we can truncate the ROC curve by varying k from 1 to tV. The corresponding area of this truncated ROC curve is called tAUROC. In the Picart et al. benchmark, the value of t was set to 0.05 and 0.1.

4 RESULTS

4.1 ADAGIO Outperforms Benchmarked Methods or is Competitive

Considering the recovery of known disease genes in the top 100 ranked genes (according to the TopK measure of Section 3.3.1), ADAGIO performs the best across both the drug- and GWAS-based gene lists. On drug-based gene lists, ADAGIO outperforms other

ADAGIO	EGAD	RWR	SVM	Degree
1. UBC	2077	2	2	4364
2. GABARAP	65	3	52	35
3. GPHN	30	4	36	23
4. PLCL1	21	9	72	14
5. TRAK2	23	8	69	15
6. GABARAPL2	63	6	66	33
7. NSF	68	5	34	41
8. GABARAPL1	64	7	73	34
9. CHRNA3	57	13	303	12
10. CHRNB4	55	12	156	12
11. CHRNB3	52	20	1080	11
12. CHRNA2	50	14	226	11
13. CHRND	62	19	376	12
14. CHRNB1	88	15	225	18
15. SCN4B	106	33	209	19

Table 2: Top 15 ADAGIO predictions For Alzheimer’s Disease on the Open Targets drugs gene list alongside prediction ranks for the same genes using EGAD, RWR and SVM, alongside the degree of the gene within the benchmark network.

ADAGIO	EGAD	RWR	SVM	Degree
1. UBC	2511	2	8647	4364
2. CTNNB1	1359	3	8644	359
3. ACSF3	6966	10487	222	1
4. TP53	1890	4	8215	538
5. ARMT1	8829	10488	223	1
6. CACUL1	9022	10485	212	1
7. OTUD3	9538	10486	253	1
8. DDI2	10950	9558	760	1
9. ANKRD13A	4531	9567	762	1
10. RAC1	1973	6	8679	392
11. GFM2	5871	10483	231	1
12. TOR4A	8393	10484	259	1
13. ORMDL2	3860	10481	246	1
14. DNPEP	5184	10482	257	1
15. ATXN10	4117	9577	761	1

Table 3: Top 15 ADAGIO predictions For Alzheimer’s Disease on the Open Targets GWAS gene list alongside prediction ranks for the same genes using EGAD, RWR and SVM, alongside the degree of the gene within the benchmark network.

ADAGIO	EGAD	RWR	SVM	Degree
1. KCNAB1	19	4	19	42
2. KCNAB3	18	5	4	41
3. KCNAB2	21	3	101	46
4. UBC	1436	2	917	4364
5. GPHN	27	7	282	23
6. PLCL1	16	12	2	14
7. GABARAP	50	6	144	35
8. TRAK2	20	11	10	15
9. NSF	54	8	365	41
10. GABARAPL2	52	9	202	33
11. KCNE4	23	513	77	4
12. DPP10	14	683	11	2
13. KCNIP2	24	543	30	3
14. GABARAPL1	48	10	163	34
15. KCNE2	6	753	150	2

Table 4: Top 15 ADAGIO predictions For Parkinson’s Disease on the Open Targets drugs gene list alongside prediction ranks for the same genes using EGAD, RWR and SVM, alongside the degree of the gene within the benchmark network.

methods for 14 of the 22 diseases (Table 6). The next best method, SVM, performs best on 6 of the 22 diseases. On GWAS-based gene lists, ADAGIO performs the best on 9 of the 22 diseases, while the next best method, RWR, performs best on 8 of the 22 diseases (Table 7). When we consider the top 250 ranked genes, ADAGIO’s competitive performance becomes even better on the drug lists, but decreases on the GWAS lists. On the drug lists, ADAGIO becomes the best method for 18 of the 22 diseases (see Supplement). However, on the GWAS gene lists, RWR outperforms ADAGIO (see Supplement). In general, we find that ADAGIO performs better than competing methods on the top 1% of gene lists, and as we consider a larger and larger subset of the top-ranked genes to be disease genes, RWR starts to match (or exceed) ADAGIO’s performance. This is reflected in the comparative performance of the global AUROC and AUPRC, where RWR and ADAGIO show nearly identical performance, demonstrating again that the advantage of ADAGIO is at the top of the list. SVM also sometimes does well in AUROC and AUPRC, but is never competitive in TopK measures (see Supplement for full results on AUROC, AUPRC and tAUROC). An enriched top-of-rank list may be useful to biologists with limited experimental bandwidth.

We present the top 15 ranked ADAGIO predictions for Alzheimer’s and Parkinson’s diseases (removing the known disease gene sets) in Tables 2, 3, 4, and 5, along with their ranks using the other methods. The top 50 ADAGIO predictions for all 22 diseases can be found in the Supplement.

