TITLE: Constructing Ensemble Gene Functional Networks Capturing Tissue/condition-specific Co-expression from Unlabled Transcriptomic Data with TEA-GCN

AUTHORS: Peng Ken Lim[1*], Ruoxi Wang[1], Jenet Princy Antony Velankanni[1], Marek Mutwil[1*]

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

*Corresponding author:
Marek Mutwil
School of Biological Sciences,
Nanyang Technological University,60 Nanyang Drive,
637551, Singapore,
Singapore
Email: mutwil@ntu.edu.sg

Peng Ken Lim
School of Biological Sciences,
Nanyang Technological University,60 Nanyang Drive,
637551, Singapore,
Singapore
Email: pengkenlim.sbs@gmail.com

Abstract

Gene co-expression networks (GCNs) generated from public transcriptomic datasets can elucidate the co-regulatory and co-functional relationships between genes, making GCNs an important tool to predict gene functions. However, current GCN construction methods are sensitive to the quality of the data, and the interpretability of the identified relationships between genes is still difficult. To address this, we present a novel method — Two-Tier Ensemble Aggregation (TEA-) GCN. TEA-GCN utilizes unsupervised partitioning of big transcriptomic datasets and three correlation coefficients to generate ensemble GCNs in a two-step aggregation process. We show that TEA-GCN outperforms in finding correct functional relationships between genes over the current state-of-the-art across three model species, and is able to not only capture condition/tissue-specific gene co-expression but explain them through the use of natural language processing (NLP). In addition, we found TEA-GCN to be especially performant in identifying relationships between transcription factors and their activation targets, making it effective in inferring gene regulatory networks. TEA-GCN is available at https://github.com/pengkenlim/TEA-GCN.

Introduction

Gene co-expression networks (GCNs) play a pivotal role in elucidating co-regulatory and co-functional relationships between genes[1–5]. This concept, rooted in the observation that functionally related genes often exhibit coordinated expression patterns, has proven invaluable in predicting gene function and uncovering modules of co-regulated genes[3–6]. The utility of GCNs to uncover biological insights becomes especially important in non-model species where experimental validation of gene function is scarce[6,7]. The rapidly increasing affordability of RNA-seq, coupled with the fast rate at which genomes are continuously published has led to an explosion of public transcriptomic data, particularly in non-model species[7–9], and offers an unprecedented opportunity to unravel the evolutionary dynamics of gene function across diverse organisms via comparative-GCN analysis[6,7,10,11].

However, the potential of GCNs is hindered by several challenges when working with public transcriptome data. The composition of transcriptome data can significantly impact the quality of GCNs and the context of biological insights embedded within them[1,1,12,13]. For example, the overrepresentation of certain sample types (e.g. tissues, conditions) within the dataset can selectively represent some tissue/condition-specific regulatory information while ignoring others[14]. Additionally, the size and heterogeneity of the transcriptome data can also have a huge impact on the resultant GCN. For example, dataset size (i.e. the number of samples) can directly affect the statistical power and therefore, the resolving power of GCNs to differentiate authentic transcriptional co-regulation of genes from spurious correlation of expression patterns[12]. On the other hand, GCN built from a highly heterogeneous dataset consisting of diverse sample types can excel in capturing ubiquitous gene modules but can fail to capture condition-specific regulatory relationships due to insufficient sample-type replication[14]. Thus, the variability of publicly available RNA-seq data of different species, in terms of sample quality, heterogeneity, and size poses a substantial hurdle for constructing performant GCNs. To address these challenges, many studies have utilized highly curated public transcriptomic datasets that are sample-balanced and/or context-specific to construct well-performing GCNs[15–19]. However, these methods rely on labor-intensive sample annotation that might not be available for some species.

Recent advancements in obtaining performant GCNs have seen the use of ensemble approaches that combine (aggregate) multiple GCNs to enrich biological information[20–22]. In one study, gene expression omnibus (GEO) datasets of *Vitis vinifera* (Grape) were used to generate unweighted GCNs and aggregate them to form an ensemble GCN[21]. Aside from producing a more

robust GCN, this approach has achieved some degree of sample balancing as coexpression was determined from context-(GEO dataset)-specific GCNs rather than directly calculated across all samples. However, this approach might not work for cases where the number of GEO datasets is limited, even for species that possess numerous public RNA-seq accessions. Another approach (henceforth referred to as the "Subbaging" method) combined weighted GCNs calculated from random subsets of principal components of the gene expression matrix[16,17]. However, the stochastic nature of this approach due to random subsampling, as well as the ambiguous meaning of principal components, compromises the interpretability of the resultant ensemble GCN. In addition, both approaches only use Pearson's correlation coefficient (PCC), which can fail to consider non-linear co-expression relationships between genes[1,23].

In response to these challenges, we present a novel method — Two-Tier Ensemble Aggregation (TEA-) GCN. The TEA-GCN framework leverages unsupervised partitioning of publicly derived transcriptomic data and utilizes three correlation coefficients to generate ensemble CGNs in a two-step aggregation process (Figure 1). Importantly, the TEA-GCN approach does not require annotated samples for batch correction and/or sample balancing, making it readily applicable to large public transcriptomic datasets with limited preprocessing and suitable for species with scarce sample annotation. We evaluated the performance of the TEA-GCN in uncovering diverse biological relationships between genes and demonstrated its superior performance over the state-of-the-art "subbagging" ensemble GCN method[16,17] in *Saccharomyces cerevisiae* (*S. cerevisiae*; Baker's yeast), *Arabidopsis thaliana (A. thaliana*; Thale cress*)* and *Homo sapiens* (*H. sapiens*; Human*),* three model organisms of vastly different biological/organizational complexities and evolutionary lineages. Using *A. thaliana* as a model, we also demonstrated the framework's ability to capture condition/tissue-specific gene co-expression and characterized its flexibility to scale to smaller transcriptomic datasets. In addition, we found TEA-GCN to be especially performant in identifying relationships between transcription factors and their activation targets that can rival the well-established GENIE3 algorithm[24]. Finally, we show that TEA-GCN enables explainability through the use of natural language processing (NLP) to reveal the experimental contexts underpinning the co-expression relationships. Put together, we present TEA-GCN as not only a more robust method to generate ensemble GCNs than the current state-of-the-art but also a landmark method to generate explainable ensemble GCNs. TEA-GCN is available at https://github.com/pengkenlim/TEA-GCN .

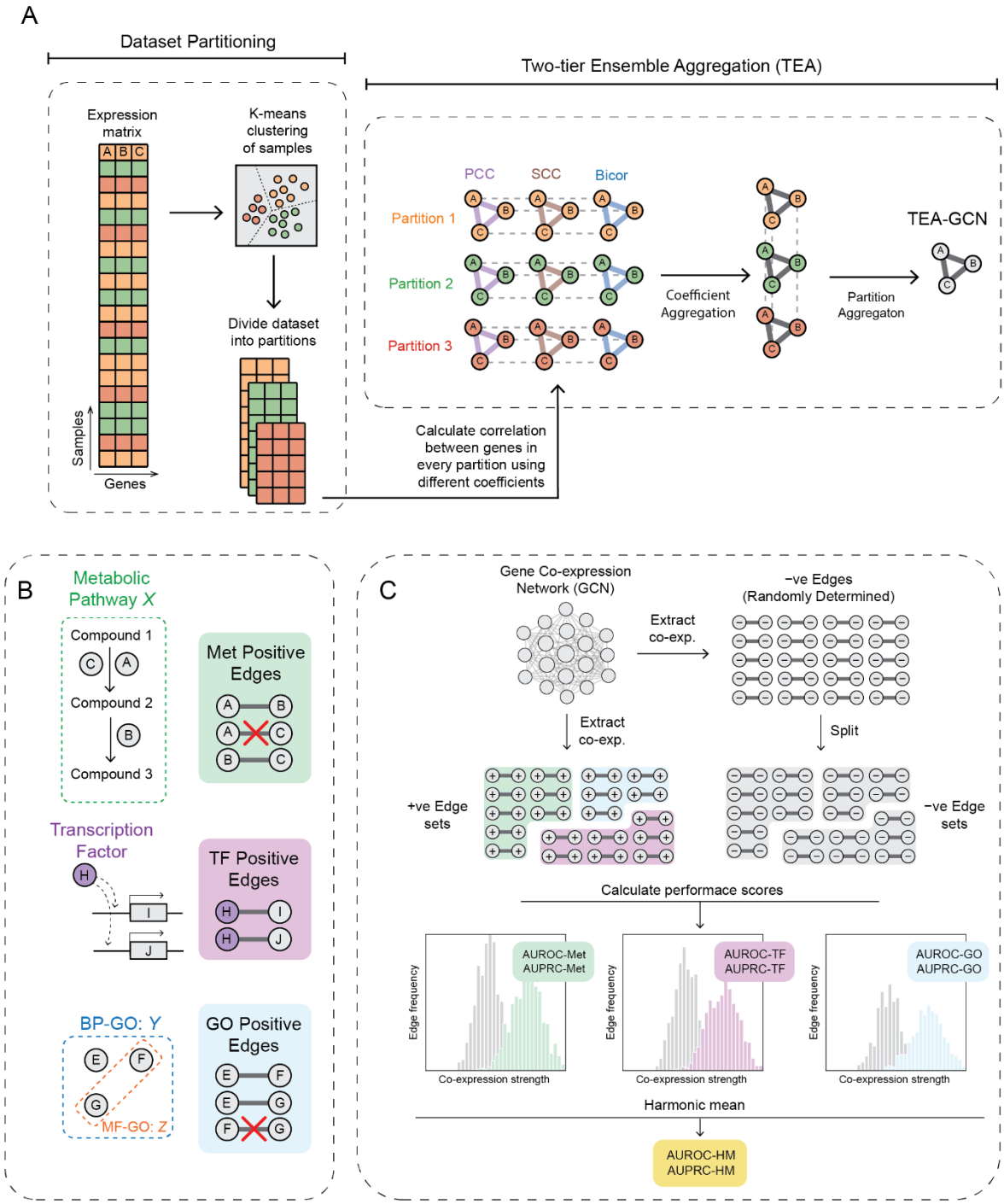


Figure 1. Overview of Two-tier Ensemble Aggregation Gene Co-expression Network (TEA-GCN) framework and GCN evaluation methodology. A) The processing of data from an initial gene expression matrix to the final ensemble GCN is visualized using a toy model for which there are only three genes (A, B, C) in the genome. The gene expression matrix is first partitioned via k-means clustering at a clustering granularity (k=3) that has been determined using a combination of semi-supervision and/or internal clustering evaluation metrics. For each dataset partition, GCNs are generated based on three different correlation coefficients and aggregated into one ensemble GCN. Resultant GCNs from each partition are again aggregated to yield the final ensemble TEA-GCN. GCNs with coexpression strength represented by Pearson's correlation coefficient (PCC), Spearman's rank correlation coefficient (SCC), and Biweight midcorrelation (Bicor) are denoted by purple, brown, and blue edges, respectively. The node color of GCNs corresponds to the particular dataset partition they represent. **B)** We define three sets of positive edges connecting genes sharing different biological relationships. Positive Metabolic (Met) edges are made by pairing experimentally validated enzymes from the same

pathway. Edges connecting enzymes that catalyze the same reaction step are removed from positive Met edges. Positive Transcription factor (TF) edges are generated between genes and their corresponding activator TFs. Positive Gene Ontology (GO) edges are made from pairing genes that were annotated for the same biological process (BP) GO term at the IDA (inferred from direct assay) evidence level. Edges connecting genes annotated with the same molecular function (MF) GO term are removed from positive GO edges. **C)** To evaluate a given GCN, co-expression strengths of positive edge sets for each biological aspect are compared to co-expression strengths of randomly determined negative edges of the same size using the Area-under-curve (AUC) of the receiver operator characteristics curve (ROC) and precision-recall curve (PRC). Scores calculated for each positive-negative edge set pair represent a GCN's performance to identify different biologically significant relationships between genes and can be aggregated using harmonic mean to infer a GCN's overall performance.

121 # Results

122 **Overview of Two-tier Ensemble Aggregation Gene Co-expression Network (TEA-**
123 **GCN) Framework**

124 TEA-GCN relies on the Two-tier Ensemble Aggregation (TEA) process consisting of two tandem
125 aggregation steps, Coefficient Aggregation and Partition Aggregation (Figure 1A). Coefficient
126 Aggregation involves the calculation of different co-expression metrics (coefficients) between a
127 given pair of genes and then unifying (aggregating) said metrics into one measure of co-
128 expression strength (Figure 1A). We used the three most popular metrics: Pearson's correlation
129 coefficient (PCC, capturing linear relationships), Spearman's rank correlation coefficient (SCC,
130 capturing monotonic relationships)[1,25,26] and biweight midcorrelation[27] (Bicor), a coefficient
131 employed in the widely used WGCNA R package[28] that is more resistant to outliers and noise as
132 compared to PCC[27]. Although there have been several studies comparing the three coefficients,
133 there is no current consensus as to which coefficient is the most ideal for generating robust GCNs
134 as different coefficients have been shown to excel in detecting different biological relationships[26].
135 Therefore, to detect more types of complex yet biologically significant co-expression relationships
136 between genes, the TEA process uses Coefficient Aggregation to derive co-expression based on
137 all three coefficients.

138 Partition Aggregation aims to capture tissue/condition-specific GCNs that might be missed
139 if all samples are used for the network construction. The tissue/condtion-specific networks have
140 been shown to capture relevant biological relationships that only appear in certain spatiotemporal
141 settings[15,18,19], or present in response to certain environmental conditions[29–31]. However, the
142 sample composition of publicly available RNA-seq samples is expected to be biased towards
143 certain sample types (e.g., easily attainable leaves in plants and certain popular mammalian cell
144 lines), which might miss co-expression relationships between underrepresented samples. To
145 address this, the TEA-GCN framework uses K-means clustering[32], an unsupervised clustering
146 algorithm to partition the transcriptomic dataset into samples with similar transcriptome profiles[33]
147 (Figure 1A). Then, the co-expression strengths between all gene pairs are aggregated across all
148 partitions to detect tissue/condition-specific co-expression. Our approach combines the
149 Coefficient and Partition aggregation in two steps. In the first step, Coefficient Aggregation
150 combines the three coefficients in every partition and in the second step, Partition Aggregation
151 combines the coefficients from the different partitions (Figure 1A).

To benchmark the TEA-GCN framework against the state of the art, we investigated whether the obtained networks can predict biologically significant relationships between genes (Figure 1B and C). To this end, we investigated whether the networks can recover experimentally-verified metabolic networks (Met) (Figure 1B; green), transcription factor-target relationships (TF) (Figure 1B; purple) and genes involved in the same biological process (GO) (Figure 1B; blue). We obtained the experimentally verified relationships (positive edges) for Met, TF and GO categories for the three species (see methods, Table S1 to S9). Since there is a tendency for co-linearity between the estimated expression abundances of sequence-similar homologs due to how multi-mapping reads are accounted for in pseudoalignment[34], edges that connect genes annotated with the same molecular function were removed from GO positive edges to prevent these edges from inflating performance scores (Figure 1B; blue). Similarly, we removed edges that connect genes coding for enzymes that catalyze the same reaction steps from Met positive edges. Using the co-expression strengths extracted for different sets of positive edges and negative edges (representing genes with no functional relationship, obtained from sampling random gene pairs), the Area-under-curves of the receiver operator characteristics curve (AUROC) and precision-recall curve (AUPRC) performance scores were calculated for each biological aspect (e.g. AUROC-Met and AUROC-Met; Figure 1C). We aggregated the scores with harmonic mean to infer a GCN's overall performance (i.e., AUROC-HM and AUROC-HM; Figure 1C). To account for the random effect on performance scores due to the random selection of negative edges, median scores from 100 evaluation iterations are reported.

173

174 **Coefficient Aggregation and Partition Aggregation improves GCN performance**

175 The TEA process uses different aggregation functions for Coefficient Aggregation and Partition
176 Aggregation. Coefficient Aggregation employs the Maximum (Max) function which outputs the
177 highest correlation magnitude reported by three different coefficients (PCC, SCC, bicor). In
178 contrast, Partition Aggregation employs the Rectified Average (RAvg) function to aggregate co-
179 expression strengths from different dataset partitions. The RAvg function takes in co-expression
180 strengths from every dataset partition and converts all negative co-expression strengths into zero
181 before averaging. Effectively, the RAvg function considers negative co-expression, which results
182 from the inverse correlation between gene expression, as the absence of co-expression (i.e. zero
183 co-expression).

184 The TEA process relies on k-means clustering[32] to generate dataset partitions from public
185 transcriptome datasets[33]. However, the optimal number of partitions (i.e. the number of
186 biologically relevant subsets of the data) is unknown. Thus, we explored a range of partitions (*k*)
187 for publicly available data for *S. cerevisiae*, *A. thaliana*, and *H. sapiens*. After removing samples
188 with poorly mapped reads (see methods), we obtained 150,096, 57,759, and 95,553 RNA-seq
189 samples for *S. cerevisiae*, *A. thaliana*, and *H. sapiens,* respectively (Table S10 to S12). To
190 investigate how partitioning granularity (*k*) affects the performance of the GCN produced by TEA,
191 we tested preselected *ks* ranging from 2 to 1000 (see methods; Table S13 to S15) (Figure 1C).

192 Across all three organisms, the performance (measured by AUROC-HM and AUPRC-HM)
193 of the TEA generally increases with partitioning granularity and saturates quickly over a large
194 window of *k* (Figure 2A; green line). The ensemble GCN inferred by the TEA can outperform GCN
195 obtained from calculating PCC from all samples (PCC GCN) at every *k* for *S. cerevisiae* and *A.*
196 *thaliana* (Figure 2A; blue dotted baseline). Although the outperformance of TEA for *H. sapiens* is
197 not observed for every *k* (low AUCPRC-HM at *ks* < 248 and > 902), TEA still shows good
198 performance for the majority (approx. 69%, 248 to 902 *k*) of the sampled *ks* (2 to 1000). More
199 importantly, the performance of TEA is stable over a wide range of *k* and does not decay
200 substantially even at high *k*, which suggests the robustness of TEA to resist over-partitioning
201 (Figure 2A). Thus, the effectiveness of TEA can be achieved by testing the different *k* values or
202 by using a sufficiently large *k* (identified using internal clustering metrics like silhouette coefficient)
203 that does not under-partition the public transcriptome dataset (Figure 1B).

204 To investigate whether aggregating different coefficients (i.e. Coefficient Aggregation)
205 improves the performance, we evaluated Partition Aggregation-only GCNs using only one of the

206 three coeffieicent (PCC, SCC, or Bicor) (Figure 2B) over the same preselected *ks* for each species
207 (Figure S1). Curiously, we found that the optimal *k* (*k* at which the sum of median AUROC-HM
208 and AUPRC-HM is highest) is dependent on the coefficient. For example, optimal *k*s for *A.*
209 *thaliana* Partition Aggregation-only GCNs were 62, 444 and 950 when PCC, SCC and Bicor were
210 used, respectively. As such, median performance scores of various Partition Aggregation-only
211 GCNs (PCC-RAvg, SCC-RAvg, and Bicor-RAvg) were compared to GCNs generated by the TEA
212 process (Max-RAvg) at their respective optimal *k*s (Figure 2C). We found that Max-RAvg GCNs
213 achieves median performance scores (AUROC-HM, AUPRC-HM) of (0.650, 0.633), (0.659,
214 0.670), and (0.554, 0.554) for *S. cerevisiae*, *A. thaliana*, and *H. sapiens* respectively (Figure 2C),
215 which is higher than the performance scores of (0.532, 0.529), (0.602, 0.622), and (0.506, 0.545)
216 of their corresponding PCC GCNs (Figure 2A). Crucially, we show that Max-RAvg outperforms
217 Partition Aggregation-only GCNs across all species in AUROC-HM and AUPRC-HM (Figure 2C,
218 heatmap section 2), which demonstrates that the combination of Coefficient Aggregation and
219 Partition Aggregation results in more performant GCNs, as compared to only using Partition
220 Aggregation alone.
221 The performance of the GCNs can be further improved by transforming the co-expression
222 strengths to rank-based metrics, such as Mutual Ranks (MR)[35]. GCNs based on MR-transformed
223 co-expression values of Max-RAvg (Max-RAvg-MR) further increased performance across all
224 biological aspects in the three species (Figure 2C). Finally, we also explored whether other
225 aggregation functions influence the network performance. To this end, we compared the
226 performances of GCNs generated by Max-Avg and Max-RAvg, which use Avg and RAvg
227 functions, respectively, for Partition Aggregation. The performance differences were minor before
228 MR transformation (Figure 2C, heatmap section 1). However, we found that RAvg for Partition
229 Aggregation yielded more performant GCNs after MR transformation (i.e. Max-Avg-MR and Max-
230 RAvg-MR) in *S. cerevisiae* and *A. thaliana*, while performance between the two variants remained
231 similar in *H. sapiens* (Figure 2C, heatmap section 2). Therefore, removing negative correlation
232 values improved the performance of the GCN and Max-RAvg-MR (Coefficient Aggregation →
233 Partition Aggregation → Mutual Rank transformation) was chosen as the most performant
234 approach for the rest of the study and designated as the TEA-GCN method.

