Knowledge-driven annotation for gene interaction enrichment analysis

Xiaoyu Liu[1, †], **Anna Jiang**[1, †], **Chengshang Lyu**[1], **and Lingxi Chen**[1,2,*]

[1]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong

[2]Tung Biomedical Sciences Centre, City University of Hong Kong, Hong Kong

[*]Corresponding: lingxi.chen@cityu.edu.hk

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Gene Set Enrichment Analysis (GSEA) is a cornerstone for interpreting gene expression data, yet traditional approaches overlook gene interactions by focusing solely on individual genes, limiting their ability to detect subtle or complex pathway signals. To overcome this, we present GREA (Gene Interaction Enrichment Analysis), a novel framework that incorporates gene interaction data into enrichment analysis. GREA replaces the binary gene hit indicator with an interaction overlap ratio, capturing the degree of overlap between gene sets and gene interactions to enhance sensitivity and biological interpretability. It supports three enrichment metrics: Enrichment Score (ES), Enrichment Score Difference (ESD) from a Kolmogorov-Smirnov-based statistic, and Area Under the Curve (AUC) from a recovery curve. GREA evaluates statistical significance using both permutation testing and gamma distribution modeling. Benchmarking on transcriptomic datasets related to respiratory viral infections shows that GREA consistently outperforms existing tools such as blitzGSEA and GSEApy, identifying more relevant pathways with greater stability and reproducibility. By integrating gene interactions into pathway analysis, GREA offers a powerful and flexible tool for uncovering biologically meaningful insights in complex datasets. The source code is available at `https://github.com/compbioclub/GREA`.

Introduction

Gene set enrichment analysis has served as a cornerstone of gene expression studies, systematically revealing pathway- and process-level biological significance (1). Building on this foundation, methodological innovations have continuously expanded its analytical scope and interpretative power. The classical GSEA method (2) ranks genes based on their differential expression (e.g., Log2FoldChange) across different phenotypes and calculates an enrichment score using a Kolmogorov-Smirnov-like statistic. This approach evaluates whether genes from a given set tend to accumulate at the top or bottom of the ranked gene list, providing insight into pathway-level activity. To enhance the computational efficiency, blitzGSEA (3) was introduced as an optimized variant of GSEA. Instead of relying on full-scale permutation testing, blitzGSEA employs Gamma distribution to fit the null distribution model by using the subset of the whole library of the pathway. Beyond group-wise enrichment analysis, single-sample methods have been developed to extend GSEA to individual samples. Single-sample GSEA (ssGSEA) (4) calculates an enrichment score for each sample separately based on the absolute expression, enabling pathway activity assessment in heterogeneous datasets without predefined case-control labels. Further refining this approach by modeling gene expression distributions, Gene Set Variation Analysis (GSVA) (5) estimates pathway activity by transforming gene expression profiles into sample-wise enrichment scores using a non-parametric kernel-based approach. Similar to the ssGSEA without labels, GSVA assesses whether a gene is highly or lowly expressed relative to the sample population and computes enrichment scores using classical Kolmogorov-Smirnov (KS) random walk statistic (6) and enrichment score difference statistic. This enables pathway-level analysis in an unsupervised manner, though its reliance on distributional assumptions may not always align with real biological variability. Additionally, tools like GSEApy, which is the Python version of the GSEA, provide flexible implementations of GSEA and GSVA, enabling users to perform enrichment analysis with various statistical frameworks and pathway databases (7).

However, most current enrichment approaches (Table 1) primarily focus on changes in individual genes, often overlooking the complex interactions between genes that drive biological functions. Since biological processes typically emerge from coordinated gene interactions rather than isolated gene activity, neglecting these relationships can limit the biological interpretability of enrichment results. While the existing methods have significantly advanced gene expression analysis, a persistent challenge remains: the inability of these tools to effectively incorporate gene interactions into the enrichment framework.

Thus, we introduce the GREA (Gene Interaction Enrichment Analysis), a novel framework that incorporates gene interaction rankings into enrichment analysis. Unlike conventional methods that focus on individual genes, GREA considers the interactions between genes, enabling a more holistic assessment of target gene set enrichment and improving the detection of subtle pathway signals that might be missed by existing approaches. Specifically, GREA replaces the conventional binary gene hit indicator with an interaction overlap ratio, quantifying the degree of overlap between the target gene set and each gene interaction. This adaptation allows the enrichment analysis, particularly the KS-based run-

ning sum statistic, to be applied effectively to gene interaction rankings. In addition to the traditional KS method, GREA introduces the recovery curve approach and computes the Area Under the Curve (AUC), which highlights the early enrichment of pathway-relevant interactions. Moreover, GREA enhances statistical significance estimation by combining permutation-based testing with gamma distribution modeling, offering more reliable p-values—especially in sparse datasets where permutation tests alone may yield limited biological insight. By integrating interaction-aware metrics and robust significance testing, GREA provides a comprehensive and flexible framework for identifying biologically meaningful signals in complex data. To demonstrate its effectiveness, we benchmarked GREA using publicly available transcriptomic datasets related to respiratory viral infections, validating its ability to uncover relevant pathways with greater sensitivity and interpretability than existing methods.

Methods

Algorithm of GREA.