4.2 ADAGIO is Robust to Network Noise

ADAGIO introduces the top-scoring GLIDE links centered around seed disease genes. The main argument for the addition of these edges is to compensate for existing noise from the generation of the network. Combined with the GLIDE-based edge re-weighting, new links should be able to reach more functionally enriched neighborhoods as opposed to being dispersed into hub genes or areas of the network that are not captured by the original network’s edges. To test this hypothesis, we randomly introduced noise to the network. Specifically, the original benchmark network was augmented by removing 20% and 40% of the existing edges at random. Additionally, we ran experiments by adding 20% and 40% new edges at random. These modifications were done in order to simulate a noisy network. We re-ran experiments on these new networks, comparing

ADAGIO	EGAD	RWR	SVM	Degree
1. UBC	1018	2	10190	4364
2. ACSF3	6360	10274	642	1
3. ARMT1	8458	10275	643	1
4. CACUL1	8687	10270	648	1
5. OTUD3	9338	10271	614	1
6. GFM2	5099	10267	629	1
7. TOR4A	7921	10268	607	1
8. ORMDL2	2604	10265	615	1
9. DNPEP	4275	10266	609	1
10. ATP8B3	5923	10262	630	1
11. DALRD3	7564	10263	616	1
12. TARS3	7165	10260	645	1
13. ESR1	971	8	9480	274
14. HSPA8	312	9	9538	199
15. DDI2	11028	9321	117	1

Table 5: Top 15 ADAGIO predictions For Parkinson's Disease on the Open Targets GWAS gene list alongside prediction ranks for the same genes using EGAD, RWR and SVM, alongside the degree of the gene within the benchmark network.

ADAGIO with the top performing competitors: RWR, SVM and DADA. Figures 2a, 2b, 2c, and 2d show the performance of these methods at varying levels of noise. In the case where new edges are added at random to the network (2c, 2a), ADAGIO is uniquely able to maintain its performance over other methods, whereas we find that all methods continue to perform well when we remove edges at random (note that for RWR in 2d, removing 40% of edges at random actually improves performance).

4.3 All methods are better at recovering the drug-based disease gene lists than the GWAS gene lists

The performance of all the prioritization algorithms varies significantly between drug based gene lists and GWAS gene lists. In particular we see that across every disease, correctly recovering GWAS disease genes is much harder (see Tables 6 and 7). The only disease where prioritization algorithms' performance is on par with that of the drug based gene list is Parkinson's disease.

We do not know why GWAS genes seem much harder to recover using any prioritization algorithm. One reason might be that drugs targeted by clinical studies are often associated with well-understood disease pathways, making their organization in the underlying network more localized. In comparison, the GWAS genes may be distributed more uniformly across the network.

4.4 Finding new Parkinson's and Alzheimer's genes using ADAGIO

We chose two of the brain diseases, Parkinson's and Alzheimer's and performed a deeper analysis of the top scoring genes. Table 8 shows the top 15 ADAGIO predictions for Alzheimer's and Parkinson's disease, using both the Open Targets drug and GWAS gene lists. We also list the network degrees of the predicted genes, and their corresponding placement in the ranked list using competing algorithms (for this, we used EGAD, RWR and SVM).

The tables show that the ADAGIO predictions are mostly composed of low-degree genes in Parkinson's diseases, for both GWAS and drug gene lists. The same is not true for Alzheimer's, where we notice a significant number of top predictions are genes of degree > 10. This is likely due to the common weighted neighbor portion of GLIDE introducing edges to more hub genes.

For Alzheimer's disease, we noticed that the ADAGIO predictions are often also ranked highly by the competing algorithms. This is more evident in Table 8a (Alzheimer's with drug gene list), where we see all of the top 8 ADAGIO predictions occupying positions 2 to 9 in RWR rankings. In order to summarize the difference between ADAGIO and competing methods, and also to have a rigorous definition of novel genes found by our method and not others, we introduce a new measure called the "Witness Measure", which we describe next.

4.5 Witness Measure

Define M as the ranked list of preferred genes predicted by ADAGIO (i.e. $M(i)$ represents the gene placed by ADAGIO at position i). For the k competing algorithms, let $R_1, \ldots, R_k$ be the mapping that associates each gene with its corresponding positioning (i.e. $R_i(g)$ denotes the rank of the gene g for the competing method i). Then, we regard a gene g at position p in M (or $g = M(p)$) to be "witnessed" or "unique" if, for any competing algorithm R_k, the following inequality is satisfied

$$R_k(g) \geq ap + b \tag{6}$$

where in this paper we set $a = 2$ and $b = 10$ as defaults. If the expression in (6) is satisfied for all the competing algorithms for a given gene g, this indicates that ADAGIO is uniquely ranking g with higher preference compared to other methods, making the placement of g more surprising. Define $I(g)$ as 1 if g satisfies (6) and 0 otherwise.

Now we define the witness measure $W(k)$ for a position k, as the ratio

$$W(k) = \frac{\sum_{i=1}^{k} I(M(i))}{k} \tag{7}$$

$W(k)$ finds the proportion of the "witnessed" genes ranked at positions 1 to k by ADAGIO. Its value being close to 1 implies that the significant proportion of the top ranked predictions are unique to ADAGIO. The graphical description of the witness measure is shown in Figure 4.

Figure 3 shows the variations in $W(k)$ for Parkinson's and Alzheimer's diseases, for both GWAS and drug-based gene lists. A more detailed breakdown of the witness analysis results are provided below.