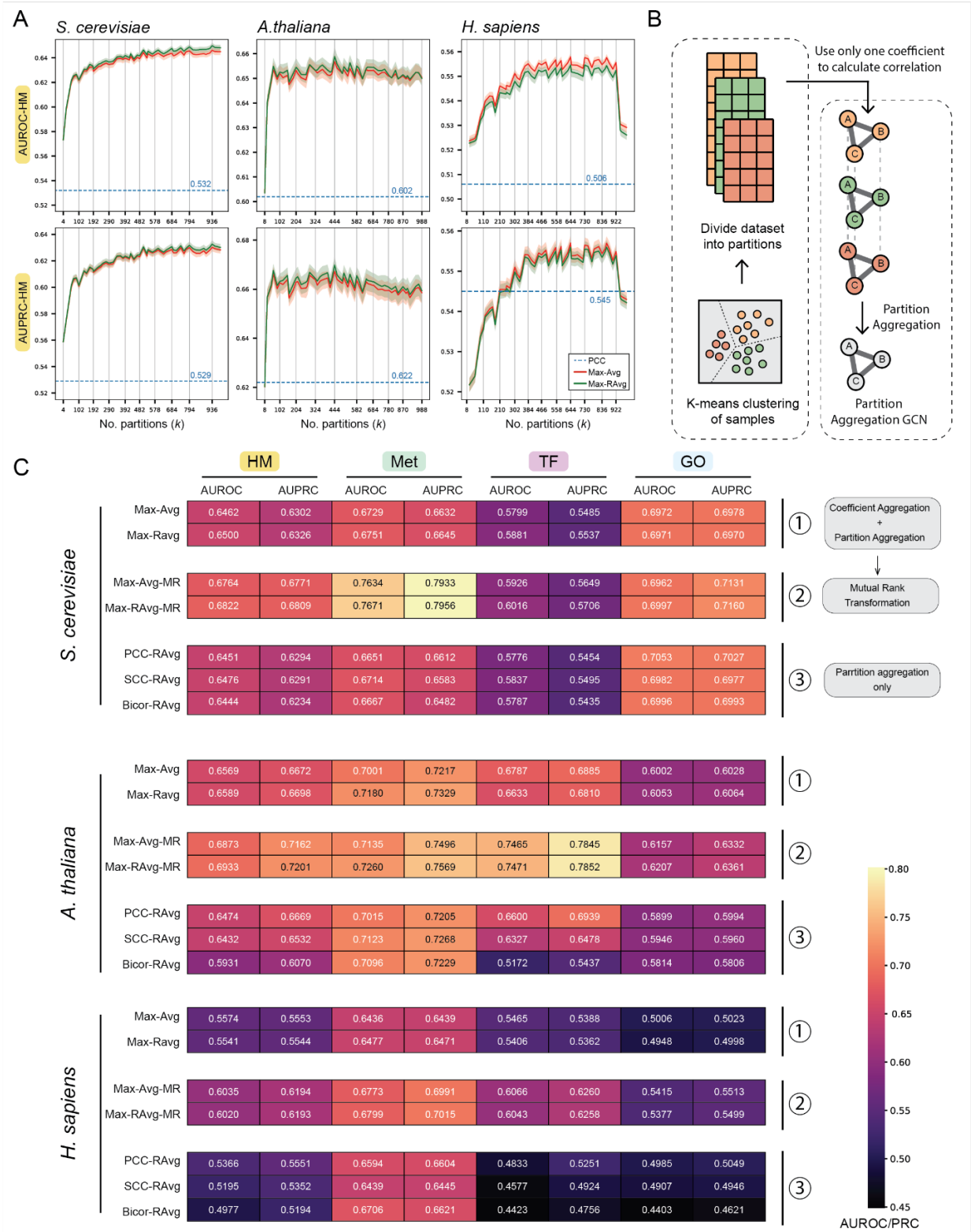


Figure 2. Benchmarking performance of TEA-GCN across Yeast, Arabidopsis and Human. A) Performance of TEA-GCNs in *S. cerevisiae*, *A. thaliana* and *H. sapiens* across different levels of partitioning granularity (*k*). The colors of solid lines correspond to the median performance scores of TEA-GCNs across 100 evaluations using different aggregation functions for Partition aggregation. Shaded regions denote upper and lower quartiles of performance scores. The blue baseline corresponds to median performances of non-ensemble GCN using Pearson's correlation coefficient (PCC). **B)** Performance of the Partition aggregation-only GCN generated

from one correlation metric. For each dataset partition, GCNs were generated based on the selected metric and subsequently aggregated across partitions. **C)** Comparison of median harmonic mean-aggregated performances and biological aspect-specific performances between TEA-GCNs (section 1), TEA-GCNs with mutual rank transformation (section 2) and Partition aggregation-only GCNs (section 3) across the three different model organisms.

245 **TEA-GCN achieves better overall performance over the state-of-the-art**

246 To benchmark the performance of the TEA-GCN (i.e. Max-RAvg-MR), we compare the
247 performances of *S. cerevisiae*, *A. thaliana*, and *H. sapiens* GCNs generated by our approach to
248 networks from two industry-leading databases, COXPRESdb[16] (https://coxpresdb.jp/download/)
249 and ATTED-II[16] (https://atted.jp/download/). According to the authors, the *A. thaliana* ATTED-II
250 GCN, *S. cerevisiae* COXPRESdb GCN, and *H. sapiens* COXPRESdb GCN were generated by
251 first constructing gene expression matrices of selected public batch-corrected RNA-seq samples.
252 The GCNs were subsequently calculated from subsampled principal components of the matrices,
253 and aggregated into one GCN[16,17]. As expected, these "Subbaging" ensemble GCNs have a
254 better overall performance than non-ensemble PCC GCNs across all three species (ΔAUROC-
255 HM and ΔAUPRC-HM > 0.06-0.16) (Figure 3A-B; Table S16 to 18). However, GCNs based on
256 our approach showed a substantially higher performance for nearly all metrics (Figure 3A, 3B,
257 and S2; Table S16 to 18).

258 To better understand the differences between our method and the state-of-the-art, we
259 compared the median AUROC and AUPRC scores and their respective curves for TF, Met and
260 GO-BP biological aspects of TEA-GCN and ATTED-II/COXPRESdb GCNs. TEA-GCNs were
261 consistently more performant in detecting the experimentally verified transcription factor-target
262 regulatory relationships (TF) across the three species (Figure 3C and 3D; Table S16 to 18),
263 especially in *A. thaliana* and *H. sapiens.* To exemplify, *A. thaliana* TEA-GCN has a TF
264 performance (AUROC-TF, AUPRC-TF) of (0.747, 0.785) as opposed to (0.680, 0.715) in *A.*
265 *thaliana* ATTED-II GCN, while *H. sapiens* TEA-GCN has a TF performance of (0.604, 0.626) as
266 opposed to (0.542, 0.587) in *H. sapiens* COXPRESdb GCN (Figure 3C and 3D; Table S16 to 18).
267 Conversely, the ATTED-II/COXPRESdb GCNs perform better in identifying enzymes involved in
268 the same metabolic pathway for *S. cerevisiae* and *H. sapiens*. *A. thaliana* TEA-GCN has a slightly
269 higher Met performance (AUROC-Met, AUPRC-Met of 0.726, 0.757) than *A. thaliana* ATTED-II
270 (0.711,0.749), and better performance in identifying genes sharing the same GO-BP term in *S.*
271 *cerevisiae* (0.700, 0.716 vs 0.673, 0.712) and *A.thaliana* (0.621,0.636 vs 0.599, 0.719) (Figure
272 3C and 3D; Table S16 to 18).

273 We speculate that the superior performance of TEA-GCN lies in its ability to detect
274 condition/tissue-specific co-expression from individual dataset partitions. As such, the observed
275 TF performance disparity between TEA-GCN and the "Subbaging" method is more pronounced
276 in multicellular *A. thaliana* and *H. sapiens* than in unicellular *S. cerevisiae* (Figure 3C and 3D;

277 Table S16 to 18), could be attributed to the presence of more specialized transcriptional programs
278 in multicellular organisms than unicellular organisms[36,37].

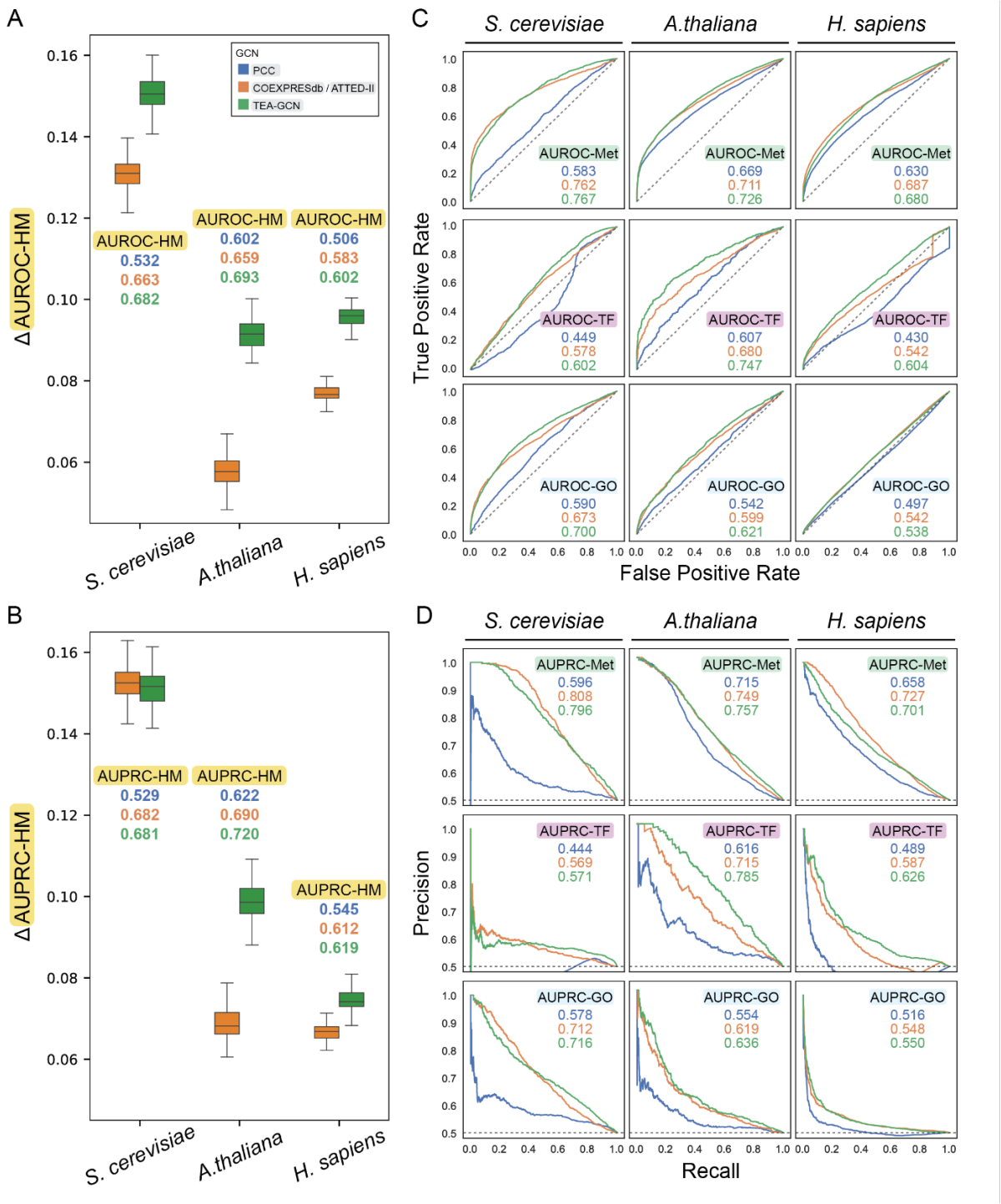


Figure 3. Benchmarking TEA-GCN across three different model organisms against "Subbaging" ensemble GCNs. Distribution of **A)** AUROC-HM performance scores and **B)** AUPRC-HM performance scores between Subbaging CGNs and TEA-GCNs across *S. cerevisiae*, *A. thaliana* and *H. sapiens*. We report delta values of performance scores (i.e. Δ AUROC-HM and Δ AUPRC-HM) used to compare GCNs on a unified axis. The delta values were calculated by subtracting performance scores with the median performance score of baseline non-ensemble PCC network of respective organisms. Median performance scores of PCC (blue) networks, Subbaging CGNs (orange) and TEA-GCNs (green) are indicated. **C)** Receiver operating characteristics curves (ROCs) and **D)** Precision-recall curves (PRCs) correspond to median performance scores for each biological aspect and organism. Across all panels, performance information relating to the PCC network, Subbaging CGN and TEA-GCN are represented in blue, orange and green, respectively.

289 **TEA-GCN captures co-expression between experimentally validated enzymes in**
290 ***Arabidopsis thaliana* metabolic pathways**

291 To characterize how well the TEA-GCN framework captures co-expression between enzymes of
292 the same pathway, we compared the co-expression strengths of Met positive edges in *A. thaliana*
293 TEA-GCN and ATTED-II GCN. We define Met positive edges as edges that connect
294 experimentally validated enzymes within individual PlantCyc pathways[38]. For a fair comparison of
295 co-expression strengths across the two CGNs, we used standardized co-expression strengths
296 (***Z***(Co-exp.)), which measure the co-expression strengths of each edge relative to the co-
297 expression strengths of all edges in the GCN (see methods). *A. thaliana* TEA-GCN performed
298 better in identifying Met positive edges over the ATTED-II GCN across all major classes (Figure
299 4A). The exception was the "Generation of Precursor Metabolites & Energy" class, where the
300 performance between the two GCNs is nearly identical (Figure 4A; Table S19), irrespective of co-
301 expression thresholds (i.e. ***Z***(Co-exp.) ≥ 0.5 and ***Z***(Co-exp.) ≥ 1). On the other hand, TEA-GCN
302 excelled in the "Metabolic Clusters" and "Biosynthesis" classes, where the percentage of positive
303 edges at both thresholds is more than 10% higher than in ATTED-II GCN (Figure 4A; Table S19).
304 The "Generation of Precursor Metabolites & Energy" class consists of conserved pathways such
305 as the TCA cycle, glycolysis, respiration, and photosynthesis that are ubiquitously expressed and
306 not likely condition/tissue-specific, which could explain the less pronounced performance of our
307 approach. Conversely, "Biosynthesis" and "Detoxification" classes, which tend to comprise
308 specialized pathways that are lineage-specific and/or condition/tissue-specific[39–44], has more co-
309 expressed edges in TEA-GCN than ATTED-II GCN at both thresholds. As such, this observation
310 is compatible with our hypothesis that the good performance of TEA-GCN could be due to its
311 ability to capture condition/tissue-specific co-expression.

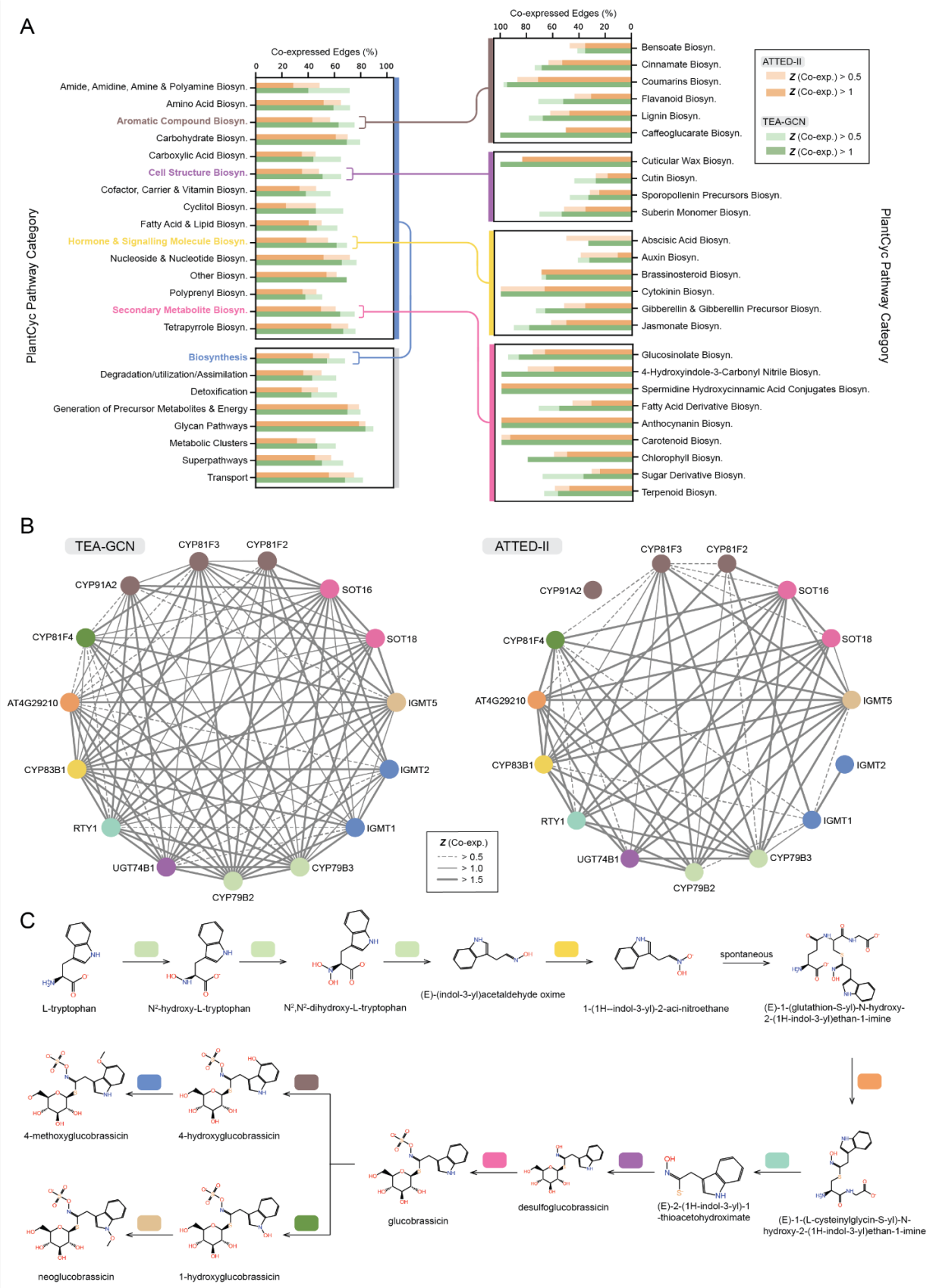


Figure 4. Characterizing TEA-GCN's ability to capture metabolic pathways in *A. thaliana*. A) Proportion of Met positive edges

of different PlantCyc pathways that are co-expressed in *A. thaliana* TEA-GCN and ATTED-II GCN at two (>0.5 and >=1) co-expression thresholds. Met positive edges connect experimentally validated enzymes of different reaction steps. Grey labeled plot corresponds to first-level ontological classes and compares between major pathway categories. The blue labeled plot breaks down second-level ontological classes within "Biosynthesis", of which third-level ontological classes consisting of "Aromatic Compound Biosynthesis", "Cell Structure Biosynthesis", "Hormone and Signalling Molecule Biosynthesis", and "Secondary Metabolite Biosynthesis" were further broken down in the brown, purple, yellow and pink plots, respectively. **B)** Co-expression between enzymes of "Glucosinolate Biosynthesis from Tryptophan" PlantCyc pathway (PWY-601) in *A. thaliana* TEA-GCN and ATTED-II GCN. Nodes represent experimentally validated enzymes according to PlantCyc and are labeled with gene symbols (if available) or gene identifiers. The thickness of grey edges connecting nodes corresponds to standardized co-expression strengths between enzymes in the respective GCNs. **C)** Reaction steps in Glucosinolate Biosynthesis from Tryptophan PWY-601. Colored blocks in the middle of reaction steps annotate enzyme nodes in panel B based on the reaction step they participate in.