Problem Formulation. Consider a gene regulatory network (GRN) represented as graph $G = (\mathcal{V}, \mathcal{E})$, where vertices $v \in \mathcal{V}$ correspond to genes and edges $(u, v) \in \mathcal{E}$ represent gene interaction between u and v. There are $N = |\mathcal{E}|$ gene interactions, and each interaction $\mathcal{I} = (u, v)$ is associated with a quantitative rank score $x_{u,v}$. When phenotypic labels are available, the score is phenotype-specific, derived from gene expression data conditioned on the phenotype. Otherwise, the score is specific for each sample.

Given a pathway-defined gene set $\mathcal{P} \subseteq \mathcal{V}$, we aim to develop a statistical framework to test whether genes in $\mathcal{P}$ exhibit significant enrichment at the extremes of the ranked interactions. The proposed methodology, implemented in GREA, operates through three principal computational phases: (1) Interaction Ranking. We compute entropy-based rank scores for all gene interactions using gene expression profiles, which may depend on phenotype labels when available. (2) Enrichment Quantification. We evaluate the distribution of gene set $\mathcal{P}$ within the ranked gene interactions using an adaptation of the Kolmogorov-Smirnov (KS) or Recovery-Curve (RC) running sum statistic. (3) Statistical Validation. We determine the statistical significance of the observed enrichment signals by fitting the Gamma distribution to approximate the null distribution, which is generated through pathway-preserving network randomization. In cases where the Gamma distribution provides a poor fit, we instead employ permutation testing to construct the null distribution and compute empirical p-values.

Calculating entropy-based rank score for gene interaction. First, we leverage the NIEE algorithm (8) to calculate an entropy-based rank score $x^s_{u,v}$ for gene interaction (u, v) in sample s. First, the top 5,000 highly variable genes (HVGs) are screened from the input dataset by the scanpy toolkit with parameter `flavor=seurat_v3` (9, 10). StringDB Protein-Protein Interaction Networks (`score>850`) are then employed as the background GRN to form the graph $G = (\mathcal{V}, \mathcal{E})$ (11).

The reference network is constructed using all reference samples $\mathcal{R}$, usually representing relatively healthy or control samples. For each gene u, its reference entropy, $H_u(\mathcal{R})$, is calculated as:

$$H_u(\mathcal{R}) = -\frac{1}{\log L} \sum_{w=1}^{L} p_{u,w}(\mathcal{R}) \log p_{u,w}(\mathcal{R}), \qquad \textbf{(1)}$$

where w represents the first-order neighbor genes of gene u, and L is the number of neighbors for gene u. $p_{u,w}(\mathcal{R})$ is the probability of the interaction between gene u and w, defined as:

$$p_{u,w}(\mathcal{R}) = \frac{|PCC_{u,w}(\mathcal{R})|}{\sum_{w=1}^{L} |PCC_{u,w}(\mathcal{R})|}, \qquad \textbf{(2)}$$

where $PCC_{u,w}(\mathcal{R})$ is the Pearson correlation coefficient between gene u and w in $\mathcal{R}$ reference samples.

The reference entropy $H_{u,v}(\mathcal{R})$ for gene interaction (u, v) can be calculated as:

$$H_{u,v}(\mathcal{R}) = \frac{H_u(\mathcal{R}) + H_v(\mathcal{R})}{2}. \qquad \textbf{(3)}$$

Similarly, the reference standard deviation $SD_{u,v}(\mathcal{R})$ for gene interaction (u, v) is:

$$SD_{u,v}(\mathcal{R}) = \frac{SD_u(\mathcal{R}) + SD_v(\mathcal{R})}{2}, \qquad \textbf{(4)}$$

where $SD_u(\mathcal{R})$ is the standard deviation for gene u across the reference samples $\mathcal{R}$.

After constructing the reference network, a sample s is added to the reference samples $\mathcal{R}$ to form the sample-specific network $\mathcal{R}^s$. The sample-specific entropy $H_{u,v}(\mathcal{R}^s)$ and the standard deviation $SD_{u,v}(\mathcal{R}^s)$ for gene interaction (u, v) are then calculated following the same procedure as in the reference network.

To quantify the fluctuations caused by sample s, we can calculate the sample-specific entropy-based rank score $x^s_{u,v}$ for gene interaction (u, v) as:

$$\begin{aligned} x^s_{u,v} = &|H_{u,v}(\mathcal{R}^s) - H_{u,v}(\mathcal{R})| \\ &\cdot |SD_{u,v}(\mathcal{R}^s) - SD_{u,v}(\mathcal{R})| \end{aligned} \qquad \textbf{(5)}$$

Given the availability of phenotype labels for the sample-specific data, we derived the phenotype-specific score $x^P_{u,v}$ by aligning phenotype labels to samples. We then performed differential gene interaction expression analysis to compare entropy-based scores of gene interaction (u, v) across different phenotypic groups. This analysis was conducted using the scanpy package (9) with the Wilcoxon statistical test. Specifically, we compute the log-fold change of $x^P_{u,v}$ between phenotypic groups to quantify the magnitude and direction of score alterations associated with different conditions. To assess statistical significance, we applied the Wilcoxon rank-sum test, obtaining test statistics that indicate the extent of these differences. The resulting test statistics are ranked in descending order to prioritize gene interactions that exhibit the most significant phenotype-associated