4.6 Witness Analysis Case Study

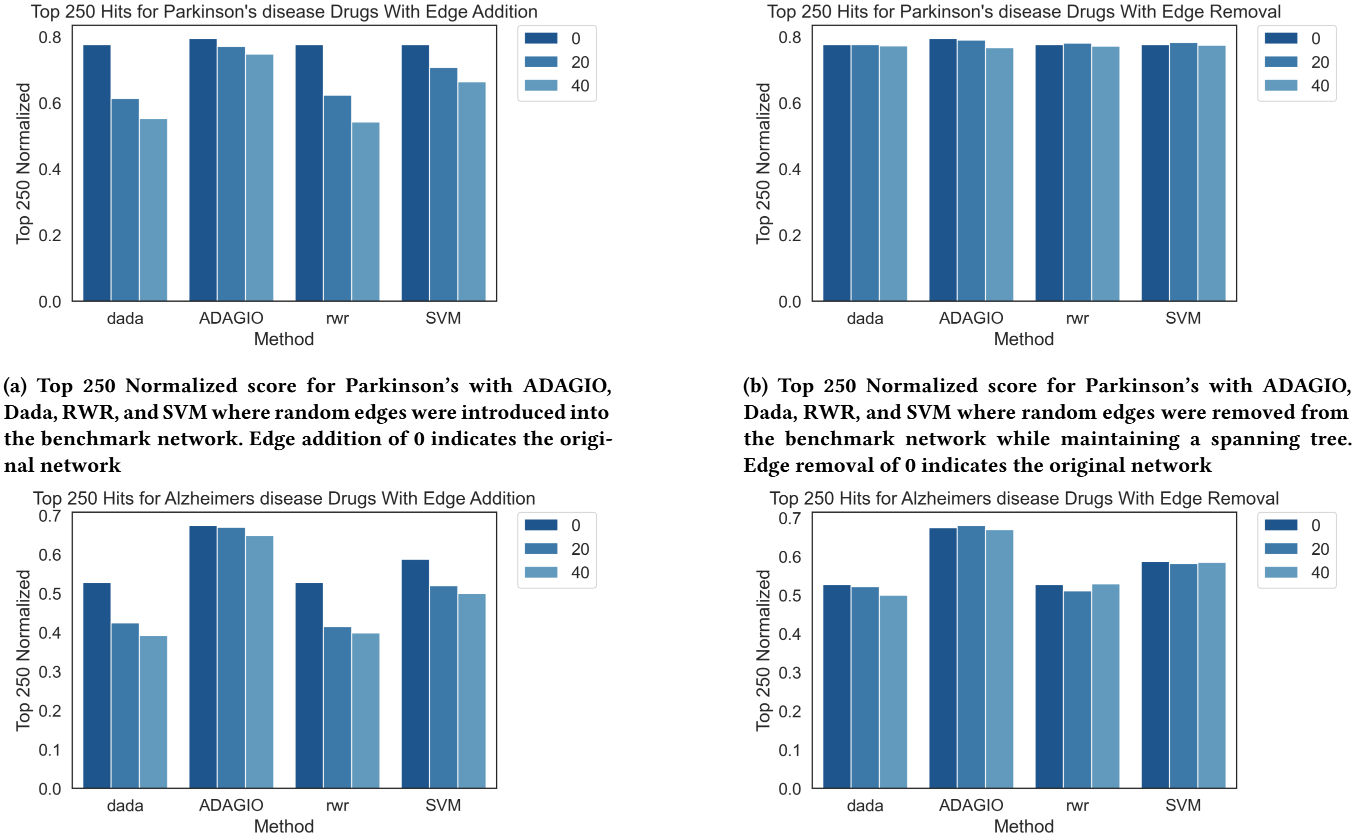
The Witness Measure, described in Section 4.5, devises a principled way of finding high-confidence ADAGIO genes that are not equivalently predicted with high certainty by competing methods. The graphs in Figures 3a and 3b, which plot the variation in $W(k)$ with position k for Alzheimer's, indicate that approximately 10% of the genes produced by ADAGIO on the drug-based gene list are considered novel by our Witness Measure. The plot for the GWAS