To better compare the *A. thaliana* GCNs of TEA-GCN and ATTED-II, we investigated the different sub-classes within the "Biosynthesis" class. TEA-GCN identified at least 10% more co-expressed edges for "Hormone and Signaling Molecule Biosynthesis", "Cyclitol Biosynthesis", "Aromatic Compound Biosynthesis", "Cell Structure Biosynthesis", "Secondary Metabolite Biosynthesis", "Amide, Amidine, Amine, and Polyamine Biosynthesis", at both co-expression thresholds (Figure 4A, Blue). Impressively, we observed an extra 23% of positive met edges belonging to "Hormone and Signaling Molecule Biosynthesis" for TEA-GCN (61.5%; 123 out of 200) compared toATTED-II GCN (38.5%; 77 out of 200) (Figure 4A, blue; Table S19) at the stricter co-expression threshold of ***Z***(Co-exp.) ≥ 1. This is largely attributed to the "Gibberellin and Gibberellin Precursor Biosynthesis", and "Jasmonate Biosynthesis" (Figure 4A, yellow) that collectively account for more than half of the edges belonging to "Hormone and Signaling Molecule Biosynthesis" (Table S19). Given that the biosynthesis of hormones is known to be highly localized to certain tissues at specific stages of development or environmental conditions[45–47], we attribute the higher performance to TEA-GCN's ability to capture condition-specific co-expression. Other notable classes where TEA-GCN excels include "Fatty Acid Derivative Biosynthesis" (TEA-GCN: 55.97% and ATTED-II: 31.45%, strict threshold; 159 total edges), "Flavonoid Biosynthesis" (TEA-GCN: 51.97% and ATTED-II: 30.71%, strict threshold; 127 total edges), "Lignin Biosynthesis" (TEA-GCN: 67.48% and ATTED-II: 47.15%, strict threshold; 123 total edges) and "Glucosinolate Biosynthesis" (TEA-GCN: 86.74% and ATTED-II: 66.85%, strict threshold; 181 total edges) (Figure 4A; Table S19).

To exemplify the utility of TEA-GCN in predicting pathway memberships of enzymes, we compared the co-expression strengths between genes that code for experimentally validated enzymes participating in the "Glucosinolate Biosynthesis from Tryptophan" PlantCyc pathway (henceforth referred to by its corresponding PlantCyc pathway accession, PWY-601; Figure 4B and C). Glucosinolates are a class of specialized natural compounds synthesized by plants in the Capparales order that consists of many agriculturally valuable crops in the Brassicaceae family like *Brassica oleracea* (*B. oleracea*; cabbage, kale, and broccoli), *Brassica juncea (B. juncea*;

353 chinese mustard) *Brassica napus* (*B. napus*; rapeseed) and *Brassica rapa* (*B. rapa*; turnip and
354 field mustard)[48–50]. Glucosinolates and their breakdown products possess a plethora of
355 bioactivities, some of which can be toxic (e.g. goitrogenic and genotoxic) and others beneficial
356 (e.g. anti-cancer)[48–50]. As such, there is a significant research interest in transgenically modifying
357 the level of specific glucosinolates in Brassicaceae food crops[51], which involves the the
358 elucidation of enzymes involved in the biosynthesis of glucosinolate pathways. The PWY-601
359 pathway converts L-tryptophan to indole glucosinolates such as Glucobrassicin (3-IMG),
360 Neoglucobrassicin (MIMG), and 4-methoxyglucobrassicin (4-MoIMG), which have anti-
361 inflammatory and anti-cancer activities in humans[48,52–54]. We show that experimentally validated
362 enzymes of PWY-601 tend to be co-expressed with more experimentally validated enzymes from
363 the same pathway in *A. thaliana* TEA-GCN than in ATTED-II GCN (Figure 4B), which
364 demonstrates the superior ability of TEA-GCN to predict enzymes involved in the said pathway.
365 Furthermore, genes coding for enzymes involved in the derivatization of 3-IMG to 4-MoIMG (i.e.
366 *IGMT1*, *IGMT2*, *CYP91A2*, *CYP81F3*, and *CYP81F2*) are notably less co-expressed with other
367 PWY-601 enzymes in ATTED-II GCN, as opposed to the more connected PWY-601 network of
368 TEA-GCN (Figure 4B and C). This scarce connection is especially prominent for *CYP91A2* and
369 *IGMT2*, which codes for monooxygenase and methyltransferase that catalyze the hydroxylation
370 of 3-IMG to 4-hydroxylglucobrassicin (4-OhIMG) and the methylation of 4-OhIMG to 4-MoIMG,
371 respectively[55]; *CYP91A2* is not co-expressed with any genes in PWY-601, while *IGMT2* is only
372 connected to *IGMT1* which codes for a methyltransferase catalyzing the same reaction step
373 (Figure 4B). To summarize, we have shown that TEA-GCN can better capture co-expression
374 between genes coding for enzymes within specialized metabolic pathways than the state-of-the-
375 art “Subbagging” method, possibly due to TEA-GCN’s ability to detect tissue/conditional-specific
376 gene co-expression.

377 **TEA-GCN can accurately reveal condition/tissue-specific gene regulatory networks**

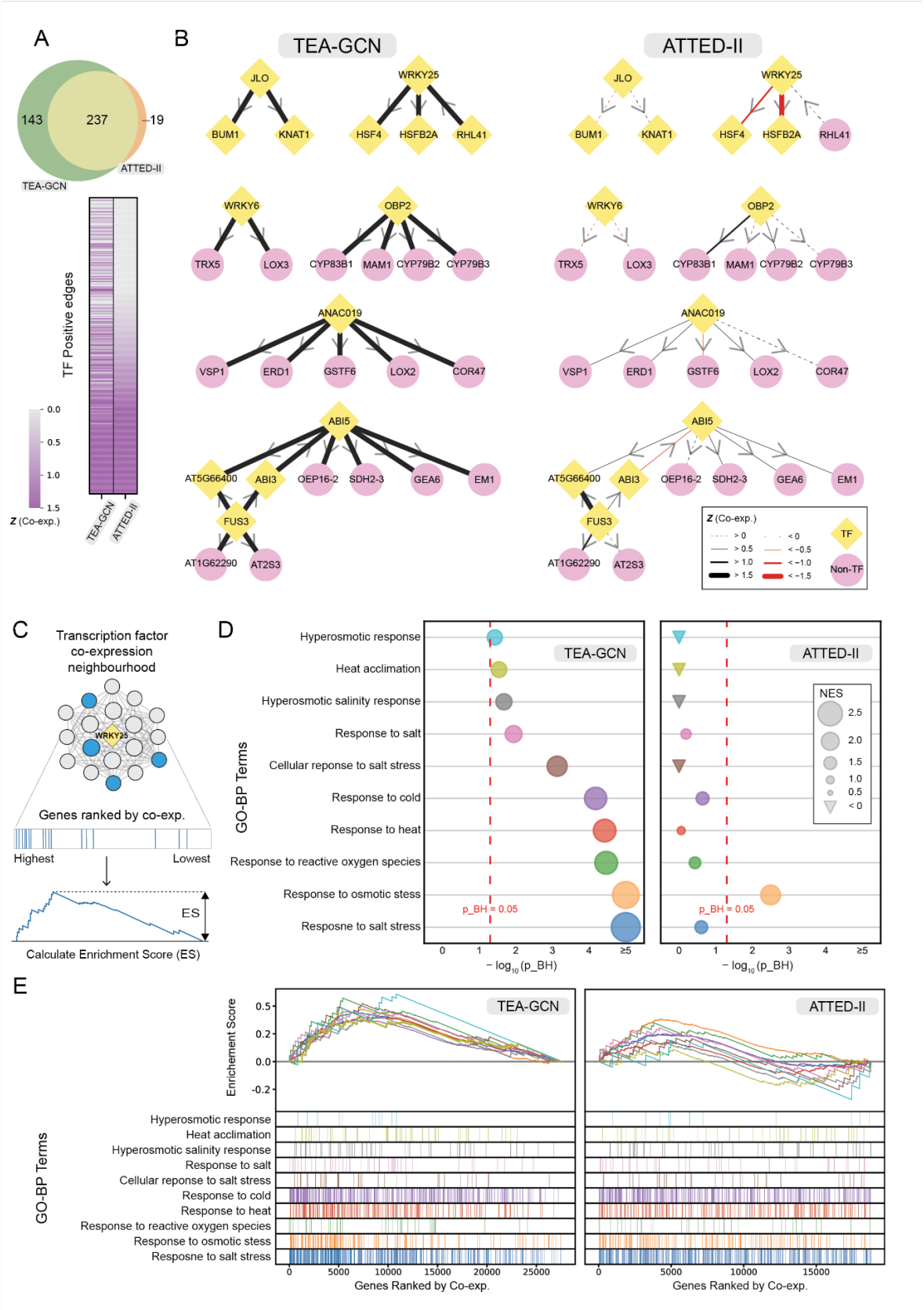
378 Our analysis revealed that TEA-GCN is particularly performant in revealing transcription factors
379 and their gene targets (Figure 3C and D; TF aspect). While a large portion of TF positive edges
380 is co-expressed in both GCNs (237 out of 673, 35.2%, Figure 5A), a substantial portion (143 out
381 of 673, 21.2%; Figure 5A) was found only in TEA-GCN. Conversely, only 19 TF positive edges
382 were uniquely co-expressed in ATTED-II GCN (2.8%; Figure 5A). Thus, TEA-GCN contains
383 nearly all edges found in ATTED-II and can identify additional positive edges.

384 To further characterize the difference between the gene regulatory networks identified by
385 TEA-GCN and ATTED-II, we analyzed transcription factors where TEA-GCN, but not ATTED-II,
386 showed a high co-expression (***Z***(Co-exp.) >1.5) with their transactivation gene targets (Figure
387 5B). The identified transcription factor genes are known to be highly tissue/condition-specific in
388 their expression, which further supports our hypothesis that TEA-GCN excels in detecting
389 tissue/condition-specific co-expression. To exemplify, *JAGGED LATERAL ORGANS*'s (*JLO*)
390 expression is specific to embryos[56], while *ABA INSENSITIVE 5*'s (*ABI5*) and *FUSCA3*'s (*FUS3*)
391 expression are seed-specific[57,58]. On the other hand, *OBF-BINDING PROTEIN 2*'s (*OBP2*)
392 expression is known to be induced by wounding and phytohormone Methyl jasmonate (MeJA)[59] ,
393 *NAC DOMAIN CONTAINING PROTEIN 19*'s (*ANAC019*) expression is induced by drought, high-
394 salt stress and phytohormone abscisic acid[60], and *WRKY6*'s expression is induced by phosphate-
395 starvation[61]. Curiously, *WRKY25* is found by ATTED-II GCN to be negatively co-expressed with
396 its transactivation targets of *HEAT SHOCK FACTOR 4* (*HSF4*) and *HEAT SHOCK*
397 *TRANSCRIPTION FACTOR B2A* (*HSFB2A)*[62] (Figure 5B). *WRKY25* is known to be involved
398 in responses against a multitude of abiotic stresses such as heat, cold, osmotic, and high-salt
399 conditions[62–65]. Given the multimodal functions of *WRKY25* and its participation in various abiotic
400 responses, we speculate that its mode of action might be condition-specific, which might be the
401 reason why its co-expression with *HSF4* and *HSFB2A* were not detected in ATTED-II GCN.

402 Next, to investigate whether the transcription factors are co-expressed with biologically
403 relevant genes, we compared the functional enrichment of genes co-expressed with the selected
404 transcription factors for *A. thaliana* TEA-GCN and ATTED-II GCN. To this end, we encapsulated
405 the co-expression neighborhoods of transcription factors as lists of genes ranked by co-
406 expression and used them as input for Gene Set Enrichment Analysis (GSEA)[66,67] using GO-BP
407 gene sets (Figure 5C). Conceptually, if genes involved in a particular biological function are
408 overrepresented in highly ranked genes (i.e. strongly co-expressed with the transcription factor),

409 the calculated enrichment score (ES) for the biological function will be large and statistically
410 significant (Figure 5C). The statistical significance of said overrepresentation can be determined
411 in the form of a null-hypothesis p-value (p_BH, after multiple-testing correction using the
412 Benjamini-Hochberg method) by comparing ESs calculated from permuted rankings (permuted
413 ESs) to the ES calculated from the original ranking[66], while the magnitude of overrepresentation
414 can be inferred from the Normalized Enrichment Score (NES) calculated as the fold-change of
415 ES from the mean of permuted ESs[66].

416 Convincingly, we show that the functional enrichment profiles of co-expression
417 neighborhoods of selected transcription factors are significant for many of their known biological
418 functions in TEA-GCN, but not for ATTED-II (Figure 5D and E, Figure S3). For example, we show
419 genes involved in "Hyperosmotic response" (GO:0006972), "Hyperosmotic salinity response"
420 (GO:0042538), "Response to salt" (GO:1902074), "Cellular response to salt stress"
421 (GO:0071472), "Response to salt stress" (GO:0009651), "Heat acclimation" (GO:0010286),
422 "Response to heat" (GO:0009408), "Response to cold" (GO:0009409) and "Response to reactive
423 oxygen species" (GO:0000302) are enriched (p_BH ≥ 0.05) in the *WRKY25* co-expression
424 neighborhood of TEA-GCN but not ATTED-II GCN (Figure 5D and E). Although both *WRKY25*
425 co-expression neighborhoods from *A. thaliana* TEA-GCN and ATTED-II GCN are enriched for
426 genes involved in "Response to osmotic stress" (GO:0006970), the former has a higher
427 magnitude of enrichment according to NES scores (Figure 5D and E). Thus, in addition to
428 showcasing the robust gene function prediction abilities of the TEA-GCN, we have exemplified its
429 potential to predict the multimodal functions of genes.


A
143
237
19
TEA-GCN
ATTED-II
TF Positive edges
0.0
0.5
1.0
1.5
Z (Co-exp.)
TEA-GCN
ATTED-II
B
TEA-GCN
ATTED-II
JLO
BUM1
KNAT1
WRKY25
HSF4
HSFB2A
RHL41
WRKY6
TRX5
LOX3
OBP2
CYP83B1
MAM1
CYP79B2
CYP79B3
ANAC019
VSP1
ERD1
GSTF6
LOX2
COR47
ABI5
AT5G66400
ABI3
OEP16-2
SDH2-3
GEA6
EM1
FUS3
AT1G62290
AT2S3
Z (Co-exp.)
> 0
> 0.5
> 1.0
> 1.5
< 0
< −0.5
< −1.0
< −1.5
TF
Non-TF
C
Transcription factor co-expression neighbourhood
WRKY25
Genes ranked by co-exp.
Highest
Lowest
ES
Calculate Enrichment Score (ES)
D
GO-BP Terms
Hyperosmotic response
Heat acclimation
Hyperosmotic salinity response
Response to salt
Cellular reponse to salt stress
Response to cold
Response to heat
Response to reactive oxygen species
Response to osmotic stess
Resposne to salt stress
TEA-GCN
ATTED-II
NES
2.5
2.0
1.5
1.0
0.5
< 0
p_BH = 0.05
0
1
2
3
4
≥5
− log10 (p_BH)
E
Enrichement Score
0.5
0.2
0.0
-0.2
GO-BP Terms
Hyperosmotic response
Heat acclimation
Hyperosmotic salinity response
Response to salt
Cellular reponse to salt stress
Response to cold
Response to heat
Response to reactive oxygen species
Response to osmotic stess
Resposne to salt stress
0
5000
10000
15000
20000
25000
Genes Ranked by Co-exp.

430

Figure 5. Characterization of transcription factor-target predictions in *A. thaliana* with TEA-GCN. A) Comparing co-expression of 673 TF positive edges in TEA-GCN and ATTED-II GCN. Venn diagram (top) shows common and unique TF positive edges that are co-expressed (***Z***(Co-exp.) ≥ 1) in *A. thaliana* TEA-GCN and ATTED-II GCN. Heatmap (bottom shows) co-expression strengths of TF positive edges in the two GCNs. Edges are sorted based on their co-expression in ATTED-II GCN. Color intensities correspond to co-expression strength (i.e. ***Z***(co-exp.)). **B)** Co-expression between selected TFs and their known targets in TEA-GCN and ATTED-II GCN. TF and non-TF genes are represented by yellow diamond-shaped nodes and purple circular-shaped nodes, respectively. Edges with different colors and thicknesses represent co-expression strengths between TFs and their known targets. Grey chevrons within the edges indicate the direction of regulation between TFs and their targets. **C)** Functional enrichment of the co-expression neighborhood of TFs is determined using Gene Set Enrichment Analysis (GSEA). For a given TF, its co-expression neighborhood information is encapsulated by genes ranked by their co-expression with said TF and used as input for GSEA. Gene Ontology (GO) biological processes (BP) were used as gene sets. **D)** Dot plots comparing the enrichment statistics of selected GO terms in the neighborhood of *WRKY25* in TEA-GCN and ATTED-II GCN. GO terms were selected based on the known functions of *WRKY25*. Dot sizes for GO terms correspond to normalized enrichment scores (NES), while inverted triangles represent GO terms with negative NES scores. For GO terms with positive NES scores, the ax-axis indicates the statistical significance in the form of Benjamini-Hochberg-corrected p-values (p_BH). The red dotted lines denote the statistical significance threshold (p_BH ≥ 0.05). **E)** Running-sum profiles for enrichment scores and annotated gene rankings of selected GO terms between co-expression neighborhoods of *WRKY25* in TEA-GCN and ATTED-II GCN. For panels D and E, colors of dots, running sum profiles, and annotated gene rankings correspond to the different selected GO terms.

449 **TEA-GCN remains performant on smaller datasets**

450 So far, we showed that the TEA-GCN framework performs well on large transcriptome datasets
451 of *S. cerevisiae* (*n* = 150,096), *A. thaliana* (*n* = 57,759), and *H. sapiens* (*n* = 95,553). Given that
452 available public RNA-seq data might be less abundant for some species, we investigated TEA-
453 GCN perforamnce on smaller datasets. To this end, we generated GCNs for downsampled
454 datasets consisting of 5,000 samples and 500 samples that were randomly selected (without
455 replacement) from the *A. thaliana* public transcriptome dataset. We show that while the overall
456 performance of the *A. thaliana* TEA-GCNs markedly decreases as the datasets become smaller,
457 the *A. thaliana* TEA-GCN at *n* = 500 is still substantially more performant (AUROC-HM: 0.649,
458 AUPRC-HM: 0.677; Figure 6A; Table S20) than the non-ensemble PCC GCN (AUROC-HM:
459 0.602, AUPRC-HM: 0.622; Figure 3A and B; Table S17) generated from the complete *A. thaliana*
460 public transcriptome dataset (*n* = 57,759). Crucially, we show that *A. thaliana* TEA-GCN at *n* =
461 5,000 (AUROC-HM: 0.669, AUPRC-HM: 0.699; Figure 6A; Table S20) can still outperform
462 ATTED-II GCN (AUROC-HM: 0.659, AUPRC-HM: 0.690; Figure 6A; Table S17), despite using a
463 downsampled dataset that is more than one order of magnitude smaller than the complete
464 dataset. This outperformance is made more impressive considering that ATTED-II GCN is
465 generated via the "subbaging" method using 14,741 curated, sample-balanced, and batch-
466 corrected RNA-seq samples (according to https://atted.jp/download/), as opposed to *A. thaliana*
467 TEA-GCN at *n* = 5,000 which uses 5,000 non-batch corrected RNA-seq samples that have been
468 randomly selected from the public transcriptomic dataset. Thus, we have demonstrated that the
469 performance of TEA-GCN is not only translatable to smaller datasets, but can likely outperform
470 the "subbaging" methods, which require the time-consuming and complex annotation and
471 balancing of samples.