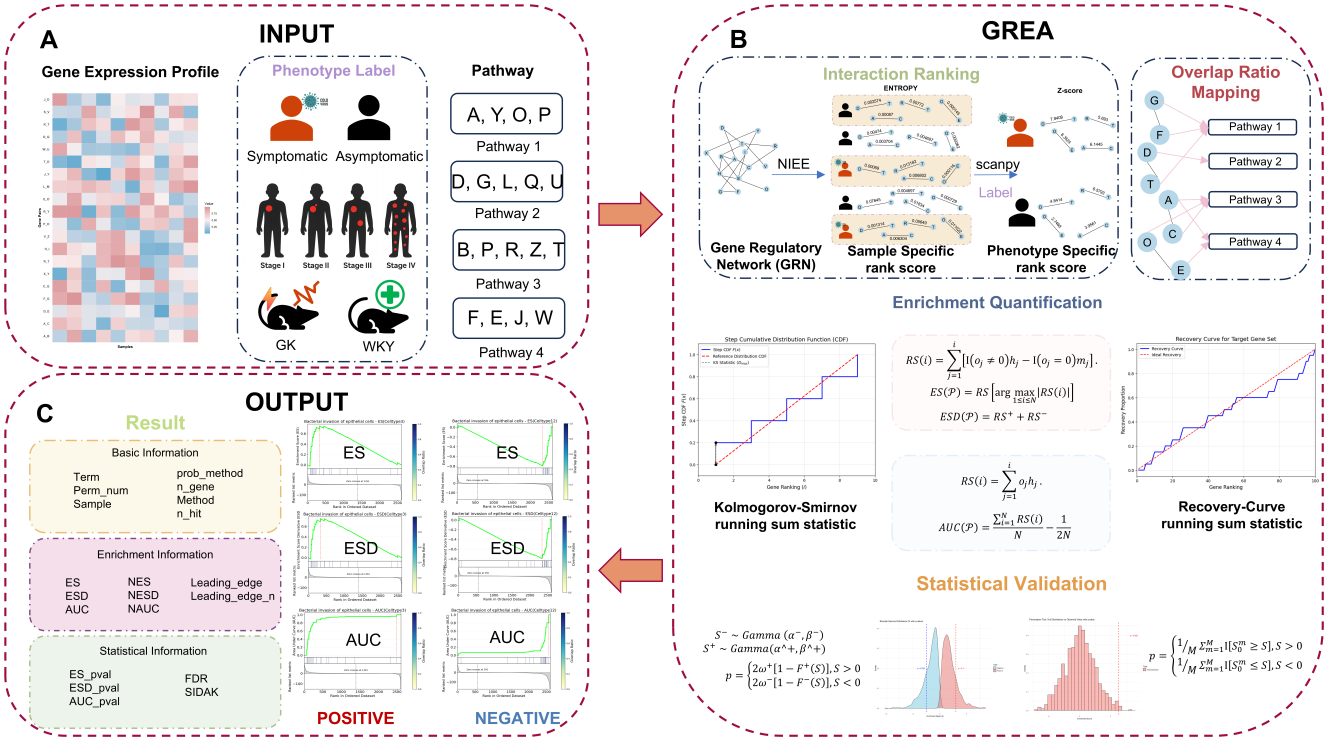

Fig. 1. Graphic overview of GREA algorithm. (A) The input for the GREA algorithm is gene expression profiles, phenotype labels (such as symptomatic and asymptomatic), and pathway information. (B) The process of analyzing gene interactions with GREA. The process begins with Interaction Ranking based on gene interaction expression values, yielding sample-specific scores. These scores are then combined with phenotype labels to generate phenotype-specific scores. Overlap ratio mapping illustrates the relationships between different pathways. Then, enrichment quantification employs statistical methods such as Kolmogorov-Smirnov running sum statistic and Recovery-Curve running sum statistic to evaluate gene set enrichment. Finally, Statistical validation is performed using Gamma distributions and permutation testing to derive p-values to determine the significance of the results. (C) The outputs for the algorithm ar tabular files containing metrics quantifying the degree of enrichment for specific gene sets. The output also includes visualization plots of ES, ESD and AUC

		GREA	GSEApy	blitzGSEA
Rank Signature	Phenotype-specific	✓	✓	✓
	Rank unit	gene interaction	gene	gene
Enrichment	Rank value	DT log2FC/z-score of entropy score	DT log2FC/z-score of gene expression	DT log2FC/z-score of gene expression
	WKS approach	ES, ESD	ES, ESD	ES, ESD
	RC approach	AUC	-	-
Significant Testing	Permutation	✓	✓	-
	Analytical	Gamma	-	Gamma
Software	Efficiency	H,A	M	H,A
	Access	Python	Python	Python

Table 1. Functionality comparison of enrichment methods. DT: Differential Test, WKS: Weighted Kolmogorov-Smirnov, RC: Recover Curve, ES: Enrichment Score, NES: Normalized Enrichment Score, ESD: Enrichment Score Difference, AUC: Area under Curve, L: Low M: Moderate. H: High. A: Accurate

entropy shifts. These ranked test statistics serve as input for the subsequent phenotype-based analysis in our algorithmic framework.