	Top 100							
	ADAGIO	SVM	EGAD	DADA	gm	mc	RWR	GLIDE Reweighted
allergy	**0.27±0.09**	0.23±0.1	0.14±0.08	0.22±0.06	0±0	0.23±0.08	0.21±0.07	0.24±0.09
Alzheimer's disease	**0.35±0.09**	0.33±0.11	0.13±0.06	0.34±0.1	0.04±0.03	0.2±0.07	0.29±0.08	0.32±0.09
arthritis	**0.3±0.06**	0.29±0.05	0.06±0.02	0.26±0.05	0.01±0.01	0.16±0.03	0.17±0.06	0.21±0.06
asthma	0.38±0.06	0.33±0.07	0.11±0.05	0.22±0.07	0.01±0.02	0.24±0.08	**0.41±0.07**	0.39±0.07
bipolar disorder	**0.52±0.09**	0.49±0.07	0.14±0.05	0.49±0.06	0.03±0.03	0.28±0.07	0.38±0.12	0.42±0.11
cardiac arrhythmia	**0.61±0.06**	0.6±0.07	0.56±0.07	0.59±0.07	0.45±0.05	0.44±0.08	0.58±0.06	**0.61±0.05**
chronic obstructive pulmonary disease	0.29±0.08	0.25±0.06	0.13±0.07	0.25±0.06	0.03±0.02	0.31±0.07	0.31±0.06	**0.33±0.08**
coronary heart disease	0.42±0.08	**0.43±0.07**	0.15±0.04	0.4±0.07	0±0.01	0.25±0.06	0.32±0.07	0.35±0.06
drug dependence	**0.35±0.07**	0.34±0.05	0.19±0.06	0.33±0.04	0.04±0.03	0.31±0.09	0.32±0.05	0.32±0.05
hypertension	0.22±0.07	**0.28±0.06**	0.09±0.04	0.25±0.07	0.02±0.02	0.18±0.05	0.18±0.06	0.22±0.07
multiple sclerosis	**0.53±0.07**	0.48±0.07	0.39±0.06	0.46±0.07	0.36±0.06	0.33±0.06	0.5±0.06	0.51±0.07
obesity	0.41±0.11	**0.43±0.06**	0.16±0.04	0.41±0.06	0.02±0.02	0.26±0.07	0.25±0.08	0.29±0.08
Parkinson's disease	**0.53±0.06**	0.51±0.06	0.43±0.06	0.52±0.06	0.34±0.04	0.37±0.08	0.5±0.06	0.52±0.05
psoriasis	**0.32±0.08**	0.24±0.06	0.14±0.05	0.25±0.05	0±0.01	0.15±0.05	0.31±0.09	**0.32±0.09**
rheumatoid arthritis	**0.39±0.08**	0.32±0.06	0.05±0.04	0.31±0.06	0±0	0.2±0.05	0.36±0.08	0.37±0.07
schizophrenia	**0.43±0.07**	**0.43±0.08**	0.36±0.09	0.33±0.1	0.04±0.03	0.31±0.08	0.37±0.07	0.4±0.08
stroke	**0.57±0.05**	0.52±0.1	0.46±0.1	0.52±0.1	0.4±0.09	0.37±0.07	0.54±0.07	0.53±0.07
systemic lupus erythematosus	**0.27±0.14**	**0.27±0.09**	0.13±0.05	0.27±0.09	0±0	0.16±0.1	0.22±0.11	0.26±0.13
type I diabetes mellitus	0.21±0.07	0.08±0.05	0.07±0.05	0.1±0.06	0±0	0.18±0.06	**0.23±0.07**	0.22±0.09
type II diabetes mellitus	0.4±0.09	**0.41±0.09**	0.14±0.05	0.38±0.08	0.02±0.02	0.26±0.07	0.28±0.08	0.31±0.08
ulcerative colitis	**0.4±0.14**	0.31±0.1	0.23±0.09	0.31±0.08	0±0	0.28±0.13	0.32±0.1	0.36±0.13
unipolar depression	0.33±0.06	0.32±0.07	0.11±0.04	0.23±0.05	0.04±0.03	0.28±0.07	**0.34±0.07**	0.33±0.07

Table 6: Benchmarking results on 22 Open Targets gene lists originated from drug studies. The left half of the table are the results for the Top 100 Normalized score. Bolded cells indicate the best performing method for that scoring metric. ADAGIO outperforms in 15 of the 22 disease gene lists for the Top 100 Metric.

	Top 100							
	ADAGIO	SVM	EGAD	DADA	gm	mc	RWR	GLIDE Reweighted
allergy	**0.09±0.06**	0.04±0.04	0.03±0.04	0.03±0.03	0±0	**0.09±0.06**	**0.09±0.06**	**0.09±0.06**
Alzheimer's disease	0.12±0.07	0.05±0.05	0±0	0.01±0.02	0±0.01	0.04±0.05	**0.15±0.06**	0.14±0.06
arthritis	**0.1±0.03**	0.02±0.02	0.05±0.03	0.02±0.02	0±0.01	0.09±0.03	0.09±0.03	0.08±0.03
asthma	0.08±0.06	0.01±0.02	0.01±0.02	0±0.01	0±0	**0.18±0.09**	0.1±0.05	0.08±0.04
bipolar disorder	0.06±0.04	0.02±0.02	0.04±0.06	0.02±0.02	0±0	**0.08±0.06**	0.06±0.03	0.05±0.03
cardiac arrhythmia	0.46±0.11	**0.52±0.06**	0.33±0.16	0.47±0.13	0.02±0.02	0.43±0.15	0.44±0.11	0.46±0.11
chronic obstructive pulmonary disease	0.04±0.03	0.01±0.01	0±0	0.01±0.01	0±0.01	**0.05±0.07**	0.04±0.03	0.04±0.03
coronary heart disease	**0.08±0.04**	0.04±0.05	0.04±0.04	0.05±0.04	0±0.01	0.05±0.04	0.07±0.04	0.06±0.03
drug dependence	0.07±0.03	0.05±0.04	0.01±0.01	0.06±0.05	0±0.01	0.04±0.03	0.05±0.03	**0.08±0.04**
hypertension	0.09±0.04	0.03±0.03	0.04±0.02	0.03±0.02	0±0	0.06±0.04	0.11±0.04	**0.12±0.04**
multiple sclerosis	**0.06±0.03**	0.05±0.03	0.03±0.03	0.04±0.02	0±0.01	**0.06±0.04**	0.05±0.03	0.05±0.03
obesity	0.07±0.03	0.07±0.04	**0.17±0.09**	0.05±0.03	0±0	0.12±0.06	0.07±0.03	0.08±0.04
Parkinson's disease	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0±0	**0.02±0.01**	**0.02±0.02**	**0.02±0.02**
psoriasis	**0.08±0.04**	0.02±0.02	0.02±0.02	0.01±0.02	0±0	0.05±0.03	**0.08±0.04**	0.07±0.04
rheumatoid arthritis	0.13±0.04	0.04±0.03	0.06±0.05	0.05±0.03	0.01±0.01	0.12±0.04	**0.14±0.04**	0.12±0.03
schizophrenia	0.02±0.03	0.03±0.04	0±0.01	0.03±0.04	0±0	**0.09±0.06**	0±0.01	0.01±0.03
stroke	**0.04±0.02**	0.01±0.02	0.02±0.02	0.01±0.02	0±0.01	0.02±0.02	**0.04±0.02**	0.03±0.02
systemic lupus erythematosus	**0.14±0.08**	0.04±0.05	0.01±0.02	0.03±0.05	0±0	0.13±0.07	0.13±0.07	0.12±0.08
type I diabetes mellitus	0.21±0.07	0.09±0.07	0.04±0.06	0.06±0.04	0±0	0.21±0.09	0.21±0.08	**0.24±0.1**
type II diabetes mellitus	**0.05±0.03**	0.01±0.02	0.01±0.01	0.02±0.02	0±0	0.04±0.05	**0.05±0.04**	**0.05±0.04**
ulcerative colitis	0.15±0.08	0.05±0.05	0.06±0.05	0±0.01	0±0	0.14±0.08	**0.17±0.08**	0.15±0.08
unipolar depression	**0.09±0.05**	0.08±0.03	0.05±0.04	0.07±0.03	0±0.01	0.08±0.05	0.08±0.05	0.1±0.05