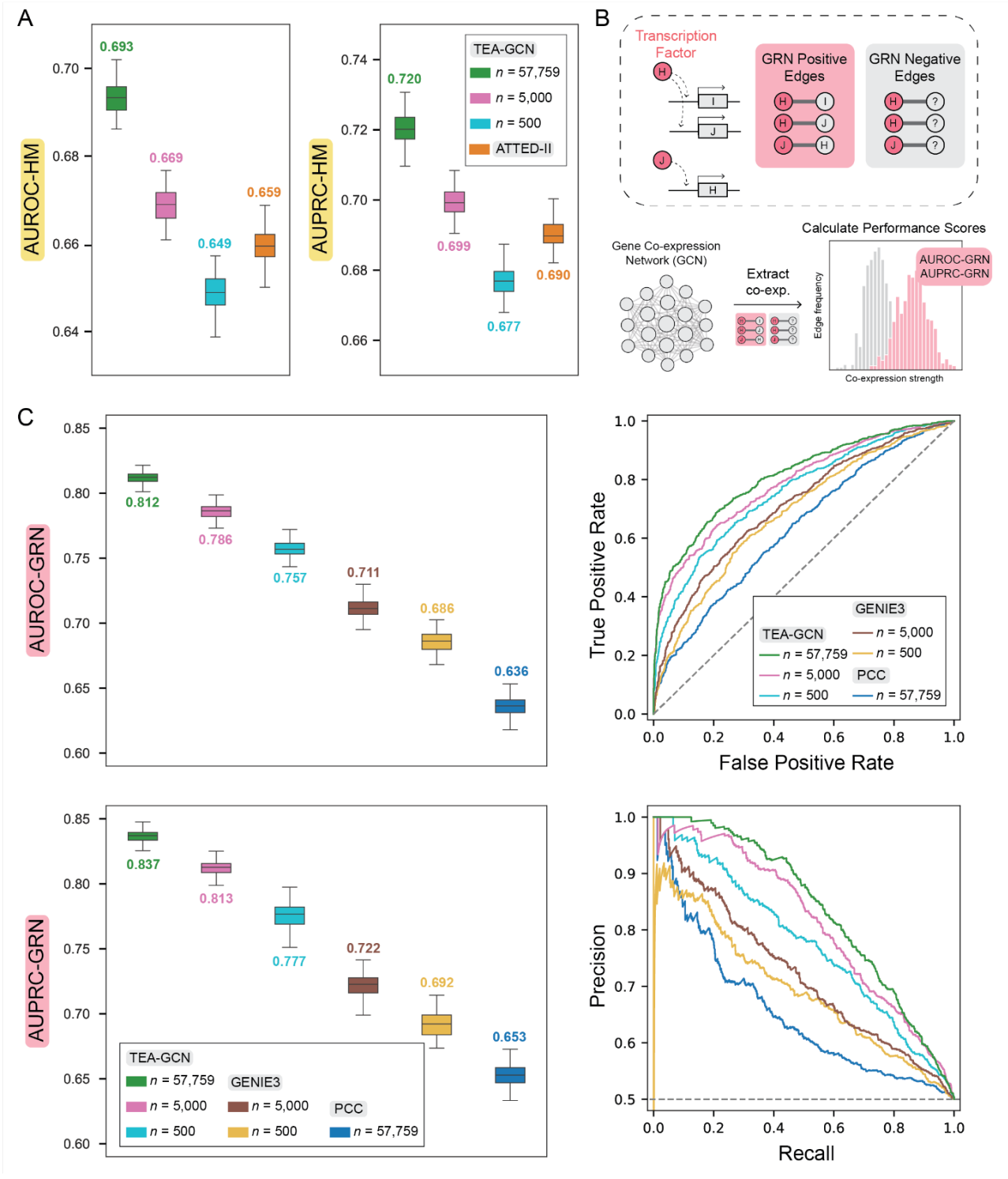


Figure 6. Scalability of TEA-GCN to smaller datasets and its GRN inference performance compared to GENIE3. A) Distribution of AUROC-HM performance scores and AUPRC-HM performance scores between ATTED-II GCNs and *A. thaliana* TEA-GCN constructed from the complete transcriptomic dataset (n = 57,759) and randomly downsampled datasets (n = 5,000 and n = 500), that were sampled 100 number of times. **B)** AUROC-GRN and AUPRC-GRN calculation relies on GRN positive edges and GRN negative edges. GRN positive edges connect transcription factors to the targets they activate. GRN positive edges are directional, differing from TF positive edges, thus transcription factors targeting each other are connected bytwo separate positive edges. GRN negative edges are generated by randomizing genes connected to transcription factors in GRN positive edges. **C)** Distribution of AUROC-GRN

480 and AUPRC-GRN scores of *A. thaliana* TEA-GCNs and GENIE3-GRNs constructed from the complete and downsampled
481 transcriptomic datasets. Non-ensemble PCC GCN is included as a comparison. Receiver operating characteristics curves (ROCs)
482 and Precision-recall curves (PRCs) correspond to median performance scores for each GCN.

483 **TEA-GCN outperforms state-of-the-art in GRN inference**

484 As demonstrated earlier, the TEA-GCN framework excels in revealing gene regulatory networks
485 (Figure 3C, 3D, and 5), which prompted us to compare its GRN-inference capabilities to GENIE3[24]
486 (ref), a well-established machine learning-based algorithm dedicated to inferring GRN from
487 transcriptomic data. Briefly, GRN-inference methods like GENIE3 calculate the propensities of
488 genes as regulatory targets for transcription factors defined by the user. Since GRNs are
489 directional (e.g. TF1 → TF2) and the aim is to predict TF - target gene pairs (rather than any gene-
490 gene pairs), we defined a new set of positive directional edges from experimentally verified GRNs
491 for *A. thaliana* (Figure 6B; Table S21; see methods). Similarly, the negative edges were defined
492 by randomizing target genes connected to edges in the GRN positive edge set (Figure 6B).
493 AUROC-GRN and AUPRC-GRN were calculated 100 times from propensities (for GENIE3) and
494 co-expression strengths (for TEA-GCN) of edges (Figure 6B; Table S22).

495 Comparison of TEA-GCN and GENIE3 revealed that TEA-GCN outperforms GENIE3
496 even when a smaller dataset comprising < 10% of samples is used (Figure 6C; Table S22).
497 Impressively, TEA-GCN at n = 500 (AUROC-GRN: 0.757, AUPRC-GRN: 0.777; Figure 6C; Table
498 S22) outperforms the GENIE3 GRN at n = 5,000 (AUROC-GRN: 0.711, AUPRC-GRN: 0.722;
499 Figure 6C; Table S22) in GRN inference despite a ten-fold difference in dataset size. We also
500 show that *A. thaliana* TEA-GCN generated using the complete dataset (n = 57,759) has the best
501 GRN-inference ability (AUROC-GRN: 0.812, AUPRC-GRN: 0.837; Figure 6C; Table S22). On the
502 other hand, the non-ensemble PCC GCN generated using the complete dataset (n = 57,759)
503 performed the worst (AUROC-GRN: 0.636, AUPRC-GRN: 0.653; Figure 6C; Table S22). Thus,
504 the excellent GRN-inference capabilities of TEA-GCNs demonstrate that our method can
505 outperform dedicated GRN algorithms, such as GENIE3.

506 **Experimental contexts underpinning gene co-expression are discoverable in TEA-**
507 **GCN**

508 Despite their utility in identifying functional associations between genes, current network-based
509 methods cannot explain which biological and experimental conditions (e.g., organ, tissue,
510 treatment) underpin the observed co-expression between genes, making it difficult to understand
511 in which conditions they are functionally related. This lack of explainability is common for machine-
512 learning GRN methods such as GENIE3[24] and GRNBoost2[68], which are often considered "black
513 boxes" due to the difficulty in understanding how the network edges are determined.

514 To investigate whether TEA-GCN provides explainable gene-gene relationships, we
515 analyzed the datasets underpinning the condition/tissue-specific co-expression (Figure 7). To this
516 end, we developed a Natural Language Processing (NLP) method that analyzes the metadata for
517 each of the data partitions identified by k-means (Figure 7A). Specifically, we identified statistically
518 overrepresented lemmas within the metadata of the *A. thaliana* RNA-seq samples (Table S14) in
519 each dataset partition (Figure 7A; see methods). As a result, dataset partitions can be annotated
520 by the overrepresented lemmas that might describe the experimental conditions and/or tissue of
521 samples contained within partitions (Table S23). Next, the overrepresented lemmas of the
522 partitions were used to annotate TEA-GCN edges by a statistical approach similar to GSEA. To
523 achieve this, for each edge in TEA-GCN, we first ranked the underpinning co-expression
524 strengths stemming from each the dataset partition (step immediately after Coefficient
525 Aggregation).(Table S24). Then, we used these ranks to identify the overrepresented lemmas of
526 dataset partitions with strongly observed co-expression (Figure 7B, see methods).

527 To test this approach, we applied this method on GRN positive edges that exhibited strong
528 co-expression in *A. thaliana* TEA-GCN (i.e. ***Z***(Co-exp.) ≥ 1.5). The majority of identified edges
529 (293 out of 400 edges; 73.25%; Table S25) could be annotated with at least one lemma, and
530 these annotations largely correspond with the known functions of the gene pairs (Figure 7C; Table
531 S23). To exemplify, we selected two well-characterized transcription factors, ABSSISIC ACID
532 INSENSITIVE 5 (ABI5) and MALE STERILE 188 (MS188). Five positive edges of *ABI5*
533 (connected to *GEA6*, *EM1*, *SDH2-3,* and *AT5G66400)* were annotated with the "ABA" co-
534 expression context (Figure 7C; Table S26), which is a common acronym for the abscisic acid
535 phytohormone that mediates important physiological processes such as seed dormancy
536 maintenance and abiotic stress response[69–71]. This is consistent with the prominent regulatory
537 role of ABI5 in ABA signaling[57,72,73]. Interestingly, *ABI5* edges connecting *AT5G66400*, *EM1*, and
538 *GEA6* wereannotated with experimental contexts reflecting low-light/dark growth conditions (i.e.

539 “Shade”, “Etiolate”, “Darkness”; Figure 7C; Table S26), which represents novel information that
540 can be used to further dissect the regulatory wirings underpinning ABA signaling of ABI5. This is
541 supported by recent studies that have revealed darkness-specific regulatory players such as
542 COP1 E3 ligase[74] and phytochrome-interacting factors[75] in ABA-induced signaling by modulating
543 ABI5 activity. In another example, all three edges of *MS188* (connected to *FAR2*, *MYB35*, and
544 *MS1*) were annotated with experimental contexts related to floral tissues (i.e. “Flower”,
545 “Flowering”, “Inflorescence”) which is congruent with MS188’s known function in anther
546 developmental processes[76–78] (e.g. tapetum development, pollen exine formation, and pollen
547 callose development; Figure 7C; Table S26). Thus, TEA-GCN, when combined with NLP
548 approaches, produces interpretable edges that allow the elucidation of the experimental context,
549 further boosting gene function prediction approaches.

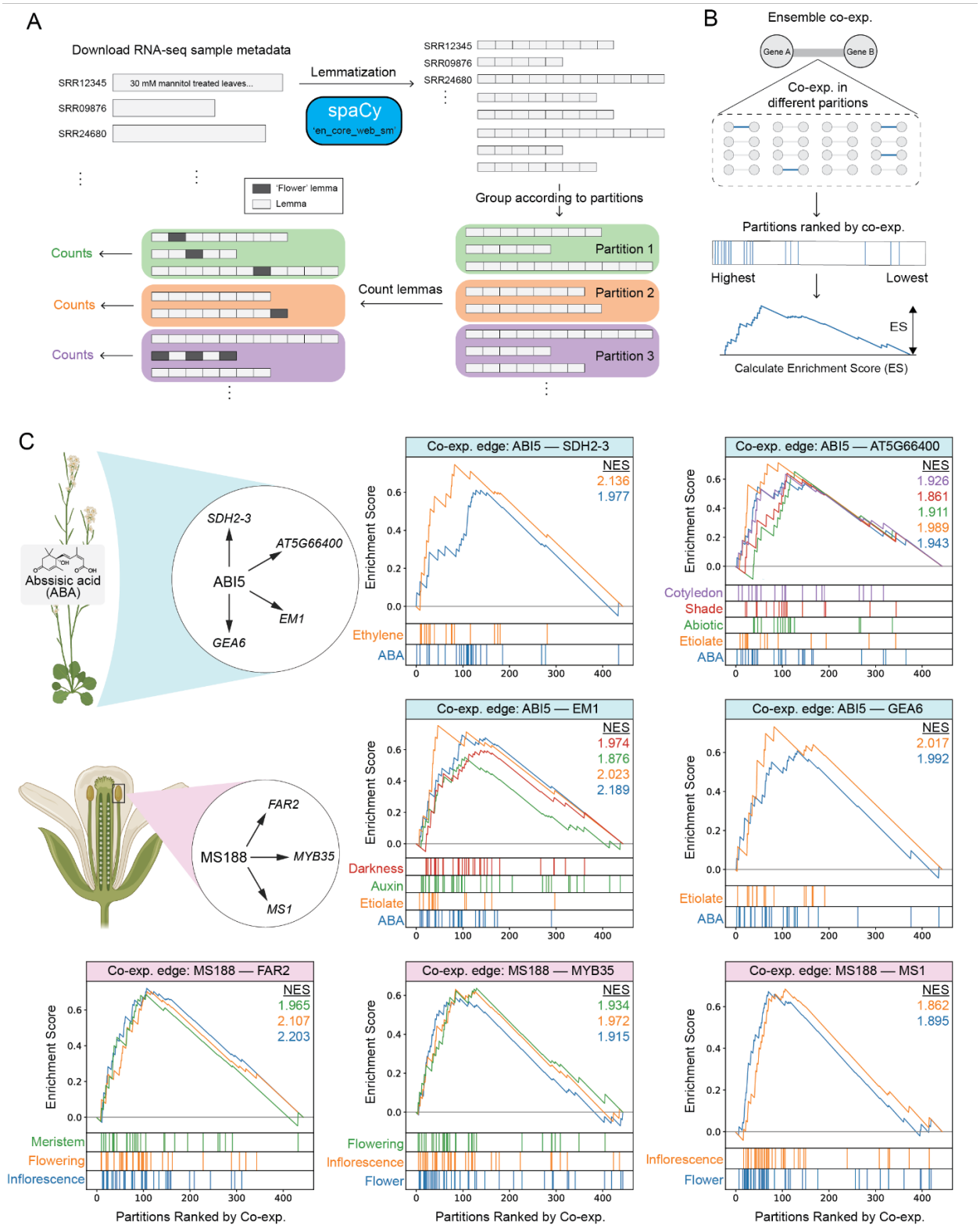


Figure 7. Annotating co-expression edges of *A. thaliana* TEA-GCN with experimental contexts. A) The metadata of RNA-seq samples are broken down into lemmas using the "en_core_web_sm" natural language processing (NLP) model from spaCy. Lemmas of RNA-seq samples belonging to the same dataset partition are then combined. For each partition, occurrences of different lemmas are counted, and enriched lemmas were determined by comparing the observed lemma occurrences to permuted occurrences when

partition assignments of RNA-seq samples are randomly shuffled. As a result, dataset partitions can be annotated by overrepresented lemmas that likely denote the experimental conditions or tissues which samples in the dataset partition were predominantly generated from. **B)** Co-expression edges of TEA-GCN can be annotated with lemmas enriched with an approach similar to Gene-set Enrichment Analysis (GSEA). The ensemble co-expression strength of a given TEA-GCN edge is first deconvoluted into co-expression strengths obtained from the different dataset partitions, and these strengths are then used to rank dataset partitions. This facilitates the calculation of enrichment scores as seen in GSEA. **C**) Lemma enrichment analysis of edges connecting *ABSCISIC ACID INSENSITIVE 5* (*ABI5*) and *MALE STERILE 188* (*MS188*) transcription factors to their known targets. Running-sum profiles for enrichment scores and annotated partition rankings of the co-expression context lemmas detected for each of the presented edges are shown. Line colors of running sum profiles correspond to colors of experimental context lemmas.

Discussion

Gene co-expression networks (GCNs) generated from transcriptomic datasets can elucidate the intricate co-regulatory and co-functional relationships between genes thus enabling gene function prediction. Consequently, GCNs have proven invaluable as a hypothesis generation tool to characterize genes in non-model species whose functions are hard to predict based on sequence similarity against characterized homologs in well-studied models (i.e., members of large gene families, orphan and lineage-specific genes/proteins)[79–83]. Due to the increased affordability of Next-generation Sequening (NGS), the number of species with transcriptomic data is rapidly increasing, which presents an unprecedented opportunity to unravel the evolutionary dynamics of gene function across diverse organisms via the construction of GCNs. However, current GCN construction methodologies [4,28,84,85] were developed for use on small in-house sequenced datasets or manually curated public RNA-seq samples and therefore, are not well equipped to handle and exploit the various complexities of public transcriptome data (e.g. high heterogeneity, sample imbalance, batch effects, poor sample quality). To address this unmet need, we developed TEA-GCN, an ensemble methodology centered around determining gene co-expression based on local co-expression observed in different dataset partitions. We show that TEA-GCN is superior to the widely-used method of calculating PCC across the whole dataset[28,84,86,87], and the state-of-the-art "Subbagging" method[16,17], across three model organisms of vastly different organizational/biological complexities and evolutionary lineages.

Through the characterization of co-expression edges, we show that TEA-GCN excels in capturing co-expression relationships that manifest in certain conditions and/or in specific tissues which enables the discovery of complex regulatory mechanisms that can be elusive. Inspired by the growing field of Explainable Artificial intelligence[88] (XAI) towards explaining the inner workings of complex machine-learning models, we demonstrated TEA-GCN's ability to explain experimental contexts underpinning gene co-expression through the use of NLP. This was possible by capitalizing on our method's ability for capturing condition/tissue-specific co-expression. Thus, we have showed that TEA-GCN can not only outperform more complex state-of-the-art methods, but also for the first time ever, report biologically significant information as experimental contexts that can supercharge gene function prediction. We made the TEA-GCN method available for use via https://github.com/pengkenlim/TEA-GCN as a easy to install and well documented standalone program for users to generate high quality GCNs from their own large transcriptomic datasets.

Methods

Generation and preprocessing of *S. cerevisiae*, *A. thaliana,* and *H. sapiens* public transcriptomic datasets

S. cerevisiae and *H. sapiens* mRNA sequences were downloaded from the NCBI RefSeq annotations of their reference genome assemblies (GCF_000146045.2 and GCF_000001405.40). Representative mRNA sequences of protein-coding genes (primary transcripts) for *S. cerevisiae* and *H. sapiens* were then obtained by selecting the longest transcript isoform for every protein-coding gene loci. Primary transcripts of *A. thaliana* were obtained by downloading the "Araport11_cdna_20220914_representative_gene_model.gz" file from TAIR database[89]. 6,612 *S. cerevisiae*, 19,379 *H. sapiens*, and 27,562 *A. thaliana* primary transcrtipts were used as pseudoalignment references for gene expression estimation of RNA-seq sample Illumina sequencing reads (read data) downloaded from ENA using Kallisto's quant functionality in single-end mode with default parameters (ref). Fastq files of 188,189, 66,048, and 142,828 RNA-seq samples were download and pseudoaligned for *S. cerevisiae, A. thaliana, and H. sapiens,* respectively. Transcripts per million (TPM) values of primary transcripts of RNA-seq samples estimated by Kallisto were used as gene expression abundances to construct gene expression matrices for the different species. RNA-seq samples with poor read alignment were identified based its %_pseudoaligned statistic (reported by Kallisto) below a certain threshold (5% for *H. sapiens*; 20% for *S. cerevisiae* and *A. thaliana*) as described[90–92] (Table S10 to 12). After removal RNA-seq samples with poor read alignment, gene expression matrices comprising of gene expression abundances from 150,096 (79.8%), 57,759 (87.5%), and 95,553 (66.9%) retained RNA-seq samples were designated as the public transcriptomic datasets for *S. cerevisiae, A. thaliana* and *H. sapiens,* respectively.

Dataset partitioning

TPM values of each RNA-seq sample in the gene expression matrices were standardized to unit variance (within-sample standardization) to control for batch effects at the sample level as described[33,90]. Next, the standardized gene expression values were subjected to dimension reduction using Principal Component Analysis (PCA; decomposition.IncrementalPCA function from scikit-learn version 1.3.0) where only the first 1,000 Principal components (PCs) were used as high-level features to embed RNA-seq samples in PC space (PC embeddings) that describes their relative transcriptome similarity between each other. PC embeddings of RNA-seq samples

626 were then used as features for k-means clustering (cluster.Kmeans function from scikit-learn
627 version 1.3.0) using k-means++ as initialization method using random seed of 42 (init = k-
628 means++, n_init = 5, max_iter = 1000, random_state = 42).