Calculating gene set enrichment signal for ranked gene interactions. Recall there are N gene interactions, which are ranked based on the sample-specific or phenotype-specific rank scores described in the previous section. For simplicity, let interaction $\mathcal{I}_i$ be the i-th ranked interaction, and its rank score is x_i. Given a pathway-defined gene set $\mathcal{P}$, if interaction $\mathcal{I}_i$ overlap with $\mathcal{P}$ (i.e., $\mathcal{I}_i \cap \mathcal{P} \neq \emptyset$), we define the overlap ratio as $o_i = \frac{|\mathcal{I}_i \cap \mathcal{P}|}{|\mathcal{I}_i|}$, where $|\cdot|$ denotes the cardinality. Clearly, the overlap o_i is determined by the presence of genes in interaction I_i within the gene set: $o_i = 0.5$ if only one gene is included, and $o_i = 1$ if both genes are present.

Next, we define the hit score h_i as the absolute rank score $|x_i|$ of list $\mathcal{I}_i$ weighted by the overlap ratio o_i, normalized by the sum of the absolute rank scores of all hit lists:

$$h_i = \frac{o_i |x_i|}{\sum_{i'=1}^{N} o_{i'} |x_{i'}|}. \tag{6}$$

In contrast, if neither gene from interaction $\mathcal{I}_i$ is in the gene set, $\mathcal{P}$ (i.e., $\mathcal{I}_i \cap \mathcal{P} = \emptyset$), we set the overlap ratio $o_i = 0$ and define its miss score m_i as the proportion of interactions com-

pletely absent from the gene set:

$$m_i = \frac{1}{\sum_{i'=1}^{N} \mathbb{1}(o_{i'} = 0)}, \tag{7}$$

where $\mathbb{1}(\cdot)$ is an indicator function that equals 1 if the condition is true and 0 otherwise.

For each interaction $\mathcal{I}_i$, we perform enrichment analysis using two running sum (RS) statistics: Kolmogorov–Smirnov (KS) (6) approach and recovery curve (RC) approach (12).

The RS in the KS-like approach accumulates hit and miss scores from the top-ranked interactions up to a given interaction $\mathcal{I}_i$:

$$RS(i) = \sum_{j=1}^{i} [\mathbb{1}(o_j \neq 0)h_j - \mathbb{1}(o_j = 0)m_j]. \tag{8}$$

There are two ways to compute the enrichment signal of a given gene set $\mathcal{P}$ over a list of pre-ranked interactions from the RS. In the first way, the enrichment score (ES) of the entire gene set $\mathcal{P}$ is then defined as the maximal absolute RS across all interactions:

$$ES(\mathcal{P}) = RS\left[\arg\max_{1 \leq i \leq N} |RS(i)|\right]. \tag{9}$$

The second way computes the difference between the positive and negative maximum derivatives of scores, which we call the enrichment score difference (ESD):

$$ESD(\mathcal{P}) = RS^{+} + RS^{-} \tag{10}$$

For a more detailed definition formulation, please refer to the full expression in Supplementary Methods.

In the RC approach, the recovery curve is defined as the RS that accumulates only the hit scores from the top-ranked interactions up to a given interaction $\mathcal{I}_i$:

$$RS(i) = \sum_{j=1}^{i} o_j h_j. \tag{11}$$

Next, we compute the area under the curve (AUC) as the enrichment signal of a given gene set $\mathcal{P}$ over a list of pre-ranked interactions from the RS. This AUC value has a range between 0 and 1, and we numerically estimate it using trapezoidal rule (13):

$$AUC(\mathcal{P}) = \frac{\sum_{i=1}^{N} RS(i)}{N} - \frac{1}{2N}. \tag{12}$$

Statistical significance test. To test whether the pathway-defined gene set $\mathcal{P}$ is significantly enriched at the extremes of the ranked interactions, we assume a null hypothesis (H_0) is that the gene set $\mathcal{P}$ is randomly distributed across the ranked interactions, meaning there is no significant enrichment of $\mathcal{P}$ at either the top or bottom of the ranked interactions. We adopt two methods to approximate the null distribution and calculate the p-value: analytical-based and permutation-based.

In the analytical approach, we randomly shuffle the interaction ranks M times (e.g. $M = 1000$) and recalculate the enrichment signal S_0^m for each shuffle m. Next, we approximate the null distribution S_0^m using a bimodal model that separately accounts for positive and negative signals. Specifically, it is modeled as a mixture of two Gamma distributions: one representing all positive signals, S^+, and the other representing all negative signals, S^-:

$$\begin{aligned} S^+ &\sim \text{Gamma}(\alpha^+, \beta^+), \\ S^- &\sim \text{Gamma}(\alpha^-, \beta^-), \end{aligned} \tag{13}$$

where α^+, β^+ and α^-, β^- are the shape and scale parameters for the positive and negative Gamma distributions, respectively.