Table 7: Benchmarking results on 22 Open Targets gene lists originated from GWAS studies. The left half of the table are the results for the Top 100 Normalized score. Bolded cells indicate the best performing method for that scoring metric. ADAGIO outperforms in 8 of the 22 disease gene lists for the Top 100 Metric.



(a) Top 250 Normalized score for Parkinson's with ADAGIO, Dada, RWR, and SVM where random edges were introduced into the benchmark network. Edge addition of 0 indicates the original network

(b) Top 250 Normalized score for Parkinson's with ADAGIO, Dada, RWR, and SVM where random edges were removed from the benchmark network while maintaining a spanning tree. Edge removal of 0 indicates the original network

(c) Top 250 Normalized score for Alzheimer's with ADAGIO, Dada, RWR, and SVM where random edges were introduced into the benchmark network. Edge addition of 0 indicates the original network

(d) Top 250 Normalized score for Alzheimer's with ADAGIO, Dada, RWR, and SVM where random edges were removed from the benchmark network while maintaining a spanning tree. Edge removal of 0 indicates the original network

Figure 2: Top 250 Normalized Scores when simulating network noise on Alzheimer's Disease and Parkinson's Disease on the drug gene list. Figure 2c and 2a examines noise when adding edges, and figure 2d 2b examines noise when removing edges. We find that ADAGIO is robust to missing edges and edge addition. Many of the comparable methods are not as robust as ADAGIO to edge addition as seen in Tables 2c,2a

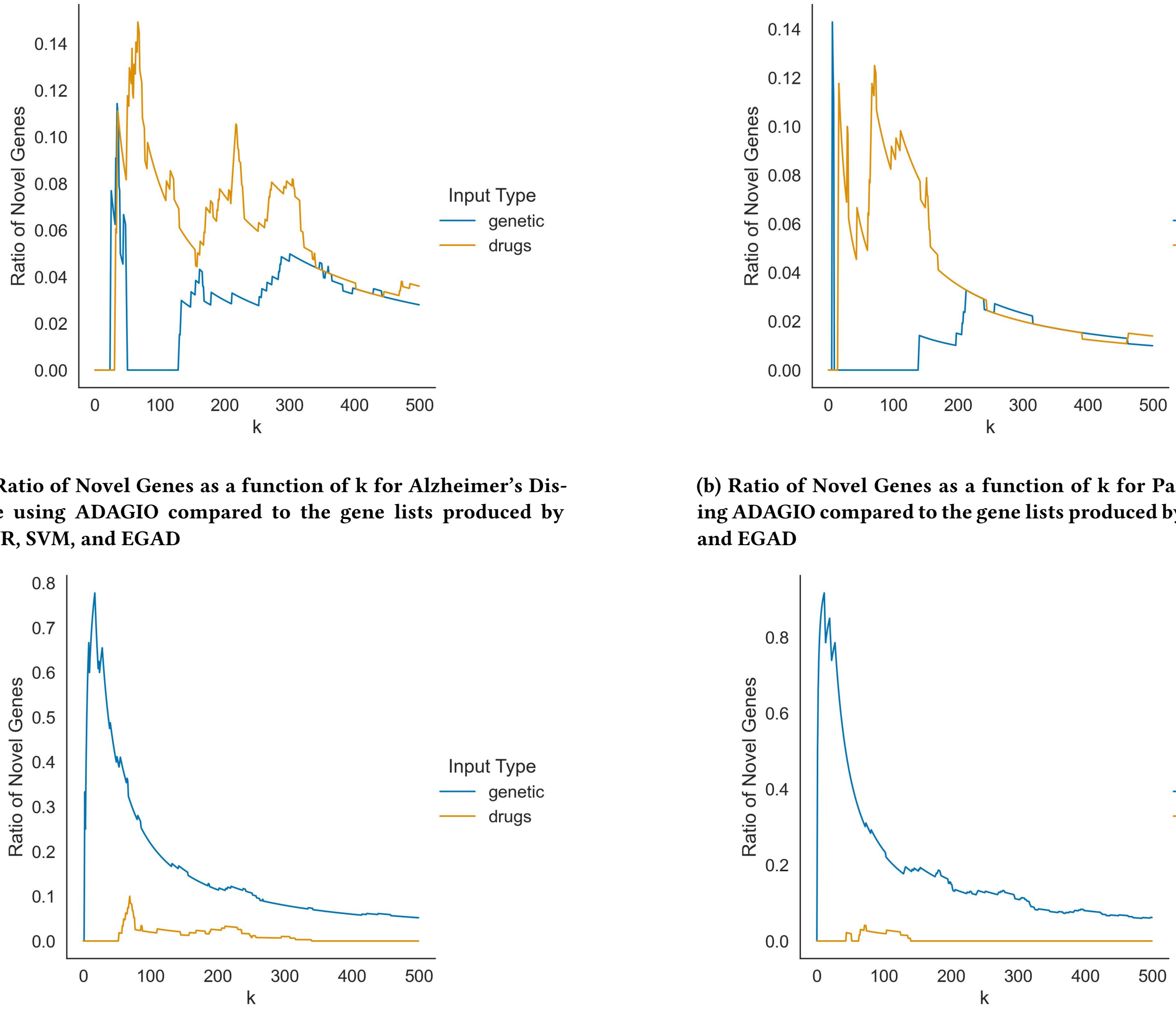
gene list, however, shows that $W(k)$ has a sharp rise and fall at $k \approx 0$, but starts growing again when $k > 150$.