629 Preselection of partitioning granularities (*ks*) for optimal *k* determination

630 For each species's public transcriptomic dataset, at least 50 *k* values were preselected at fixed
631 intervals based on the silhouette coefficient, an internal clustering metric commonly used to
632 determine the clustering performance of unsupervised clustering algorithms like k-means[32]. To
633 ensure preselected *k*s were sampled at regular intervals across the *k* range from 2 to 1000, we
634 sectioned said range into *k* windows of length 19. For each *k* window, K-means clustering was
635 iteratively used to generate dataset partitions, and silhouette coefficients was calculated from the
636 resultant partition assignments. *k*s that produced the highest silhouette coefficient in each window
637 were selected (Table S13 to 13).

638 Data sources for establishing positive edges

639 *S. cerevisiae* TF positive edges are generated by downloading activation regulatory links from
640 the YEASTRACT+ database[93] (http://www.yeastract.com/) that fufill both"Documented" and
641 "Potential" criteria. According to the authors, "Documented " regulatory links are infered from
642 experimental transcription factor-binding evidence while "Potential" regulatory links connects
643 transcription factors with target genes that contains corresponding consensus sequences in
644 upstream promoter region[93]. *A. thaliana* TF positive edges are generated by downloading
645 activation regulatory links from the PlantTFDB database[94] (https://planttfdb.gao-lab.org/). *H.*
646 *sapiens* TF positive edges are generated by downloading activation regulatory links from
647 TRRUST version 2 database[95] (https://www.grnpedia.org/trrust/) and then eliminating regulatory
648 links that does not connect to any known *H. sapiens* transcription factors. *H. sapiens* transcription
649 factors are obtained from the HumanTFs database[96] (https://humantfs.ccbr.utoronto.ca/). GO
650 annotations for S. *cerevisiae*, *A. thaliana* and *H. sapiens* were downloaded from EBI Gene
651 Ontology Annotation (GOA) Database[97], The Arabidopsis Information Resource[89] (TAIR), and
652 *Saccharomyces* Genome Database[98] (SGD), respectively. Pathway annotations for S.
653 *cerevisiae*, *A. thaliana* and *H. sapiens* were webscraped from BioCyc[99], PlantCyc[38], and
654 HumanCyc[100] databases, respectively.

Correlation coefficients and mutual rank calculation

PCC, SCC, and Bicor were calculated based on previous publications detailing their use in measuring gene coexpression[1,27] (ref). In the event where correlation coefficients cannot be determined (e.g. expression values of one gene are all 0), correlation coefficients are set to 0. MR transformation of determined co-expression strengths of a GCN are calculated as described[35].

Aggregation functions

The mathematical description of the 3 ensemble aggregation functions, namely the Maximum (Max), Average (Avg), and Rectified Average(RAvg) are detailed below.

Let $V = [v_1, v_2, v_3, \ldots, v_n]$ and be a vector of co-expression strengths between a given pair of genes from different (n) GCNs. The Max aggregation method involves taking the maximum co-expression strength value across co-expression strengths in V and can be expressed as a function where:

$$Max(V) = max_i v_i$$

The Avg aggregation method involves taking the arithmetic mean of V and can be expressed as a function where:

$$Avg(V) = \frac{\sum_i v_i}{n}$$

The RAvg aggregation method involves first rectifying negative edges in V to zero. The arithmetic mean is then calculated from the resultant rectified vector, W. RAvg can be expressed as a function where:

$$W = \frac{|V| + V}{2}$$

$$RAvg(V) = \frac{\sum_i w_i}{n}$$

Standardizing co-expression strengths within GCNs

To compare the co-expression strengths of edges connecting the same gene pair across different GCNs, the co-expression strengths of edges were standardized before comparison. Because different CGNs do not necessarily contain the same genes and edges, the co-expression strengths of each GCN were standardized based on the mean and standard deviation of the co-

expression strengths from a subset of common edges shared between the GCNs. The mathematical description of said standardization is detailed below.

Let $X = [x_1, x_2, x_3 \ldots, x_N]$ be a vector of co-expression strengths of common edges in one of the GCNs to be compared. The mean co-expression strength (μ) and standard deviation (σ) of X can be calculated as follows:

$$\mu = \frac{\sum_i x_i}{N}$$

$$\sigma = \sqrt{\frac{\sum(X - \mu)}{N}}$$

The co-expression strength of any given edge in the GCN (x_i) can then be standardized as follows:

$$Z(x_i) = \frac{x_i - \mu}{\sigma}$$

For interpretability, ***Z***(co-exp.) is used to refer to standardized co-expression strengths throughout the text.

Inferring *A. thaliana* GRNs using GENIE3

The Python implementation of GENIE3[24] was used to infer GRNs used in this study with default settings (i.e. ntrees = 1000, tree-method = Random Forest, K = square root method). A list of 324 *A. thaliana* transcription factors (Table S21) was obtained from PlantTFDB[94] and used as regulators for GENIE3.

Downloading of *A. thaliana* RNA-seq metadata

The metadata of RNA-seq samples was downloaded from ENA via the ENA Browser API using "sample_description", "dev_stage", "study_title", "sample_title", "tissue_lib" and "tissue_type" as search fields (Table S13 to 15).

Identifying overrepresented lemmas in *A. thaliana* dataset partitions

The "en_core_web_sm" NLP model (version 3.7.1) from the SpaCy (version 3.7.1) Python package was used to lemmatized metadata of *A. thaliana* RNA-seq samples. Before lemmatization, the metadata of RNA-seq samples was preprocessed. First, metadata returned from all search fields of every RNA-seq sample were concatenated. Next, stop words (e.g. about, quite, the) were detected from the concatenated metadata using the token.is_stop attribute of

712 word tokens generated by the “en_core_web_sm” model. Digits and punctuations were also
713 removed from the concatenated metadata.

714 To obtain overrepresented lemmas from the different dataset partitions, metadata lemmas
715 of RNA-seq samples belonging to the same partition were grouped and combined to represent
716 lemmas of each dataset partition. The occurrence of lemmas in each dataset partition was then
717 counted and stored as observed lemma occurrences. Permuted lemma occurrences were
718 similarly obtained by repeating the counting process on dataset partition lemmas combined from
719 randomly shuffled sample-partition assignments. A null hypothesis p-value indicating the
720 overrepresentation significance of a particular lemma in a dataset partition can be mathematically
721 defined below.

722 Let $S = [s_1, s_2, s_3, \ldots s_n]$ be a vector of permuted occurrences of a lemma in a dataset
723 partition from n random shufflings of sample-partition assignments and o be the observed
724 occurrence. The p-value (P) can be calculated as follows:

725 $$P = \frac{1 + \sum_{i=1}^{n} I(s_i \geq o)}{n}$$

726 Where I is an indicator function that takes a value of 1 or 0 if the statement is true or false
727 respectively.

728 P-values determined for every lemma in each dataset partition were then subjected to
729 multiple-testing correction using the Benjamini-Hochberg method[101], yielding Benjamini-Hochberg
730 corrected p-values (p_BH). Based on 1000 permutations, 2,333 overrepresented lemmas (p_BH
731 ≤ 0.05, o > 20) were identified for *A. thaliana* dataset partitions . After manual selection to remove
732 lemmas that are unlikely to contribute to biological insight (e.g. “arabidopsis”, “bottle”, “clean”,
733 “device”, etc.), 276 lemma labels remain (Table S23).

734 Enrichment-based analyses

735 GSEA of co-expression neighborhoods of selected *A. thaliana* transcription factors was
736 implemented by the GSEApy python package[67] (version 1.1.2). For each transcription factor,
737 genes were ranked based on co-expression strength. The ranks were centered and gene sets
738 were constructed using *A. thaliana* GO annotations (biological process catergory). Together, the
739 centered ranks and constructed gene sets were used as input for the prerank sub-command of
740 GSEApy (weight = 1). Annotating edges between co-expressed genes with experimental
741 context lemmas also uses the prerank sub-command of GSEApy in the same way. However,
742 centered partition rankings determined from ranking dataset partitions based on co-expression

743 strengths after Coefficient Aggregation were used in place of gene rankings. Instead of gene
744 sets, sets of dataset partitions for which lemma labels were assigned were used.

745 ## Data and code availability

746 TEA-GCN is available at https://github.com/pengkenlim/TEA-GCN for the scientific community to
747 generate high-quality GCNs from their transcriptomic datasets. S. cerevisiae, A. thaliana, and H.
748 sapiens TEA-GCNs generated by this study can be found at FigShare DOIs
749 10.6084/m9.figshare.26347018, 10.6084/m9.figshare.26347078, and
750 10.6084/m9.figshare.26347297, respectively. Transcriptomic datasets in the form of gene
751 expression matrices used in this study can be found under 10.6084/m9.figshare.26346916,
752 10.6084/m9.figshare.26346964, and 10.6084/m9.figshare.263470105, respectively.
753 Supplemental tables referenced in this study are available at FigShare DOI
754 10.6084/m9.figshare.26347507.

755 ## References

756 1. Usadel, B. *et al.* Co-expression tools for plant biology: opportunities for hypothesis
757 generation and caveats. *Plant Cell Environ.* **32**, 1633–1651 (2009).
758 2. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-
759 expression analysis for functional classification and gene-disease predictions. *Brief.*
760 *Bioinform.* **19**, 575–592 (2018).
761 3. Lim, P. K., Zheng, X., Goh, J. C. & Mutwil, M. Exploiting plant transcriptomic databases:
762 Resources, tools, and approaches. *Plant Commun.* **3**, 100323 (2022).
763 4. Rao, X. & Dixon, R. A. Co-expression networks for plant biology: why and how. *Acta*
764 *Biochim. Biophys. Sin.* **51**, 981–988 (2019).
765 5. Aoki, K., Ogata, Y. & Shibata, D. Approaches for extracting practical information from gene
766 co-expression networks in plant biology. *Plant Cell Physiol.* **48**, 381–390 (2007).
767 6. Rhee, S. Y. & Mutwil, M. Towards revealing the functions of all genes in plants. *Trends*
768 *Plant Sci.* **19**, 212–221 (2014).
769 7. Julca, I., Tan, Q. W. & Mutwil, M. Toward kingdom-wide analyses of gene expression.
770 *Trends Plant Sci.* **28**, 235–249 (2023).
771 8. Burgin, J. *et al.* The European Nucleotide Archive in 2022. *Nucleic Acids Res.* **51**, D121–

772 D125 (2023).
773 9. Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species.
774 *Nucleic Acids Res.* **50**, D129–D140 (2022).
775 10. Ruprecht, C., Vaid, N., Proost, S., Persson, S. & Mutwil, M. Beyond Genomics: Studying
776 Evolution with Gene Coexpression Networks. *Trends Plant Sci.* **22**, 298–307 (2017).
777 11. Ovens, K., Eames, B. F. & McQuillan, I. Comparative Analyses of Gene Co-expression
778 Networks: Implementations and Applications in the Study of Evolution. *Front. Genet.* **12**,
779 695399 (2021).
780 12. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network
781 construction and analysis: safety in numbers. *Bioinforma. Oxf. Engl.* **31**, 2123–2130 (2015).
782 13. Vandenbon, A. Evaluation of critical data processing steps for reliable prediction of gene co-
783 expression from large collections of RNA-seq data. *PloS One* **17**, e0263344 (2022).
784 14. Gupta, C. & Pereira, A. Recent advances in gene function prediction using context-specific
785 coexpression networks in plants. *F1000Research* **8**, F1000 Faculty Rev-153 (2019).
786 15. Pearce, S., Ferguson, A., King, J. & Wilson, Z. A. FlowerNet: a gene expression correlation
787 network for anther and pollen development. *Plant Physiol.* **167**, 1717–1730 (2015).
788 16. Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an animal
789 gene coexpression database navigating from a global view to detailed investigations.
790 *Nucleic Acids Res.* **51**, D80–D87 (2023).
791 17. Obayashi, T., Hibara, H., Kagaya, Y., Aoki, Y. & Kinoshita, K. ATTED-II v11: A Plant Gene
792 Coexpression Database Using a Sample Balancing Technique by Subagging of Principal
793 Components. *Plant Cell Physiol.* **63**, 869–881 (2022).
794 18. Lin, H., Yu, J., Pearce, S. P., Zhang, D. & Wilson, Z. A. RiceAntherNet: a gene co-
795 expression network for identifying anther and pollen development genes. *Plant J. Cell Mol.*
796 *Biol.* **92**, 1076–1091 (2017).
797 19. Lancour, D. *et al.* Analysis of brain region-specific co-expression networks reveals clustering

798 of established and novel genes associated with Alzheimer disease. *Alzheimers Res. Ther.*
799 **12**, 103 (2020).
800 20. Liesecke, F. *et al.* Improved gene co-expression network quality through expression dataset
801 down-sampling and network aggregation. *Sci. Rep.* **9**, 14431 (2019).
802 21. Wong, D. C. J. Network aggregation improves gene function prediction of grapevine gene
803 co-expression networks. *Plant Mol. Biol.* **103**, 425–441 (2020).
804 22. Arshad, Z. & McDonald, J. F. A computational approach to generate highly conserved gene
805 co-expression networks with RNA-seq data. *STAR Protoc.* **3**, 101432 (2022).
806 23. Liesecke, F. *et al.* Ranking genome-wide correlation measurements improves microarray
807 and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* **8**, 10885
808 (2018).
809 24. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from
810 Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).
811 25. Hou, J. *et al.* Distance correlation application to gene co-expression network analysis. *BMC*
812 *Bioinformatics* **23**, 81 (2022).
813 26. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual
814 information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328 (2012).
815 27. Hardin, J., Mitani, A., Hicks, L. & VanKoten, B. A robust measure of correlation between two
816 genes on a microarray. *BMC Bioinformatics* **8**, 220 (2007).
817 28. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
818 analysis. *BMC Bioinformatics* **9**, 559 (2008).
819 29. Yue, L. *et al.* Gene co-expression network analysis of the heat-responsive core
820 transcriptome identifies hub genes in Brassica rapa. *Planta* **253**, 111 (2021).
821 30. Li, J.-X. *et al.* Investigation of hub gene associated with the infection of Staphylococcus
822 aureus via weighted gene co-expression network analysis. *BMC Microbiol.* **21**, 329 (2021).
823 31. Lv, L. *et al.* Gene co-expression network analysis to identify critical modules and candidate

824 genes of drought-resistance in wheat. *PloS One* **15**, e0236186 (2020).
825 32. Arthur, D. & Vassilvitskii, S. k-means++: the advantages of careful seeding. in *Proceedings*
826 *of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–1035 (Society
827 for Industrial and Applied Mathematics, USA, 2007).
828 33. Zheng, X., Lim, P. K., Mutwil, M. & Wang, Y. A method for mining condition-specific co-
829 expressed genes in Camellia sinensis based on k-means clustering. *BMC Plant Biol.* **24**,
830 373 (2024).
831 34. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
832 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
833 35. Obayashi, T. & Kinoshita, K. Rank of correlation coefficient as a comparable measure for
834 biological significance of gene coexpression. *DNA Res. Int. J. Rapid Publ. Rep. Genes*
835 *Genomes* **16**, 249–260 (2009).
836 36. Aravind, L., Anantharaman, V. & Venancio, T. M. Apprehending multicellularity: regulatory
837 networks, genomics, and evolution. *Birth Defects Res. Part C Embryo Today Rev.* **87**, 143–
838 164 (2009).
839 37. Anantharaman, V., Iyer, L. M. & Aravind, L. Comparative Genomics of Protists: New Insights
840 into the Evolution of Eukaryotic Signal Transduction and Gene Regulation*. *Annu. Rev.*
841 *Microbiol.* **61**, 453–475 (2007).
842 38. Hawkins, C. *et al.* Plant Metabolic Network 15: A resource of genome-wide metabolism
843 databases for 126 plants and algae. *J. Integr. Plant Biol.* **63**, 1888–1905 (2021).
844 39. Delli-Ponti, R., Shivhare, D. & Mutwil, M. Using Gene Expression to Study Specialized
845 Metabolism-A Practical Guide. *Front. Plant Sci.* **11**, 625035 (2020).
846 40. Brockington, S. F. *et al.* Lineage-specific gene radiations underlie the evolution of novel
847 betalain pigmentation in Caryophyllales. *New Phytol.* **207**, 1170–1180 (2015).
848 41. Wink, M. Evolution of secondary metabolites from an ecological and molecular phylogenetic
849 perspective. *Phytochemistry* **64**, 3–19 (2003).

850 42. Lim, P. K., Julca, I. & Mutwil, M. Redesigning plant specialized metabolism with supervised
851 machine learning using publicly available reactome data. *Comput. Struct. Biotechnol. J.* **21**,
852 1639–1650 (2023).
853 43. Colinas, M. & Goossens, A. Combinatorial Transcriptional Control of Plant Specialized
854 Metabolism. *Trends Plant Sci.* **23**, 324–336 (2018).
855 44. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends Plant*
856 *Sci.* **24**, 83–98 (2019).
857 45. Voß, U., Bishopp, A., Farcot, E. & Bennett, M. J. Modelling hormonal response and
858 development. *Trends Plant Sci.* **19**, 311–319 (2014).
859 46. Zhang, Y., Berman, A. & Shani, E. Plant Hormone Transport and Localization: Signaling
860 Molecules on the Move. *Annu. Rev. Plant Biol.* **74**, 453–479 (2023).
861 47. Verma, V., Ravindran, P. & Kumar, P. P. Plant hormone-mediated regulation of stress
862 responses. *BMC Plant Biol.* **16**, 86 (2016).
863 48. Prieto, M. A., López, C. J. & Simal-Gandara, J. Glucosinolates: Molecular structure,
864 breakdown, genetic, bioavailability, properties and healthy and adverse effects. *Adv. Food*
865 *Nutr. Res.* **90**, 305–350 (2019).
866 49. Barba, F. J. *et al.* Bioavailability of Glucosinolates and Their Breakdown Products: Impact of
867 Processing. *Front. Nutr.* **3**, 24 (2016).
868 50. Johnson, I. T. Glucosinolates: bioavailability and importance to health. *Int. J. Vitam. Nutr.*
869 *Res. Int. Z. Vitam.- Ernahrungsforschung J. Int. Vitaminol. Nutr.* **72**, 26–31 (2002).
870 51. Sønderby, I. E., Geu-Flores, F. & Halkier, B. A. Biosynthesis of glucosinolates--gene
871 discovery and beyond. *Trends Plant Sci.* **15**, 283–290 (2010).
872 52. Paul, S., Geng, C.-A., Yang, T.-H., Yang, Y.-P. & Chen, J.-J. Phytochemical and Health-
873 Beneficial Progress of Turnip (Brassica rapa). *J. Food Sci.* **84**, 19–30 (2019).
874 53. Higdon, J. V., Delage, B., Williams, D. E. & Dashwood, R. H. Cruciferous Vegetables and
875 Human Cancer Risk: Epidemiologic Evidence and Mechanistic Basis. *Pharmacol. Res. Off.*

876 *J. Ital. Pharmacol. Soc.* **55**, 224–236 (2007).
877 54. Latté, K. P., Appel, K.-E. & Lampen, A. Health benefits and possible risks of broccoli – An
878 overview. *Food Chem. Toxicol.* **49**, 3287–3309 (2011).
879 55. Pfalz, M. *et al.* Metabolic engineering in Nicotiana benthamiana reveals key enzyme
880 functions in Arabidopsis indole glucosinolate modification. *Plant Cell* **23**, 716–729 (2011).
881 56. Bureau, M. & Simon, R. JLO regulates embryo patterning and organ initiation by controlling
882 auxin transport. *Plant Signal. Behav.* **3**, 145–147 (2008).
883 57. Collin, A., Daszkowska-Golec, A. & Szarejko, I. Updates on the Role of ABSCISIC ACID
884 INSENSITIVE 5 (ABI5) and ABSCISIC ACID-RESPONSIVE ELEMENT BINDING
885 FACTORs (ABFs) in ABA Signaling in Different Developmental Stages in Plants. *Cells* **10**,
886 1996 (2021).
887 58. Moreno-Risueno, M. A. *et al.* FUSCA3 from barley unveils a common transcriptional
888 regulation of seed-specific genes between cereals and Arabidopsis. *Plant J. Cell Mol. Biol.*
889 **53**, 882–894 (2008).
890 59. Skirycz, A. *et al.* DOF transcription factor AtDof1.1 (OBP2) is part of a regulatory network
891 controlling glucosinolate biosynthesis in Arabidopsis. *Plant J. Cell Mol. Biol.* **47**, 10–24
892 (2006).
893 60. Podzimska-Sroka, D., O'Shea, C., Gregersen, P. L. & Skriver, K. NAC Transcription Factors
894 in Senescence: From Molecular Structure to Function in Crops. *Plants Basel Switz.* **4**, 412–
895 448 (2015).
896 61. Chen, Y.-F. *et al.* The WRKY6 transcription factor modulates PHOSPHATE1 expression in
897 response to low Pi stress in Arabidopsis. *Plant Cell* **21**, 3554–3566 (2009).
898 62. Li, S., Fu, Q., Huang, W. & Yu, D. Functional analysis of an Arabidopsis transcription factor
899 WRKY25 in heat stress. *Plant Cell Rep.* **28**, 683–693 (2009).
900 63. Doll, J. *et al.* Arabidopsis thaliana WRKY25 Transcription Factor Mediates Oxidative Stress
901 Tolerance and Regulates Senescence in a Redox-Dependent Manner. *Front. Plant Sci.* **10**,