Thus, as detailed in Supplementary Methods, the two-sided p-value of an observed enrichment signal S is calculated as twice the probability that S falls within the bimodal distribution formed by the positive and negative Gamma models:

$$p = \begin{cases} 2w^+[1 - F^+(S)], & S > 0, \\ 2w^-[1 - F^-(S)], & S < 0, \end{cases} \tag{14}$$

where $F^+(S)$ and $F^-(S)$ denote the cumulative distribution function (CDF) of the positive and negative Gamma distribution, respectively. w^+ and w^- represent the weight of positive and negative Gamma distributions, respectively, which are determined based on the ratio of positive to negative scores in the null distribution. Specifically, we define:

$$w^+ = \frac{N^+}{N^+ + N^-}, \quad w^- = \frac{N^-}{N^+ + N^-} \tag{15}$$

where N^+ and N^- denote the number of positive and negative enrichment scores in the null distribution, respectively. This ensures that the weights reflect the ratio of positive to negative scores in the null distribution:

$$w^+ + w^- = 1. \tag{16}$$

In contrast to the Kolmogorov–Smirnov (KS) approach, which considers both positive and negative signals, the recovery curve (RC) method focuses solely on positive signals. Following the same procedure, we randomly shuffle the interaction ranks M times (e.g., $M = 1000$) and recalculate the enrichment signal S_0^m for each shuffle m. The null distribution of the enrichment signals is then approximated using a Gamma distribution account for a positive signal, S:

$$S \sim \text{Gamma}(\alpha, \beta), \tag{17}$$

where α and β are the shape and scale parameters for the Gamma distribution. Given this model, the statistical significance of an observed enrichment signal S is assessed by calculating its position within the estimated distribution.Specifically, the one-side p-value is defined as:

$$p = 1 - F(S), \tag{18}$$

where F(S) denote the CDF of the Gamma Distribution. However, upon examining the patterns of the Recovery Curve and the Gamma distribution, we found that, in most cases, accurately fitting the null distribution of the RC method is challenging. As a result, we implement a warning system to notify users of potential model misfits and recommend using a permutation-based approach as an alternative, as described below.

In the permutation test, the nominal p-value is computed by comparing the observed enrichment score S to the null distribution of enrichment scores S_0^m obtained from the permutations. The p-value formula is:

$$p = \begin{cases} \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(S_0^m \geq S), & S > 0; \\ \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(S_0^m \leq S), & S < 0. \end{cases} \quad \textbf{(19)}$$

We compare the calculated p-value to a pre-defined significance level α (commonly set at 0.05) to determine if the gene set enrichment is statistically significant. If $p \leq \alpha$, we reject the null hypothesis, concluding that the gene set $\mathcal{P}$ is significantly enriched at the top or bottom of the ranked interactions.

Given that multiple gene sets are often tested simultaneously, we apply the False Discovery Rate (FDR) adjustment to control for multiple comparisons with Benjamini-Hochberg (14) and Sidak correction (15) approaches.

Finally, larger gene sets tend to exhibit higher enrichment signals, S, purely by chance. To account for variability in gene set size and ensure comparability across different gene sets, we compute the normalized enrichment signal for each gene set, $\mathcal{P}$, by dividing the observed enrichment signal by the mean enrichment signal derived from random permutations (null distribution):

$$N(S) = \frac{S \times M}{\sum_{m=1}^{M} S_0^m}. \quad \textbf{(20)}$$

Experiment setting.

Dataset Description. To evaluate the performance of GREA, we conducted a series of benchmarking experiments using publicly available transcriptomic datasets sourced from the NCBI Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). Specifically, we selected two datasets, GSE52428 (16) and GSE73072 (17), both of which profile host gene expression responses to experimental respiratory viral infections. GSE52428 contains whole blood transcriptional profiles from human subjects who were challenged with either influenza H1N1 or H3N2 viruses. GSE73072 comprises a broader range of respiratory virus challenges, including influenza A subtypes (H1N1 and H3N2), human rhinovirus (HRV), and respiratory syncytial virus (RSV).

Detailed application of the GREA algorithm on these datasets enables a comprehensive assessment of its capability to capture biologically meaningful enrichment signals. In the following, we present the complete analytical workflow and key findings obtained from applying GREA to the selected viral infection datasets and benchmarking between GREA and past enrichment methods.

P value Benchmarking. The primary goal of this p-value benchmark analysis is to evaluate and compare the reproducibility and similarity of different enrichment methods. To achieve this, we designed multiple preprocessing strategies that generate phenotype-specific scores by ranking z-scores derived from differential test, using gene expression or entropy-based interaction data to construct the input matrix for all methods. The preprocessing methods are summarized as follows: *Exp_Entropy* involves selecting the top 5,000 highly variable genes (HVGs) from expression data, identifying these genes within entropy-based gene interactions with sample-specific scores, and then transforming them into phenotype-specific scores. This method is used for internal evaluation of our approach. *Exp_SingleGene* also starts by selecting 5,000 HVGs from expression data and identifying corresponding entropy gene interactions but then decomposes each gene pair into individual genes by averaging their sample-specific scores before converting to phenotype-specific scores. *SingleGene* first selects 5,000 highly variable entropy gene interactions, transforms their sample-specific scores to phenotype-specific scores, and then averages the phenotype-specific scores for each gene to derive single-gene scores. *Entropy*, another internal baseline, selects 5,000 highly variable entropy gene interactions and directly transforms sample-specific scores into phenotype-specific scores without further decomposition. *HVG_SingleGene* begins by selecting 5,000 highly variable entropy gene interactions, decomposes them into single genes by averaging sample-specific scores, and then transforms these into phenotype-specific scores. *SingleGene_HVG* first decomposes gene interactions into single genes with averaging sample-specific value, then selects 5,000 HVGs based on the resulting scores, followed by phenotype-specific transformation. Finally, *EXP* selects 5,000 HVGs directly from expression data and transforms their sample-specific scores into phenotype-specific scores. These diverse preprocessing schemes enable a comprehensive evaluation of enrichment methods under various input conditions.