The graphs in Figures 3c and 3d show the variations of $W(k)$ with k, with only DADA and RWR used as the competing algorithms. ADAGIO, DADA, and RWR are all similar to one another given that they all rely on RWR to function. We see that for drug gene lists on Alzheimer's the ratio of novel genes is effectively zero until we reach $k = 70$ where it spikes to 10% after which falling to 4%. In Parkinson's disease we see even less of a spike. When we compare against figures 3a and 3b it becomes clear that ADAGIO is able to pick up genes in the DADA list. Our results in Table 6, however, show that DADA consistently performs on par with RWR. We hypothesize that despite having similar gene predictions, DADA struggles with finding genes that are higher in degree that are not hub genes.

4.6.1 Case Analysis: Parkinson's and Alzheimer's Disease. We now take a deeper look into genes that meet the criteria for novel witness by ADAGIO.

KCNC3 ranks highly among the top 15 genes witnessed by ADAGIO in Parkinon's disease on the GWAS Open Targets gene list. KCNC3 allows ions to pass over neural membranes in response to voltage differences across the membrane; this helps with the recovery of deactivated sodium channels [13]. Two variations of a mutant KCNC3 both result in nonfunctional channels, one affecting the amino acid that detects a change in membrane potential, and the other decreasing the channel's ability to open and close its pores in response to voltage sensory information [26]. Alterations in the function of these potassium channels are linked with neurodegenerative diseases such as Parkinson's [26].

Another gene that appears ranked highly in Tables 8a,8b, and 8c is STBD1. STBD1 regulates glycophagy and intracellular glycogen transport [22]. Although STBD1 is not explicitly related to brain



(a) Ratio of Novel Genes as a function of k for Alzheimer's Disease using ADAGIO compared to the gene lists produced by RWR, SVM, and EGAD

(b) Ratio of Novel Genes as a function of k for Parkinson's using ADAGIO compared to the gene lists produced by RWR, SVM, and EGAD

(c) Ratio of Novel Genes as a function of k for Alzheimer's Disease using ADAGIO compared to the gene lists produced by RWR, and DADA

(d) Ratio of Novel Genes as a function of k for Parkinson's using ADAGIO compared to the gene lists produced by RWR, and DADA

Figure 3: Graphs of the ratio of novel genes as predicted by ADAGIO, where figure 3a, and figure 3b compares ADAGIO to RWR, SVM and EGAD for Alzheimer's and Parkinson's disease. Figures 3c and 3d do the same analysis comparing ADAGIO to RWR and DADA

function, literature evidence shows that metabolic activity is closely related with aging and neurodegenerative diseases [1, 12]. In particular experimental evidence [12] shows that mutations STBD1 impair regulatory interactions with ATG8, a well known marker [15] for Parkinson's. Additionally, GWAS studies claim that there is significant evidence for the effects of STBD1 on Parkinson's [24].

KLK6 and ATG4 are both among the top 15 novel genes witnessed by ADAGIO on the Open Targets GWAS and the Open Targets drugs gene lists for Alzheimer's disease, respectively (Tables 8b, 8a).

Many studies [3, 17] find strong evidence for correlation between the presence of KLK6 in the brain and Alzheimer's disease. Mitsui et al. [17] found a significant positive correlation between aging and KLK6 levels; the same study found that the concentration of this protein in Alzheimer's patients was significantly depleted. The study concluded that KLK6 is an important aging protein and that its absence in older patients is correlated with the presence of Alzheimer's. Other sources [3] corroborate this; one finds that this difference in KLK6 levels between Alzheimer's patients and controls is most significant in the frontal cortex of the brain.

The presence of Amyloid-β plaques is a common feature of Alzheimer's [11], and a significant body [5] of research targets understanding the proteins that contribute to Amyloid-β plaque formation. According to Barnett et al. [5], Amyloid-β plaques are a result of the buildup and subsequent rupture of autophagosomes.

Gene	ADAGIO	RWR	EGAD	SVM	Degree
ARHGEF9	32	1342	3306	10345	1
ATG4A	33	824	9127	10262	4
GIMAP6	35	1922	11397	10662	1
STBD1	36	2454	2891	11600	1
CACNG5	50	137	248	1730	50
ATG7	51	281	7811	300	12
RASGRF1	54	175	608	786	80
ATG4B	58	264	10470	8283	12
NR0B2	61	182	212	158	54
CACNG1	64	663	1897	8945	62
THRA	67	167	280	451	75
CACNA2D1	82	655	1868	8812	62
TRPV1	111	2622	331	5837	10
SCNN1D	117	1027	9689	2947	18
CACNB1	159	632	1861	9925	65

(a) Top 15 Novel Genes as witnessed by ADAGIO on the Open Targets drugs gene list for Alzheimer's Disease, not witnessed by RWR, SVM, EGAD, and their ranks, with degree of the genes annotated.