902 1734 (2019).
903 64. Jiang, Y. & Deyholos, M. K. Functional characterization of Arabidopsis NaCl-inducible
904 WRKY25 and WRKY33 transcription factors in abiotic stresses. *Plant Mol. Biol.* **69**, 91–105
905 (2009).
906 65. Li, S., Fu, Q., Chen, L., Huang, W. & Yu, D. Arabidopsis thaliana WRKY25, WRKY26, and
907 WRKY33 coordinate induction of plant thermotolerance. *Planta* **233**, 1237–1252 (2011).
908 66. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
909 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–
910 15550 (2005).
911 67. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set
912 enrichment analysis in Python. *Bioinforma. Oxf. Engl.* **39**, btac757 (2023).
913 68. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene
914 regulatory networks. *Bioinforma. Oxf. Engl.* **35**, 2159–2161 (2019).
915 69. Kishor, P. B. K., Tiozon, R. N., Fernie, A. R. & Sreenivasulu, N. Abscisic acid and its role in
916 the modulation of plant growth, development, and yield stability. *Trends Plant Sci.* **27**, 1283–
917 1295 (2022).
918 70. Bulgakov, V. P., Wu, H.-C. & Jinn, T.-L. Coordination of ABA and Chaperone Signaling in
919 Plant Stress Responses. *Trends Plant Sci.* **24**, 636–651 (2019).
920 71. Nambara, E. & Marion-Poll, A. ABA action and interactions in seeds. *Trends Plant Sci.* **8**,
921 213–217 (2003).
922 72. Lopez-Molina, L., Mongrand, S. & Chua, N. H. A postgermination developmental arrest
923 checkpoint is mediated by abscisic acid and requires the ABI5 transcription factor in
924 Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4782–4787 (2001).
925 73. Pollo-Rodríguez, F., Sánchez-Vicente, I. & Lorenzo, O. The turnover of ABI5 by scaffold
926 proteins to attenuate ABA signaling. *J. Exp. Bot.* **75**, 3749–3753 (2024).
927 74. Peng, J. *et al.* COP1 positively regulates ABA signaling during Arabidopsis seedling growth

928 in darkness by mediating ABA-induced ABI5 accumulation. *Plant Cell* **34**, 2286–2308
929 (2022).
930 75. Qi, L. *et al.* PHYTOCHROME-INTERACTING FACTORS Interact with the ABA Receptors
931 PYL8 and PYL9 to Orchestrate ABA Signaling in Darkness. *Mol. Plant* **13**, 414–430 (2020).
932 76. Xiong, S.-X. *et al.* The transcription factors MS188 and AMS form a complex to activate the
933 expression of CYP703A2 for sporopollenin biosynthesis in Arabidopsis thaliana. *Plant J.*
934 *Cell Mol. Biol.* **88**, 936–946 (2016).
935 77. Wang, K. *et al.* The Regulation of Sporopollenin Biosynthesis Genes for Rapid Pollen Wall
936 Formation. *Plant Physiol.* **178**, 283–294 (2018).
937 78. Zhang, Z.-B. *et al.* Transcription factor AtMYB103 is required for anther development by
938 regulating tapetum development, callose dissolution and exine formation in Arabidopsis.
939 *Plant J. Cell Mol. Biol.* **52**, 528–538 (2007).
940 79. Gerlt, J. A., Babbitt, P. C., Jacobson, M. P. & Almo, S. C. Divergent evolution in enolase
941 superfamily: strategies for assigning functions. *J. Biol. Chem.* **287**, 29–34 (2012).
942 80. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends*
943 *Plant Sci.* **19**, 698–708 (2014).
944 81. Jeffery, C. J. Current successes and remaining challenges in protein function prediction.
945 *Front. Bioinforma.* **3**, 1222182 (2023).
946 82. Jacobson, M. P., Kalyanaraman, C., Zhao, S. & Tian, B. Leveraging structure for enzyme
947 function prediction: methods, opportunities and challenges. *Trends Biochem. Sci.* **39**, 363–
948 371 (2014).
949 83. Jeffery, C. J. Enzymes, pseudoenzymes, and moonlighting proteins: diversity of function in
950 protein superfamilies. *FEBS J.* **287**, 4141–4149 (2020).
951 84. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and
952 Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2015).
953 85. Zainal-Abidin, R.-A., Harun, S., Vengatharajuloo, V., Tamizi, A.-A. & Samsulrizal, N. H.

954 Gene Co-Expression Network Tools and Databases for Crop Improvement. *Plants Basel*
955 *Switz.* **11**, 1625 (2022).
956 86. Proost, S. & Mutwil, M. CoNekT: an open-source framework for comparative genomic and
957 transcriptomic network analyses. *Nucleic Acids Res.* **46**, W133–W140 (2018).
958 87. De Bodt, S., Hollunder, J., Nelissen, H., Meulemeester, N. & Inzé, D. CORNET 2.0:
959 integrating plant coexpression, protein-protein interactions, regulatory interactions, gene
960 associations and functional annotations. *New Phytol.* **195**, 707–720 (2012).
961 88. A., S. & R., S. A systematic review of Explainable Artificial Intelligence models and
962 applications: Recent developments and future trends. *Decis. Anal. J.* **7**, 100230 (2023).
963 89. Garcia-Hernandez, M. *et al.* TAIR: a resource for integrated Arabidopsis data. *Funct. Integr.*
964 *Genomics* **2**, 239–253 (2002).
965 90. Lim, P. K., Wang, R. & Mutwil, M. LSTrAP-denovo: Automated Generation of Transcriptome
966 Atlases for Eukaryotic Species Without Genomes. *Physiol. Plant.* **176**, e14407 (2024).
967 91. Lim, P. K. *et al.* Bacteria.guru: Comparative Transcriptomics and Co-Expression Database
968 for Bacterial Pathogens. *J. Mol. Biol.* **434**, 167380 (2022).
969 92. Villanueva, E. M. F. *et al.* Protist.guru: A Comparative Transcriptomics Database for
970 Protists. *J. Mol. Biol.* **434**, 167502 (2022).
971 93. Monteiro, P. T. *et al.* YEASTRACT+: a portal for cross-species comparative genomics of
972 transcription regulation in yeasts. *Nucleic Acids Res.* **48**, D642–D649 (2020).
973 94. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory
974 interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
975 95. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse
976 transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
977 96. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
978 97. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in
979 Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, D262–D266 (2004).

980 98. Wong, E. D. *et al.* Saccharomyces genome database update: server architecture, pan-
981 genome nomenclature, and external resources. *Genetics* **224**, iyac191 (2023).
982 99. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways.
983 *Brief. Bioinform.* **20**, 1085–1093 (2019).
984 100. Romero, P. *et al.* Computational prediction of human metabolic pathways from the
985 complete human genome. *Genome Biol.* **6**, R2 (2005).
986 101. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
987 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995)

988

Acknowledgement

The authors wish to thank SingAREN (Singapore Advanced Research and Education Network) and Mr. Soh Hwee Jin, Melvin from the High Performance Computing Centre (HPCC) of Nanyang Technological University (NTU) for providing the network infrastructure used for downloading of public sequencing data. The authors also thank Asst. Prof. Jarkko Tapani Salojarvi and Asst. Prof. Goh Wen Bin Wilson from the School of Biological Sciences and Lee Kong Chian School of Medicine, respectively, of NTU for their input in the project. Lastly, the authors thank all members of the Mutwil Lab for their support in the project and comments on the manuscript. P.K.L is supported by the NTU Research Scholarship (RSS). M.M is supported by Singaporean Ministry of Education grant MOE-T2EP30122-0017.

Author information

Authors and Affiliations

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

Peng ken Lim, Ruoxi Wang, Jenet Princy Antony Velankanni and Marek Mutwil

Contributions

P.K.L and M.M conceived the project and wrote the manuscript. P.K.L developed and programmed code for the method. P.K.L and R.W performed data analysis and generated figures. J.P.A.V helped with the construction of edges. P.K.L. was supervised by M.M, while R.W and J.P.A were supervised by P.K.L. M.M secured funding for the project.

Corresponding authors

Correspondence to Marek Mutwil or Peng Ken Lim.

Supplemental Figures

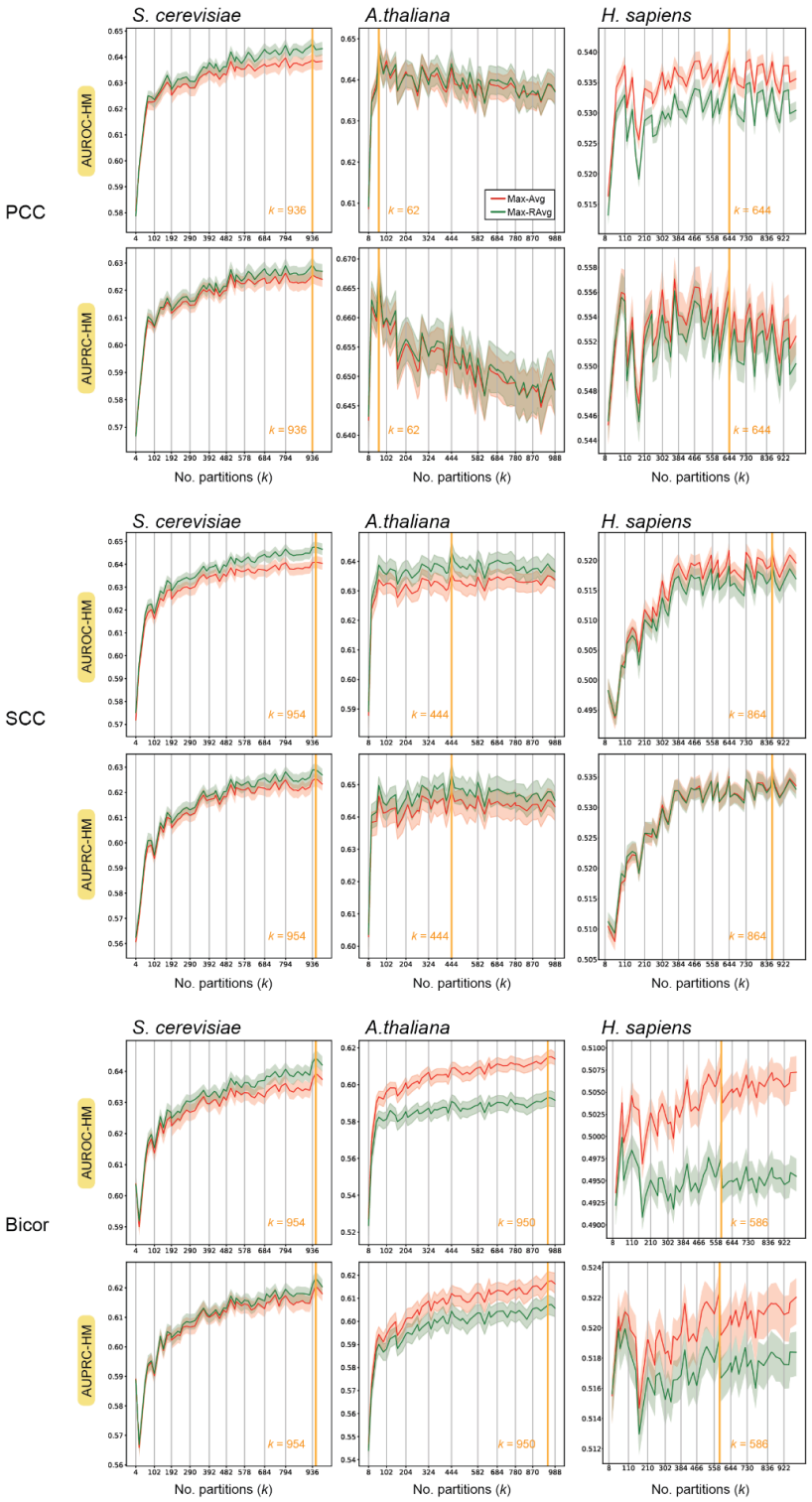


Figure S1. Overall performances of Partition Aggregation-only ensemble GCNs using different coefficients across different levels of partitioning granularity (*k*). The colors of solid lines correspond to the median AUCROC-HM and AUCPRC-HM scores of GCNs across 100 evaluations using aggregation functions RAvg or Avg for Partition Aggregation at different *k*s. Shaded regions of the same color mapping denote upper and lower quartiles of performance scores. The orange line denotes optimal *k* at which the RAvg Partition Aggregation-only GCNs has the highest sum of median AUCROC-HM and AUCPRC-HM scores.

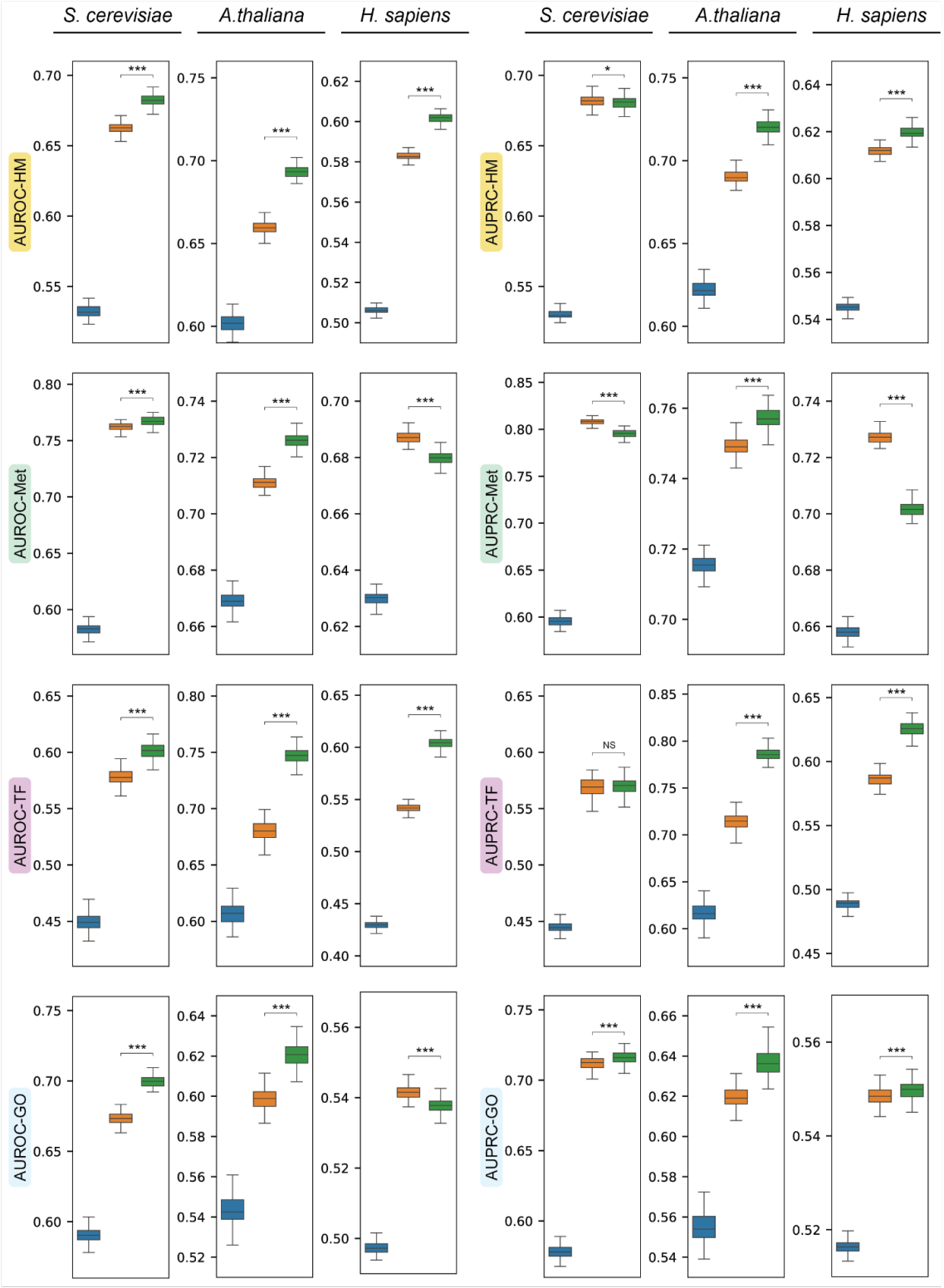


Figure S2. Comparison of TEA-GCN against "Subagging" ensemble GCNs in terms of performance across different biological aspects in every species. Blue, orange, and green boxes represents performance scores measured for non-ensemble PCC GCNs, "Subbagging" GCNs, and TEA-GCNs respectively in each species. Asterisks represents different levels of statistical significance (Two-sample student's T-test) calculated between score distributions of "Subbagging" GCNs and TEA-GCNs (NS: P＞0.05, one asterisk: 0.01 ≤ P ≤ 0.05, three asterisks: P ≤ 0.001