In the first experiment for p-value, we used the Enrichr (18) gene set library created from KEGG (19) using pre-rank in GSEApy (7), blitzGSEA (3) and GREA with 500, 1,000 and 2,000 permutations applied to different dataset related to respiratory viral infections in different preprocessing methods. We examined the consistency of p-values across 3 duplications with the same number of permutations. After calculating the Spearman correlation (corr) (20) between p-values, which means the similarity of results, we set the variances as $1 - corr$ and recorded based on the number of permutations for GREA on ES, GREA on ESD, GREA on AUC, blitzGSEA (3) and GSEApy (7). If the correlation is high (close to 1), the error approaches 0, indicating stable results. If the correlation is low, the error approaches 1, indicating unstable results.

For the second experiment, similar to the setting in the first

experiment, we computed pairwise Spearman rank correlations across all method combinations to assess how consistently they ranked pathway significance. This process was repeated for all (repetition time and Permutation number) combinations, and the resulting correlation matrices were averaged to produce a final mean Spearman correlation matrix.
In order to find the minimum p-value pathway under different methods, we set the third experiment. The color scale represents the -log10(p-value) values, ranging from light pink (-log10(p-value) ~ 0) to dark purple (-log10(p-value) ~ 3). Grey blocks indicate that a particular method did not identify a significant pathway.

Results

Overview of GREA. We developed Gene Interaction Enrichment Analysis (GREA), a novel enrichment analysis tool that quantifies the pathway-level difference by integrating gene regulatory interactions and rank-based scoring strategies. The overall workflow of GREA is illustrated in Figure 1. This method is designed to address the limitation of traditional GSEA in handling gene interaction situations. GREA takes three main components as the inputs: (1) a gene expression profile, (2) phenotype labels (e.g. disease and control), and (3) a predefined gene set library (e.g. KEGG, GO) (Figure 1A).
Figure 1B illustrates the core algorithm of GREA. First, we first derive the sample-specific gene interaction ranking score $x^s_{u,v}$ from the gene regulatory network (GRN) based on entropy-based algorithm. These sample-specific interactions are transformed into phenotype-specific ranking score $x^P_{u,v}$ for each gene interaction with differential test. GREA computes three types of enrichment signals: (1) Enrichment Score (ES) derived from the Kolmogorov–Smirnov statistic; (2) Enrichment Score Difference (ESD), which quantifies the maximum difference between positive ES and negative ES; (3) Area Under the Curve (AUC) of recovery curves that only focus on the hit gene in the predefined gene set library. Here, we propose a continuous overlap ratio metric that quantifies the degree to which a pathway is represented in high-ranking regulatory interactions. This contrasts with traditional binary hit-counting indicators and improves sensitivity for partially active pathways. After getting the enrichment signals, GREA employs both permutation test and gamma distribution fitting to evaluate the statistical significance, enabling flexible hypothesis testing across different data sizes and sparsity levels. GREA outputs a ranked list of enriched pathways for each phenotype, along with p-values and essential information. We also provide the visualization for both KS and RC approaches (Figure 1C).
Overall, GREA bridges gene interaction and pathway-level interpretation, offering improved sensitivity and robustness in complex biological settings. In the following sections, we evaluate GREA against state-of-the-art tools for its analytical power across real-word disease datasets.

P-value Benchemarking. In Figure 2, we present a comprehensive evaluation of the performance of different enrichment analysis methods across multiple dimensions. Notably, these analyses are based on a combined dataset that aggregates results from multiple disease-specific datasets. To provide further specificity, we have also generated individual figures for each dataset separately Supplementary Figures 1-3. This allows for a more detailed examination of method performance within specific biological contexts.

In Figure 2A, we present a comparative boxplot of the variance in enrichment p-values estimations reproducibility across all disease datasets for different preprocessing methods and enrichment strategies. The y-axis represents the variance of p-values in repeating three times, while the x-axis lists the combinations of preprocessing pipelines, enrichment methods, and statistical models. Each boxplot summarizes the distribution of variances under three different numbers of permutations (500, 1,000, and 2,000), shown in red, beige, and green, respectively.

On the left and middle parts of the figure, we observe results from our method GREA, under different statistical frameworks: permutation testing (Perm), gamma distribution fitting (Gamma), and hybrid approaches like KS and AUC-based scoring. In general, GREA-based methods show consistently lower variance across, with the minimum value in variance being 0.000405 compared to the alternative methods shown on the far right: blitz and GSEApy, with the minimum variance of 0.001599 and 0.001795, separately. These two commonly used enrichment tools exhibit higher and more dispersed variance, especially under different preprocessing inputs. Moreover, the blitz and GSEApy demonstrate more outliers than our GREA methods. This suggests that GREA provides more stable enrichment results, particularly when applying phenotype-specific scores derived from entropy or interaction-based features.