Gene	ADAGIO	RWR	EGAD	SVM	Degree
SLC6A2	25	117	2511	3021	2
SLC6A4	26	118	3760	3022	2
SCN3A	33	95	4767	2631	3
SCN9A	35	99	10460	3473	2
ARHGEF9	130	1413	3319	10256	1
ATG4A	131	904	9060	10173	4
GIMAP6	133	1975	11292	10564	1
STBD1	134	2501	2910	11496	1
ATG7	149	373	7761	353	12
ATG4B	156	356	10384	8225	12
CACNG1	162	750	1925	8872	62
CACNA2D1	180	743	1895	8745	62
SCNN1D	212	1106	9617	2976	18
CACNB1	254	720	1888	9840	65
KLK6	258	5769	5920	4981	1

(b) Top 15 Novel Genes as witnessed by ADAGIO on the Open Targets GWAS gene list for Alzheimer's Disease, not witnessed by RWR, SVM, EGAD, and their ranks, with degree of the genes annotated.

Gene	ADAGIO	RWR	EGAD	SVM	Degree
SLC39A1	16	2073	5606	8383	1
SLC39A2	17	1891	4964	8382	1
SCNN1B	30	83	229	316	28
CACNG4	45	125	179	4266	60
ARHGEF9	53	1282	2744	8927	1
ATG4A	62	802	8934	8179	4
CACNG1	64	358	453	9068	62
RASGRF1	66	169	286	386	80
GIMAP6	67	1875	11352	10023	1
STBD1	68	2395	2307	10231	1
CACNA2D1	72	340	325	9014	62
ATG7	98	522	7541	305	12
CACNB1	105	328	352	7315	65
ATG4B	112	509	10360	5979	12
CACNG6	151	365	330	8992	60

(c) Top 15 Novel Genes as witnessed by ADAGIO on the Open Targets drugs gene list for Parkinson's Disease, not witnessed by RWR, SVM, EGAD, and their ranks, with degree of the genes annotated.

Gene	ADAGIO	RWR	EGAD	SVM	Degree
KCNC3	7	36	39	56	41
KCNA7	9	40	38	45	41
SLC39A1	140	2206	5714	8485	1
SLC39A2	141	2024	5071	8484	1
ARHGEF9	198	1420	2858	9026	1
ATG4A	207	943	9024	8281	4
CACNG1	209	501	533	9167	62
GIMAP6	212	2008	11441	10116	1
STBD1	213	2527	2421	10323	1
ATG4B	257	651	10447	6090	12

(d) Top 10 Novel Genes as witnessed by ADAGIO on the Open Targets GWAS gene list for Parkinson's Disease, not witnessed by RWR, SVM, EGAD, and their ranks, with degree of the genes annotated.

Table 8: Tables of gene's witnessed by ADAGIO that are not witnessed by RWR, SVM, and EGAD. The ranks of each gene within their corresponding gene list is annotated, alongside the degree of the gene within the network. Tables 8a, 8b look at novel genes in the Alzheimer's gene lists. Tables 8c, 8d look at novel genes in the Parkinson's gene lists. In particular we note that ADAGIO is able to find mostly genes that are not just hub genes but genes with degree less than 20 which other methods are not able to discover.

ATG4 regulates the buildup of these autophagosomes [22]; when ATG4 is inhibited (for example, by reactive oxidative species (ROS)), autophagosomes accumulate.

5 DISCUSSION

We have presented ADAGIO, a new disease gene prioritization algorithm that does well against previous methods in a 22 disease benchmark. Additionally, we have investigated some of the novel genes that ADAGIO finds in two of the brain diseases in greater depth, namely Alzheimer's and Parkinson's diseases. One direction

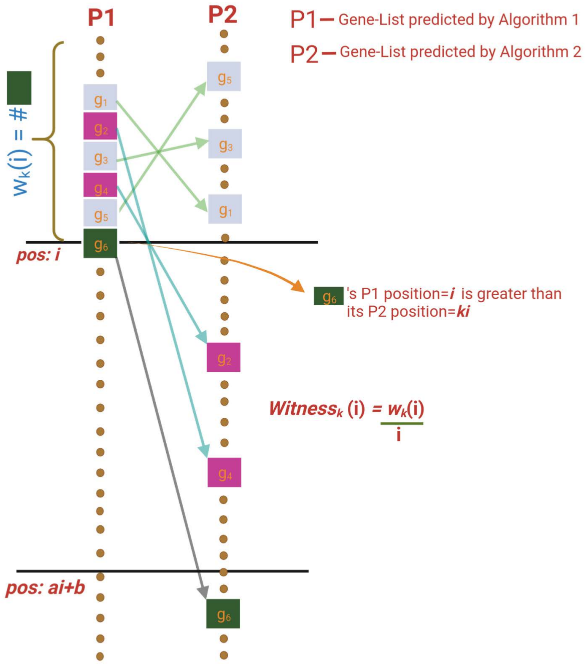

P1
P2
P1 — Gene-List predicted by Algorithm 1
P2 — Gene-List predicted by Algorithm 2
w_k(i) = #
g1
g2
g3
g4
g5
g6
pos: i
g6 's P1 position=i is greater than its P2 position=ki
Witness_k (i) = w_k(i) / i
pos: ai+b

Figure 4: Graphical description of the Witness measure, capturing novelty in the difference between two gene lists as produced by prioritization algorithms.

for future work might be to instead consider tissue-specific networks [16] that consider the level of gene expression in, for example, brain tissue, to modify the edge weights. We tried something very simple, scaling edges down by an exponential factor depending on whether or not either gene connected by the edge was brain related. Results were underwhelming; a more sophisticated analysis with different types of brain cells would be necessary to take advantage of this kind of context.