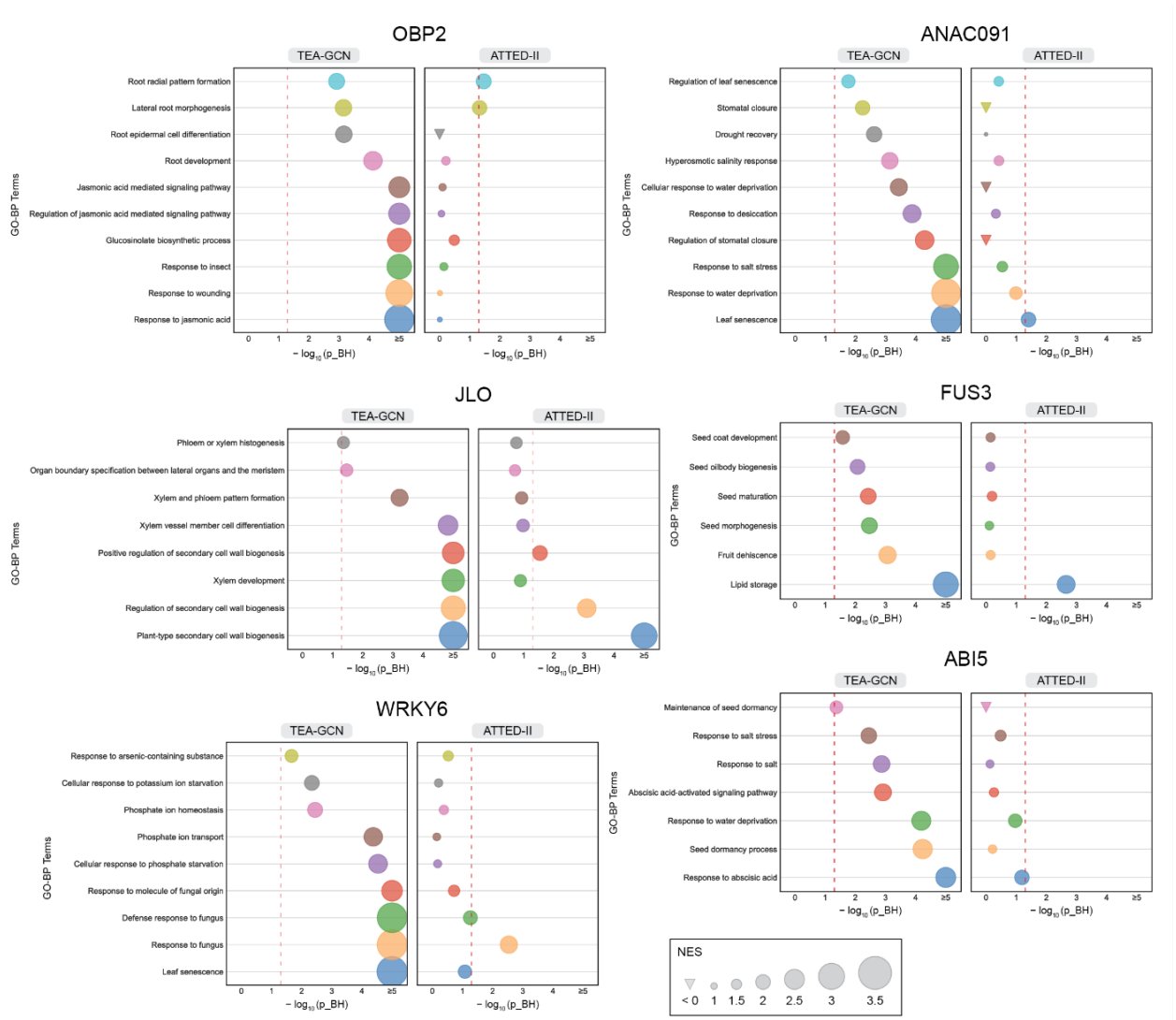
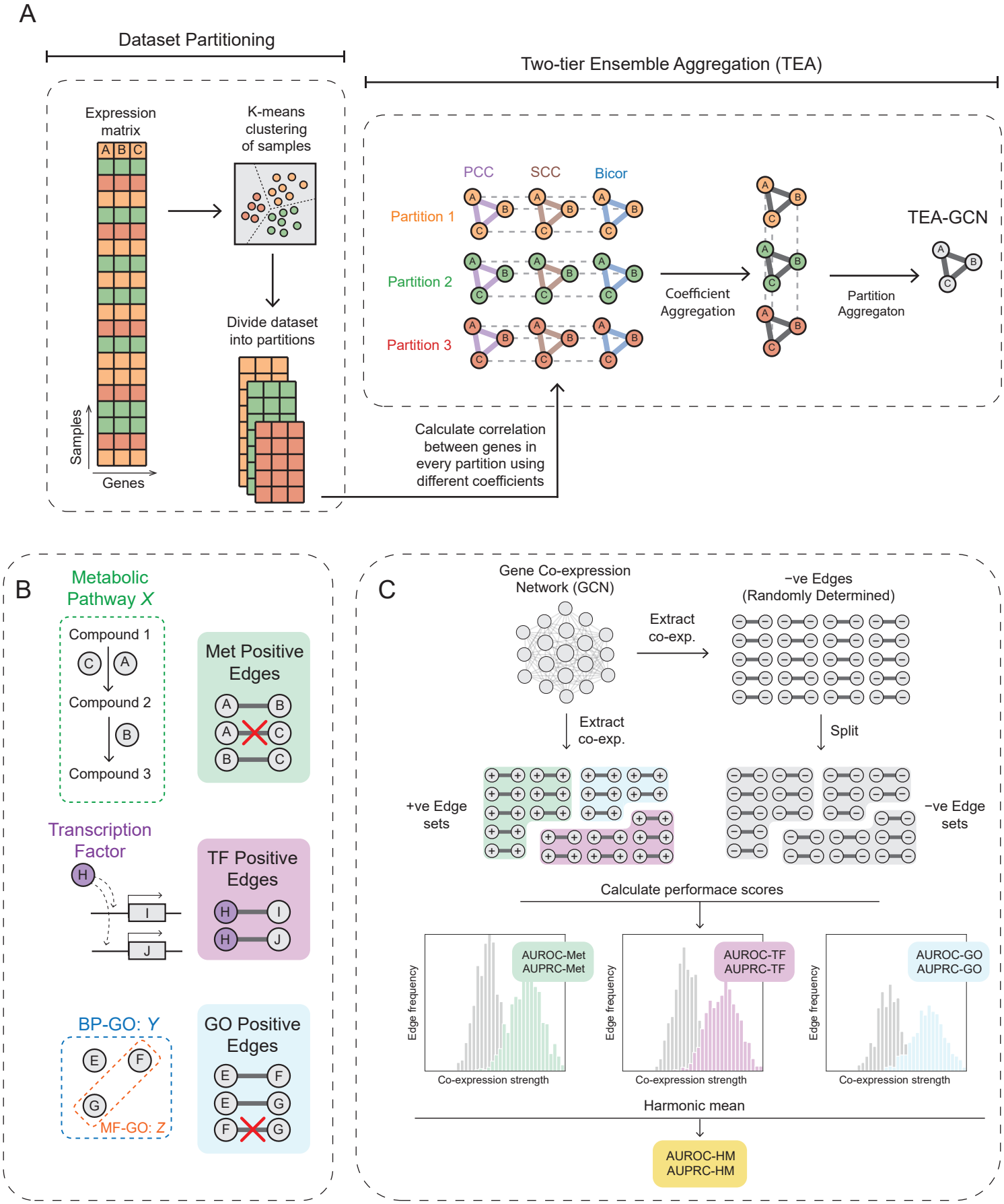
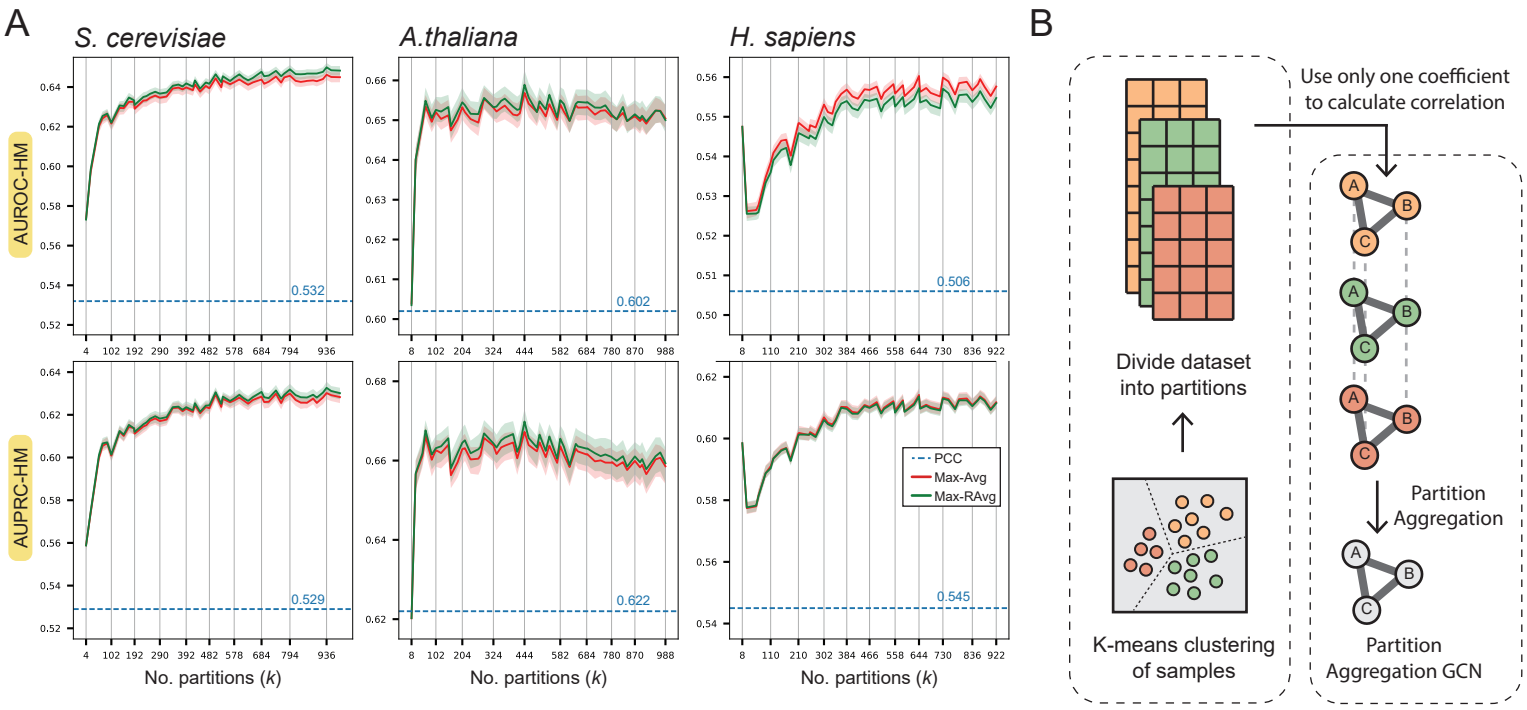


Figure S3. Comparing *A. thaliana* TEA-GCN and ATTED-II GCN co-expression neighbourhoods of selected transcription factors in terms of their functional enrichment profiles. GO terms were selected based on the known functions of the selected transcription factors. Dot sizes for GO terms correspond to normalized enrichment scores (NES) while inverted triangles represent CO terms with negative NES scores. For GO terms with positive NES scores, statistical significance in the form of Benjamini-Hochberg-corrected p-values (p_BH) can be inferred from dot placement on the x-axis. The red dotted lines denote the statistical significance threshold (p_BH ≥ 0.05).

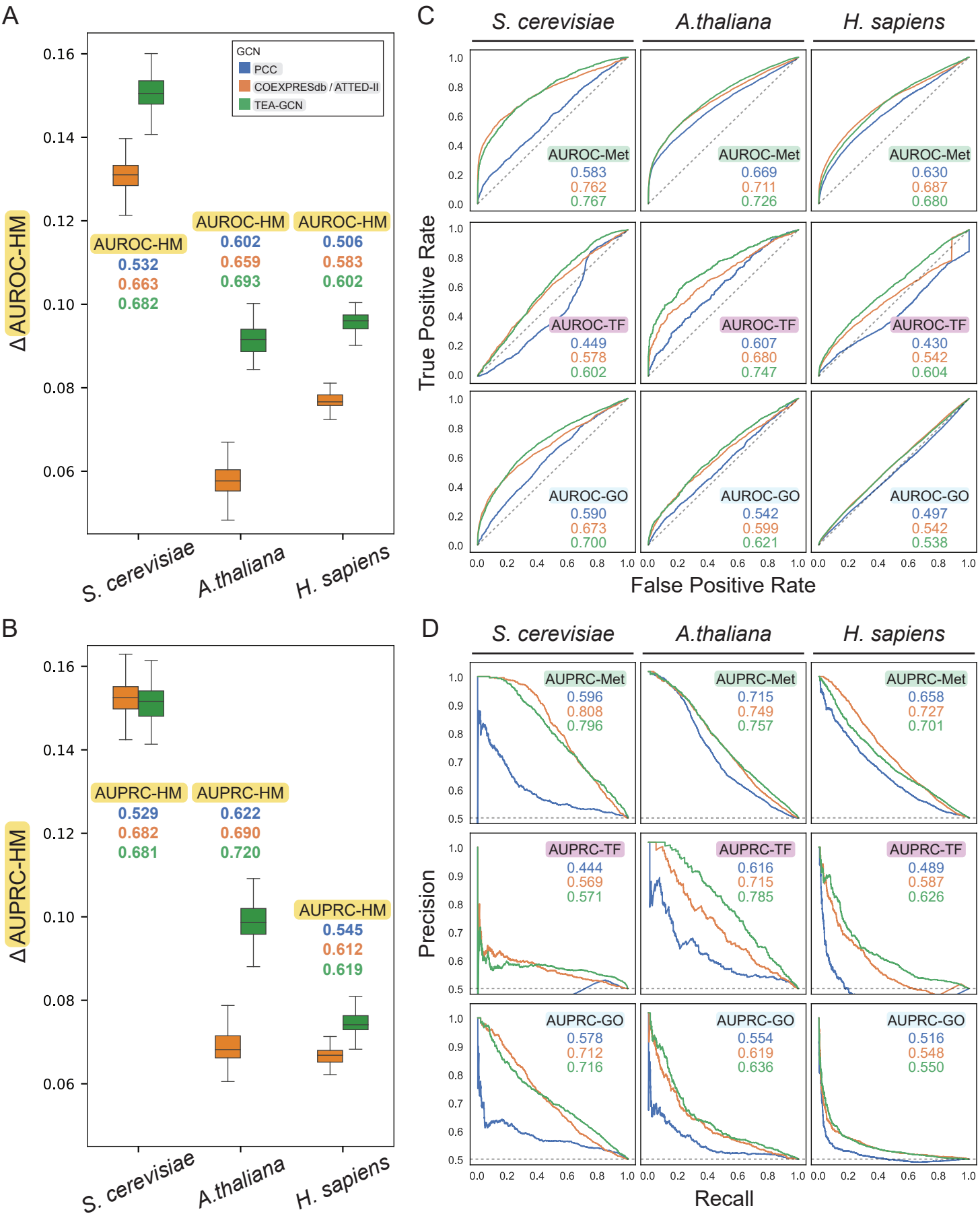

A
Dataset Partitioning
Two-tier Ensemble Aggregation (TEA)
Expression matrix
A B C
Samples
Genes
K-means clustering of samples
Divide dataset into partitions
PCC
SCC
Bicor
Partition 1
Partition 2
Partition 3
Coefficient Aggregation
Partition Aggregaton
TEA-GCN
Calculate correlation between genes in every partition using different coefficients
B
Metabolic Pathway X
Compound 1
Compound 2
Compound 3
Met Positive Edges
Transcription Factor
TF Positive Edges
BP-GO: Y
MF-GO: Z
GO Positive Edges
C
Gene Co-expression Network (GCN)
Extract co-exp.
−ve Edges (Randomly Determined)
Extract co-exp.
Split
+ve Edge sets
−ve Edge sets
Calculate performace scores
AUROC-Met
AUPRC-Met
AUROC-TF
AUPRC-TF
AUROC-GO
AUPRC-GO
Edge frequency
Co-expression strength
Harmonic mean
AUROC-HM
AUPRC-HM