Within the GREA method, the figure highlights that preprocessing methods like Exp_Entropy and Entropy generally lead to higher variance, 0.044748, than other methods. Gamma-based models also appear to reduce variability further in certain cases, like all preprocessing methods in AUC. These internal comparisons clearly show the gamma distribution fitting adds more stability to p-value production than permutation testing.

Same dataset for Figure 2B, we want to evaluate the similarity between our method and other methods in past research. Spearman correlation(20) indicates that the two methods assign similar levels of significance to the same sets of pathways across diverse biological contexts.

Similar to initial expectations, it is worth noting that our unique part of analyzing gene interactions in GREA is not very similar to the past two methods, both of which are $\rho < 0.5$. The results in blitz are highly connected with our enrichment in ES, ESD in gamma fitting and AUC in permutation testing. By contrast, methods based on the GSEApy enrichment method (e.g., SingleGene-GSEApy, HVG_SingleGenr-GSEApy) exhibit low or even negative correlation with nearly all other methods ($\rho < 0.3$), including both gamma- and permutation-based GREA configurations. This indicates that GSEApy scoring yields a different enrichment rank way.

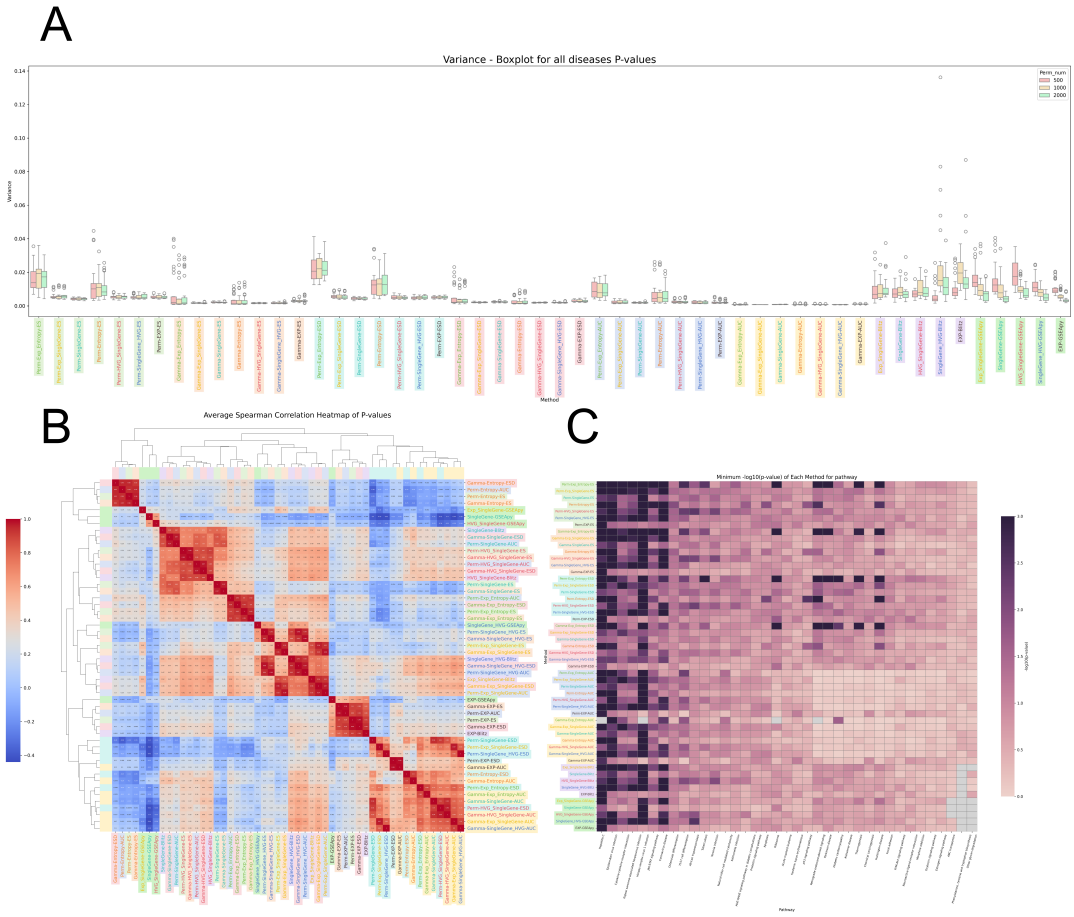

Fig. 2. P-value Benchmarking. (A) Box plot comparing the reproducibility of p-values across three repeated experiments under different permutations, measured by variance. (B) Heatmap showing the average Spearman correlation of p-values between each pair of different methods. Both axes represent different preprocessing methods, and each cell indicates the correlation between the p-values obtained from the corresponding pair. (C) Heatmap displaying the minimum -log10(p-value) obtained by each preprocessing method for its most significantly enriched pathway. The y-axis lists preprocessing methods, and the x-axis shows the corresponding pathways identified by each method. Color intensity reflects the statistical significance, with higher values indicating stronger enrichment.