The benchmark regime we used in this paper is most appropriate in cases that consider each disease independently. However, beginning with PRINCE [25], there exists a growing body of work (e.g. [6, 23]) that considers the known relationships between *diseases* while performing gene prioritization. For example, it is known that there is some overlap between some pathways involved in the pathology of Alzheimer's and Parkinson's diseases. In future work we would like to adapt ADAGIO to this setting.

6 ACKNOWLEDGEMENTS

Thanks to NSF grants CCF-1934553 and DMS-1812503. Thanks also to the T-Tripods DIAMONDS program (part of CCF-1934553) for supporting MJS, Lord Laidlaw and the Tufts Laidlaw Scholars Program for support of SH and KL, and the NSF DREAM program for support of MG.

REFERENCES

[1] Y. Aman, T. Schmauck-Medina, et al. Autophagy in healthy aging and disease. *Nature aging*, 1(8):634–650, 2021.

[2] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486, 2006.

[3] E. Ashby, P. Kehoe, and S. Love. Kallikrein-related peptidase 6 in Alzheimer's disease and vascular dementia. *Brain research*, 1363:1–10, 2010.

[4] S. Ballouz, M. Weber, P. Pavlidis, and J. Gillis. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*, 33(4):612–614, 2017.

[5] A. Barnett and G. Brewer. Autophagy in aging and Alzheimer's disease: pathologic or protective? *Journal of Alzheimer's Disease*, 25(3):385–394, 2011.

[6] A. Cornish, A. David, and M. Sternberg. Phenorank: reducing study bias in gene prioritization through simulation. *Bioinformatics*, 34(12):2087–2095, 2018.

[7] L. Cowen, T. Ideker, B. Raphael, and R. Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551–562, 2017.

[8] K. Devkota, J. Murphy, and L. Cowen. GLIDE: combining local methods and diffusion state embeddings to predict missing interactions in biological networks. *Bioinformatics*, 36(Supplement_1):i464–i473, 2020.

[9] S. Erten, G. Bebek, R. Ewing, and M. Koyutürk. DA DA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, 4(1):19, Jun 2011.

[10] M. Ghoussaini, E. Mountjoy, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49(D1):D1311–D1320, 10 2020.

[11] G. Gouras, T. Olsson, and O. Hansson. β-amyloid peptides and amyloid plaques in Alzheimer's disease. *Neurotherapeutics*, 12(1):3–11, 2015.

[12] Z. Han, W. Zhang, et al. Model-based analysis uncovers mutations altering autophagy selectivity in human cancer. *Nature communications*, 12(1):1–18, 2021.

[13] S. Khare, K. Galeano, et al. C-terminal proline deletions in KCNC3 cause delayed channel inactivation and an adult-onset progressive SCA13 with spasticity. *The Cerebellum*, 17(5):692–697, 2018.

[14] S. Köhler, S. Bauer, D. Horn, and P. Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[15] M. Lynch-Day, K. Mao, K. Wang, M. Zhao, and D/ Klionsky. The role of autophagy in Parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(4):a009357, 2012.

[16] O. Magger, Y. Waldman, E. Ruppin, and R. Sharan. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLOS Computational Biology*, 2012.

[17] S. Mitsui, A. Okui, H. Uemura, T. Mizuno, T. Yamada, Y. Yamamura, and N. Yamaguchi. Decreased cerebrospinal fluid levels of neurosin (KLK6), an aging-related protease, as a possible new risk factor for Alzheimer's disease. *Annals of the New York Academy of Sciences*, 977(1):216–223, 2002.

[18] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, 9(1):1–15, 2008.

[19] S. Picart-Armada, S. Barrett, et al. Benchmarking network propagation methods for disease gene identification. *PLoS computational biology*, 15(9):e1007276, 2019.

[20] S. Picart-Armada, F. Fernández-Albert, et al. Null diffusion-based enrichment for metabolomics data. *PloS one*, 12(12):e0189012, 2017.

[21] S. Picart-Armada, W. Thompson, A. Buil, and A. Perera-Lluna. diffustats: an R package to compute diffusion-based scores on biological networks. *Bioinformatics*, 34(3):533–534, February 2018.

[22] M. Safran, I. Dalah, et al. Genecards version 3: the human gene integrator. *Database*, 2010, 2010.

[23] D. Smedley et al. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature protocols*, 10(12):2004–2015, 2015.

[24] A. Soto-Ortolaza, M. Heckman, et al. GWAS risk factors in parkinson's disease: LRRK2 coding variation and genetic interaction with PARK16. *American journal of neurodegenerative disease*, 2(4):287, 2013.

[25] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

[26] M. Waters, N. Minassian, et al. Mutations in voltage-gated potassium channel kcnc3 cause degenerative and developmental central nervous system phenotypes. *Nature genetics*, 38(4):447–451, 2006.

[27] H. Zhang, A. Ferguson, et al. Benchmarking network-based gene prioritization methods for cerebral small vessel disease. *Briefings in bioinformatics*, 22(5):bbab006, 2021.