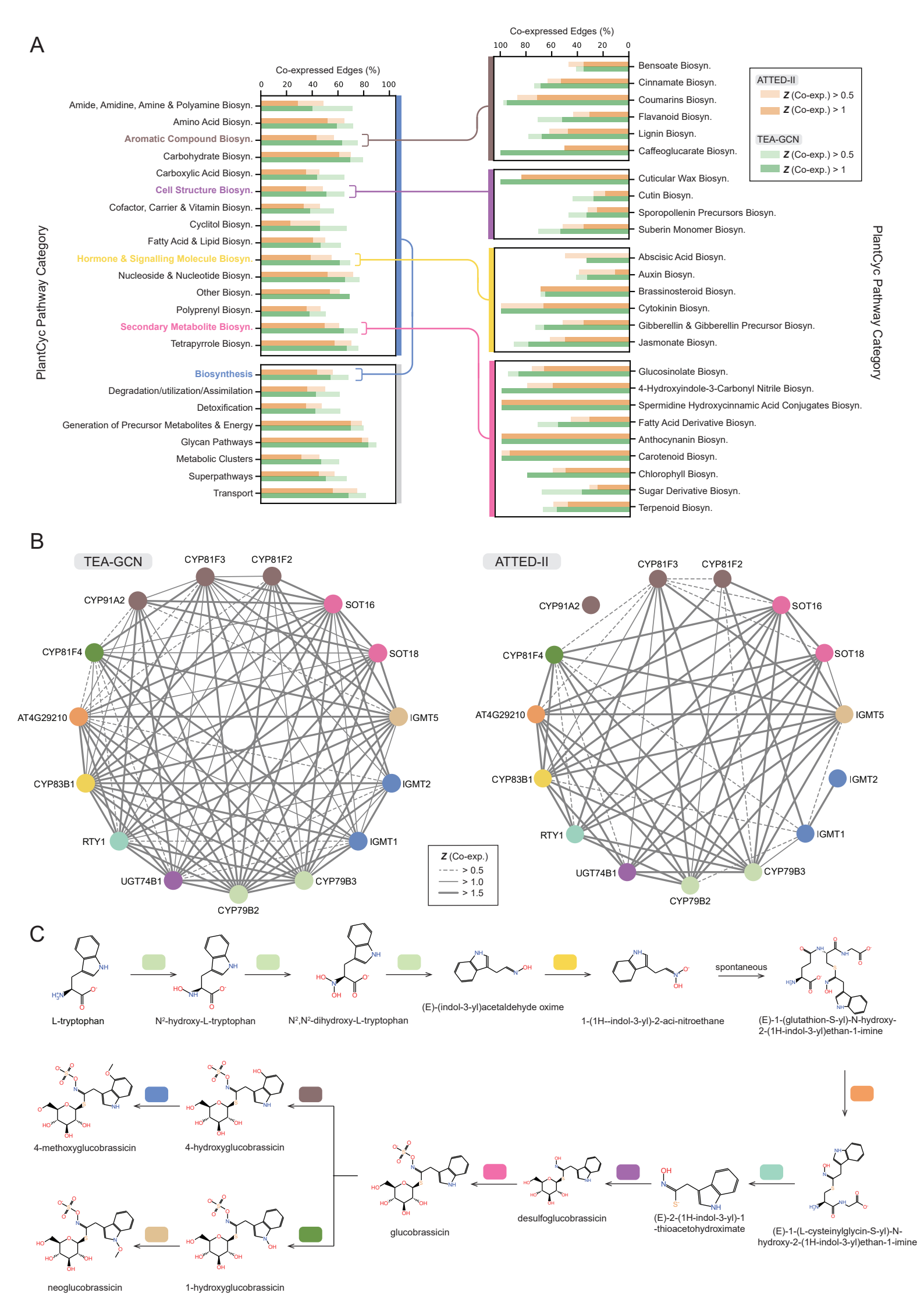


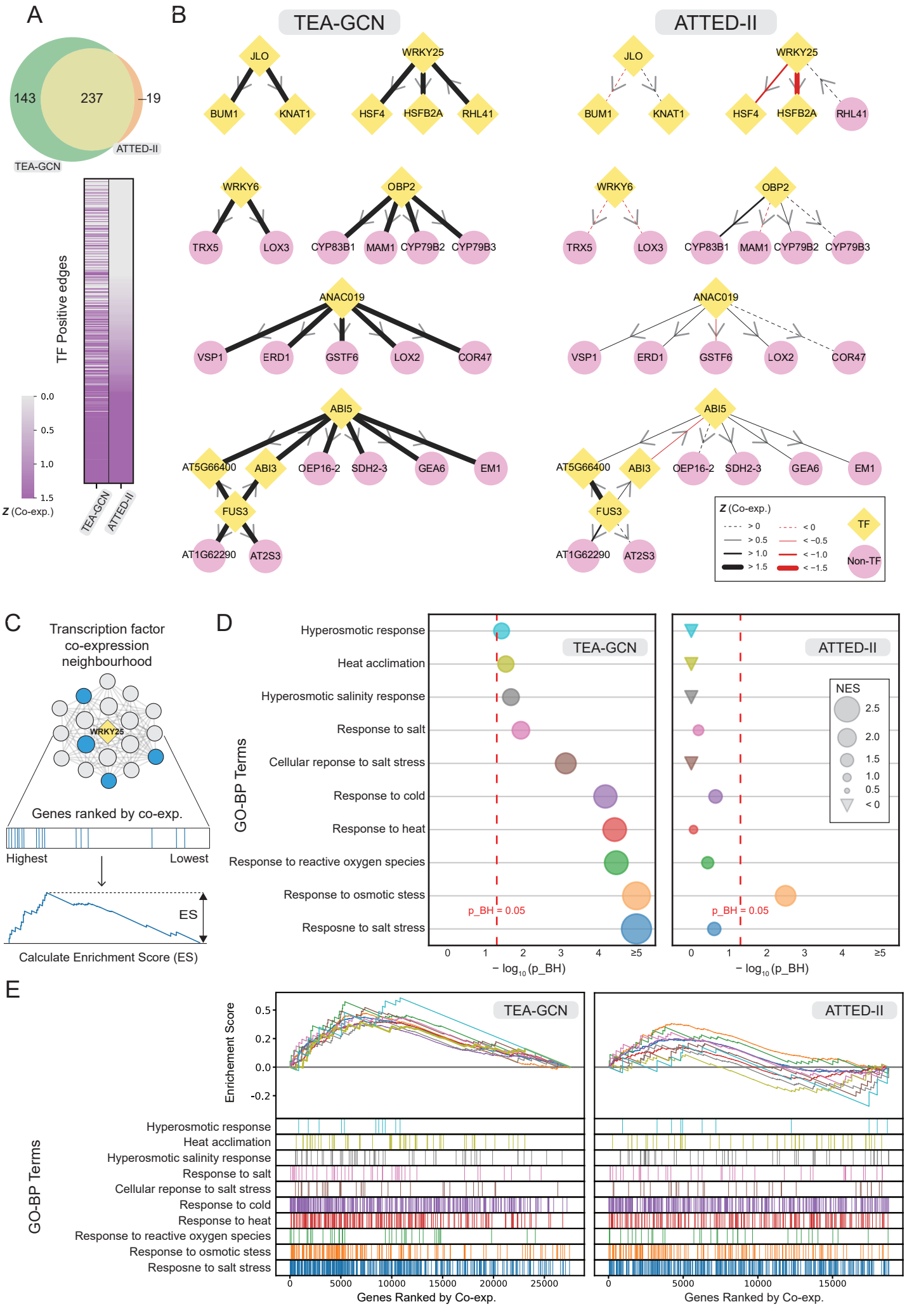
C

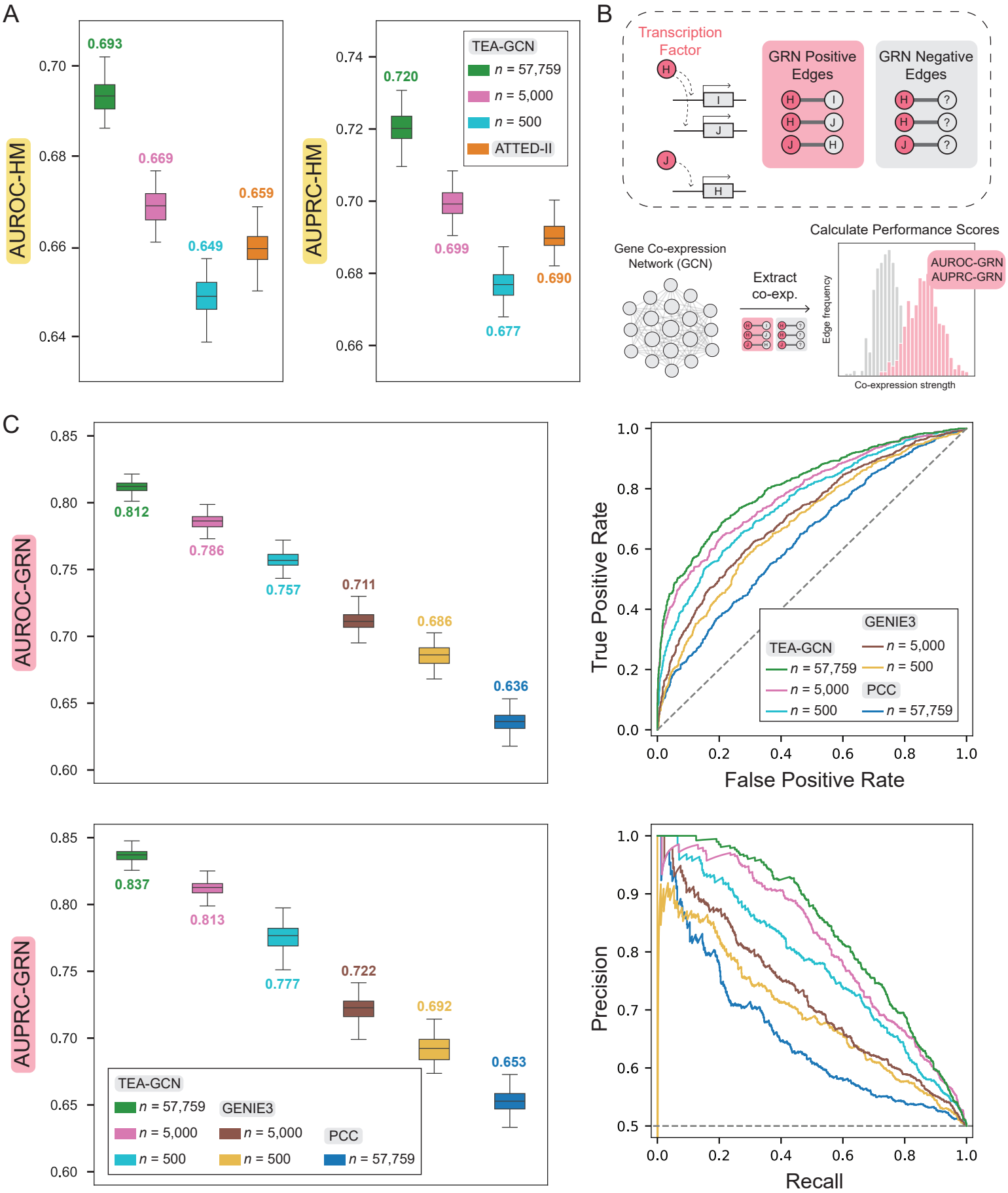
		HM		Met		TF		GO		
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	
S. cerevisiae	Max-Avg	0.6462	0.6302	0.6729	0.6632	0.5799	0.5485	0.6972	0.6978	Coefficient Aggregation + Partition Aggregation
	Max-Ravg	0.6500	0.6326	0.6751	0.6645	0.5881	0.5537	0.6971	0.6970	
	Max-Avg-MR	0.6764	0.6771	0.7634	0.7933	0.5926	0.5649	0.6962	0.7131	Mutual Rank Transformation
	Max-RAvg-MR	0.6822	0.6809	0.7671	0.7956	0.6016	0.5706	0.6997	0.7160	
	PCC-RAvg	0.6451	0.6294	0.6651	0.6612	0.5776	0.5454	0.7053	0.7027	Parrtition aggregation only
	SCC-RAvg	0.6476	0.6291	0.6714	0.6583	0.5837	0.5495	0.6982	0.6977	
	Bicor-RAvg	0.6444	0.6234	0.6667	0.6482	0.5787	0.5435	0.6996	0.6993	
A. thaliana	Max-Avg	0.6569	0.6672	0.7001	0.7217	0.6787	0.6885	0.6002	0.6028	
	Max-Ravg	0.6589	0.6698	0.7180	0.7329	0.6633	0.6810	0.6053	0.6064	
	Max-Avg-MR	0.6873	0.7162	0.7135	0.7496	0.7465	0.7845	0.6157	0.6332	
	Max-RAvg-MR	0.6933	0.7201	0.7260	0.7569	0.7471	0.7852	0.6207	0.6361	
	PCC-RAvg	0.6474	0.6669	0.7015	0.7205	0.6600	0.6939	0.5899	0.5994	
	SCC-RAvg	0.6432	0.6532	0.7123	0.7268	0.6327	0.6478	0.5946	0.5960	
	Bicor-RAvg	0.5931	0.6070	0.7096	0.7229	0.5172	0.5437	0.5814	0.5806	
H. sapiens	Max-Avg	0.5574	0.5553	0.6436	0.6439	0.5465	0.5388	0.5006	0.5023	
	Max-Ravg	0.5541	0.5544	0.6477	0.6471	0.5406	0.5362	0.4948	0.4998	
	Max-Avg-MR	0.6035	0.6194	0.6773	0.6991	0.6066	0.6260	0.5415	0.5513	
	Max-RAvg-MR	0.6020	0.6193	0.6799	0.7015	0.6043	0.6258	0.5377	0.5499	
	PCC-RAvg	0.5470	0.5777	0.6718	0.6860	0.4994	0.5502	0.5017	0.5218	
	SCC-RAvg	0.5195	0.5352	0.6439	0.6445	0.4577	0.4924	0.4907	0.4946	
	Bicor-RAvg	0.5687	0.5631	0.6781	0.6746	0.5759	0.5548	0.4842	0.4894	

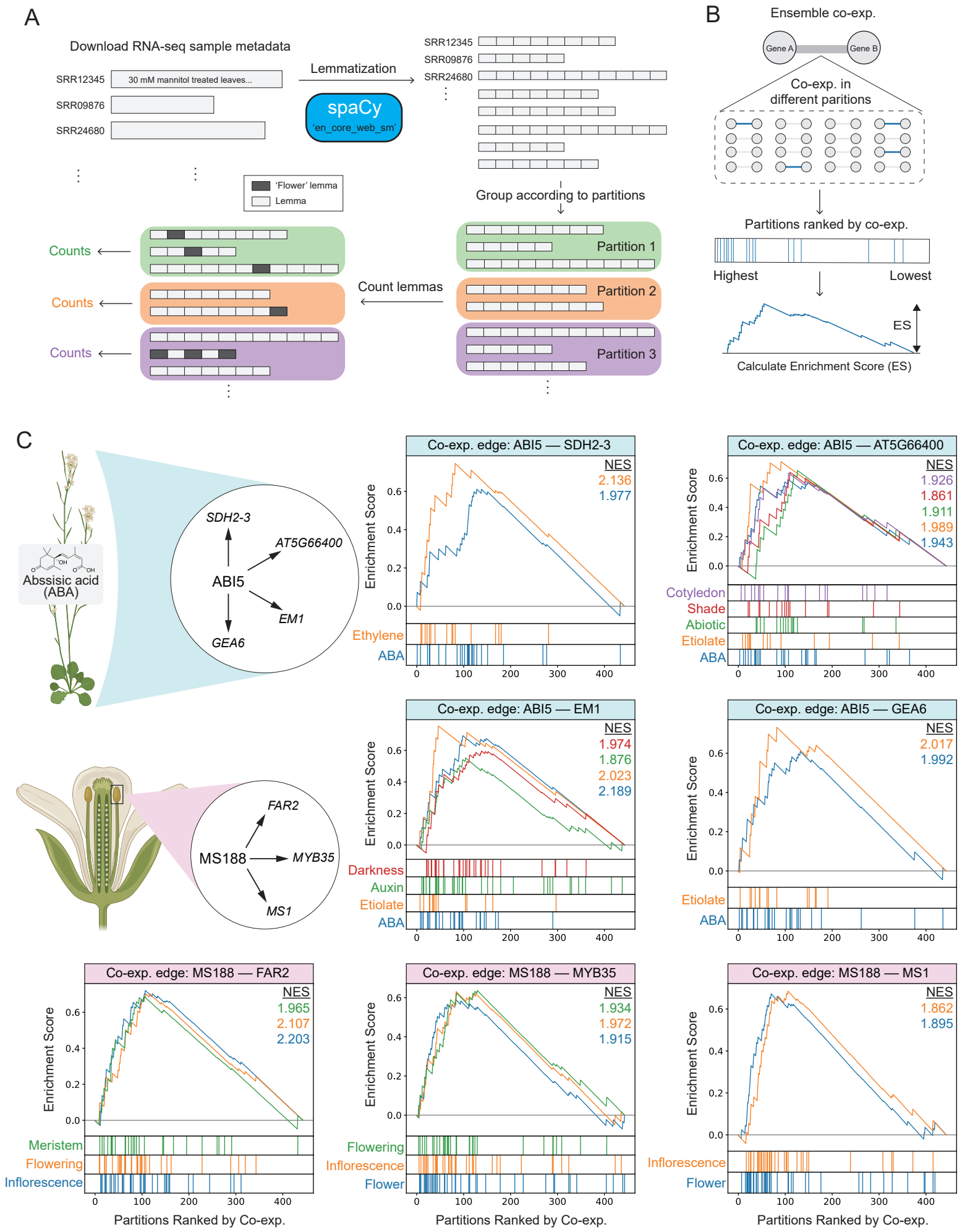
0.80 0.75 0.70 0.65 0.60 0.55 0.50 0.45
AUROC/PRC


A
GCN
PCC
COEXPRESdb / ATTED-II
TEA-GCN
ΔAUROC-HM
AUROC-HM
0.532
0.663
0.682
AUROC-HM
0.602
0.659
0.693
AUROC-HM
0.506
0.583
0.602
S. cerevisiae
A.thaliana
H. sapiens
B
ΔAUPRC-HM
AUPRC-HM
0.529
0.682
0.681
AUPRC-HM
0.622
0.690
0.720
AUPRC-HM
0.545
0.612
0.619
S. cerevisiae
A.thaliana
H. sapiens
C
S. cerevisiae
A.thaliana
H. sapiens
True Positive Rate
False Positive Rate
AUROC-Met
0.583
0.762
0.767
AUROC-Met
0.669
0.711
0.726
AUROC-Met
0.630
0.687
0.680
AUROC-TF
0.449
0.578
0.602
AUROC-TF
0.607
0.680
0.747
AUROC-TF
0.430
0.542
0.604
AUROC-GO
0.590
0.673
0.700
AUROC-GO
0.542
0.599
0.621
AUROC-GO
0.497
0.542
0.538
D
S. cerevisiae
A.thaliana
H. sapiens
Precision
Recall
AUPRC-Met
0.596
0.808
0.796
AUPRC-Met
0.715
0.749
0.757
AUPRC-Met
0.658
0.727
0.701
AUPRC-TF
0.444
0.569
0.571
AUPRC-TF
0.616
0.715
0.785
AUPRC-TF
0.489
0.587
0.626
AUPRC-GO
0.578
0.712
0.716
AUPRC-GO
0.554
0.619
0.636
AUPRC-GO
0.516
0.548
0.550


A
Co-expressed Edges (%)
0 20 40 60 80 100
Amide, Amidine, Amine & Polyamine Biosyn.
Amino Acid Biosyn.
Aromatic Compound Biosyn.
Carbohydrate Biosyn.
Carboxylic Acid Biosyn.
Cell Structure Biosyn.
Cofactor, Carrier & Vitamin Biosyn.
Cyclitol Biosyn.
Fatty Acid & Lipid Biosyn.
Hormone & Signalling Molecule Biosyn.
Nucleoside & Nucleotide Biosyn.
Other Biosyn.
Polyprenyl Biosyn.
Secondary Metabolite Biosyn.
Tetrapyrrole Biosyn.
Biosynthesis
Degradation/utilization/Assimilation
Detoxification
Generation of Precursor Metabolites & Energy
Glycan Pathways
Metabolic Clusters
Superpathways
Transport
PlantCyc Pathway Category
Co-expressed Edges (%)
100 80 60 40 20 0
Bensoate Biosyn.
Cinnamate Biosyn.
Coumarins Biosyn.
Flavanoid Biosyn.
Lignin Biosyn.
Caffeoglucarate Biosyn.
Cuticular Wax Biosyn.
Cutin Biosyn.
Sporopollenin Precursors Biosyn.
Suberin Monomer Biosyn.
Abscisic Acid Biosyn.
Auxin Biosyn.
Brassinosteroid Biosyn.
Cytokinin Biosyn.
Gibberellin & Gibberellin Precursor Biosyn.
Jasmonate Biosyn.
Glucosinolate Biosyn.
4-Hydroxyindole-3-Carbonyl Nitrile Biosyn.
Spermidine Hydroxycinnamic Acid Conjugates Biosyn.
Fatty Acid Derivative Biosyn.
Anthocynanin Biosyn.
Carotenoid Biosyn.
Chlorophyll Biosyn.
Sugar Derivative Biosyn.
Terpenoid Biosyn.
ATTED-II
Z (Co-exp.) > 0.5
Z (Co-exp.) > 1
TEA-GCN
Z (Co-exp.) > 0.5
Z (Co-exp.) > 1
PlantCyc Pathway Category
B
TEA-GCN
CYP81F3
CYP81F2
CYP91A2
SOT16
CYP81F4
SOT18
AT4G29210
IGMT5
CYP83B1
IGMT2
RTY1
IGMT1
UGT74B1
CYP79B3
CYP79B2
Z (Co-exp.)
> 0.5
> 1.0
> 1.5
ATTED-II
CYP81F3
CYP81F2
CYP91A2
SOT16
CYP81F4
SOT18
AT4G29210
IGMT5
CYP83B1
IGMT2
RTY1
IGMT1
UGT74B1
CYP79B3
CYP79B2
C
L-tryptophan
N²-hydroxy-L-tryptophan
N²,N²-dihydroxy-L-tryptophan
(E)-(indol-3-yl)acetaldehyde oxime
1-(1H--indol-3-yl)-2-aci-nitroethane
spontaneous
(E)-1-(glutathion-S-yl)-N-hydroxy-2-(1H-indol-3-yl)ethan-1-imine
4-methoxyglucobrassicin
4-hydroxyglucobrassicin
glucobrassicin
desulfoglucobrassicin
(E)-2-(1H-indol-3-yl)-1-thioacetohydroximate
(E)-1-(L-cysteinylglycin-S-yl)-N-hydroxy-2-(1H-indol-3-yl)ethan-1-imine
neoglucobrassicin
1-hydroxyglucobrassicin


A
143
237
19
TEA-GCN
ATTED-II
TF Positive edges
0.0
0.5
1.0
1.5
Z (Co-exp.)
TEA-GCN
ATTED-II
B
TEA-GCN
ATTED-II
JLO
BUM1
KNAT1
WRKY25
HSF4
HSFB2A
RHL41
WRKY6
TRX5
LOX3
OBP2
CYP83B1
MAM1
CYP79B2
CYP79B3
ANAC019
VSP1
ERD1
GSTF6
LOX2
COR47
ABI5
AT5G66400
ABI3
OEP16-2
SDH2-3
GEA6
EM1
FUS3
AT1G62290
AT2S3
Z (Co-exp.)
> 0
> 0.5
> 1.0
> 1.5
< 0
< −0.5
< −1.0
< −1.5
TF
Non-TF
C
Transcription factor co-expression neighbourhood
WRKY25
Genes ranked by co-exp.
Highest
Lowest
ES
Calculate Enrichment Score (ES)
D
GO-BP Terms
Hyperosmotic response
Heat acclimation
Hyperosmotic salinity response
Response to salt
Cellular reponse to salt stress
Response to cold
Response to heat
Response to reactive oxygen species
Response to osmotic stess
Resposne to salt stress
TEA-GCN
ATTED-II
NES
2.5
2.0
1.5
1.0
0.5
< 0
p_BH = 0.05
0
1
2
3
4
≥5
− log10 (p_BH)
E
Enrichement Score
0.5
0.2
0.0
-0.2
GO-BP Terms
Genes Ranked by Co-exp.
0
5000
10000
15000
20000
25000


A
AUROC-HM
0.693
0.669
0.649
0.659
AUPRC-HM
0.720
0.699
0.677
0.690
TEA-GCN
n = 57,759
n = 5,000
n = 500
ATTED-II
B
Transcription Factor
GRN Positive Edges
GRN Negative Edges
Gene Co-expression Network (GCN)
Extract co-exp.
Calculate Performance Scores
AUROC-GRN
AUPRC-GRN
Edge frequency
Co-expression strength
C
AUROC-GRN
0.812
0.786
0.757
0.711
0.686
0.636
True Positive Rate
False Positive Rate
GENIE3
TEA-GCN
n = 5,000
n = 57,759
n = 500
PCC
n = 5,000
n = 500
n = 57,759
AUPRC-GRN
0.837
0.813
0.777
0.722
0.692
0.653
Precision
Recall
TEA-GCN
n = 57,759
n = 5,000
n = 500
GENIE3
n = 5,000
n = 500
PCC
n = 57,759


A
Download RNA-seq sample metadata
SRR12345
30 mM mannitol treated leaves...
SRR09876
SRR24680
Lemmatization
spaCy
'en_core_web_sm'
SRR12345
SRR09876
SRR24680
'Flower' lemma
Lemma
Group according to partitions
Partition 1
Partition 2
Partition 3
Count lemmas
Counts
Counts
Counts
B
Ensemble co-exp.
Gene A
Gene B
Co-exp. in different paritions
Partitions ranked by co-exp.
Highest
Lowest
ES
Calculate Enrichment Score (ES)
C
Abssisic acid (ABA)
SDH2-3
AT5G66400
ABI5
EM1
GEA6
MS188
FAR2
MYB35
MS1
Co-exp. edge: ABI5 — SDH2-3
NES
2.136
1.977
Ethylene
ABA
Co-exp. edge: ABI5 — AT5G66400
NES
1.926
1.861
1.911
1.989
1.943
Cotyledon
Shade
Abiotic
Etiolate
ABA
Co-exp. edge: ABI5 — EM1
NES
1.974
1.876
2.023
2.189
Darkness
Auxin
Etiolate
ABA
Co-exp. edge: ABI5 — GEA6
NES
2.017
1.992
Etiolate
ABA
Co-exp. edge: MS188 — FAR2
NES
1.965
2.107
2.203
Meristem
Flowering
Inflorescence
Co-exp. edge: MS188 — MYB35
NES
1.934
1.972
1.915
Flowering
Inflorescence
Flower
Co-exp. edge: MS188 — MS1
NES
1.862
1.895
Inflorescence
Flower
Enrichment Score
Partitions Ranked by Co-exp.