Within our GREA methods, a clear block of high internal similarity is observed among GREA configurations between AUC in gamma fitting and ESD in permutation testing, the same as in AUC in permutation testing and ESD in gamma fitting. The detailed results of the variance, mean and sample number for different repetitions, permutations, and methods are provided in the Supplementary Table 1

As shown in Figure 2C, the performance of different enrichment analysis methods varies significantly in terms of pathway detection sensitivity, as measured by the minimum -log10(p-value) across pathways. Notably, past methods blitz and GSEApy, particularly those using HVG- or SingleGene-based gene sets, exhibit poor performance on the right-hand side of the heatmap, where a substantial number of pathways show extremely low significance levels or are entirely undetected (grey or pale pink regions). This indicates a limited capacity of these methods to identify relevant biological pathways under these settings.

In contrast, our proposed method, GREA, demonstrates a clear advantage in both sensitivity and coverage. When combined with the EXP_Entropy gene selection strategy and either the ES or ESD enrichment approach, GREA consistently identifies more pathways with significantly higher -log10(p-values), especially in pathway Ribosome, Retrograde endocannabinoid signaling, Parkinson disease, Thermogenesis and Huntington disease. These results highlight the effectiveness of incorporating interaction gene pair information in enhancing signal detection and biological relevance. The ability of GREA to capture additional significant pathways not detected by other methods suggests that its design better reflects the underlying regulatory or functional relationships in the data. Overall, this comparison underscores the robustness and improved interpretability of GREA in complex pathway enrichment scenarios. Detailed information on the minimum p-values, maximum -log10(p-values), and corresponding pathways for each method can be found in Supplementary Table 2.

Discussion

In this study, we introduced GREA, a novel framework that integrates gene regulatory interactions into pathway library enrichment analysis. Our results demonstrate that GREA offers several key advantages over traditional GSEA methods, particularly in its ability to capture subtle but biologically

meaningful signals in gene interactions.

Through p-value benchmarking experiments, we showed that GREA produces more stable and reproducible results compared to past enrichment tools like blitzGSEA (3) and GSEApy (7). The incorporation of continuous overlap ratios and different statistical testing approaches (permutation testing and gamma distribution fitting) significantly enhances the robustness of enrichment signal detection. Notably, GREA consistently identified more significant pathways across diverse preprocessing strategies, suggesting its versatility in handling different types of input data.

One of the most important contributions of GREA is its focus on gene interactions rather than individual genes. This approach gave a novel horizon of biological systems, where subtle difference is often found from coordinated interactions between genes. The ability of GREA to detect partial pathway activation and subtle changes in regulatory networks represents a substantial improvement over binary hit-counting methods used in traditional GSEA.

Despite these advances, some limitations remain. While the gamma distribution fitting generally improved stability, we observed challenges in accurately modeling the null distribution for recovery curve-based enrichment. Future work should focus on refining these aspects and extending the capabilities of GREA to single-sample analysis, which would be particularly valuable for biological pathway identification applications.

In conclusion, GREA represents a significant step forward in gene set enrichment analysis by explicitly considering gene interactions and employing advanced statistical methodologies. As biological research continues to generate increasingly complex datasets, tools like GREA will be essential for extracting meaningful insights and advancing our understanding of cellular processes.

Code Availability

The source code of GREA can be accessed in https://github.com/compbioclub/GREA.

Author Contributions Statement

XYL: Software; Formal analysis; Validation; Visualization; Investigation; Methodology; Writing—original draft; Writing—review and editing. ANJ: Software; Formal analysis; Validation; Visualization; Investigation; Methodology; Writing—original draft; Writing—review and editing. CSL: Formal analysis; Validation; Investigation; Writing—original draft; Writing—review and editing. LXC: Conceptualization; Supervision; Software; Methodology; Writing—original draft; Writing—review and editing.

Competing Interests

The authors declare no competing interests.

References

1. Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.
2. Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.
3. Alexander Lachmann, Zhuorui Xie, and Avi Ma'ayan. blitzgsea: efficient computation of gene set enrichment analysis through gamma distribution approximation. *Bioinformatics*, 38(8):2356–2357, 02 2022.
4. David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 2009.
5. Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14:1–15, 2013.
6. Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
7. Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.
8. Chengshang Lyu, Lingxi Chen, and Xiaoping Liu. Detecting tipping points of complex diseases by network information entropy. *Briefings in bioinformatics*, 25(4):bbae311, 2024.
9. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.
10. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.
11. Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.
12. Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.
13. Shi-Tao Yeh et al. Using trapezoidal rule for the area under a curve calculation. *Proceedings of the 27th Annual SAS® User Group International (SUGI'02)*, 4:1, 2002.
14. JA Ferreira and AH Zwinderman. On the benjamini–hochberg method. 2006.
15. Hervé Abdi et al. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3(01):2007, 2007.
16. Christopher W Woods, Micah T McClain, Minhua Chen, Aimee K Zaas, Bradly P Nicholson, Jay Varkey, Timothy Veldman, Stephen F Kingsmore, Yongsheng Huang, Robert Lambkin-Williams, et al. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza h1n1 or h3n2. *PloS one*, 8(1):e52198, 2013.
17. Tzu-Yu Liu, Thomas Burke, Lawrence P Park, Christopher W Woods, Aimee K Zaas, Geoffrey S Ginsburg, and Alfred O Hero. An individualized predictor of health and disease using paired reference and target samples. *BMC bioinformatics*, 17:1–15, 2016.
18. Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 05 2016.
19. M. Kanehisa. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(90001):277D–280, January 2004.
20. Charles Spearman. The proof and measurement of association between two things. 1961.