Usefulness of scRNA-seq data in predicting plant metabolic pathway genes

Jingwei Ma[1,2‡], Zhenglin Wang[1‡], Liting Zou[1‡], Xiaoxue Wang[1,2], Xinyi Zuo[1,3], Fei Wang[1,3], Zhiqing Wang[1,4], Zhimei Li[1], Lin Li[2]*, Peipei Wang[1]*

[1] Kunpeng Institute of Modern Agriculture at Foshan, Foshan, Guangdong 528200, China

[2] College of Life Science, Shanxi Normal University, Taiyuan, Shanxi 030031, China

[3] College of Horticulture, Northwest A&F University, Yangling, Shaanxi 712100, China

[4] College of Plant Protection, South China Agricultural University, Guangzhou, Guangdong 510642, China

Abstract

It is an ever challenging task to make genome-wide predictions for plant metabolic pathway genes (MPGs) encoding enzymes that catalyze the biosynthesis of plant natural products. Here, starting from 1,129 benchmark MPGs that have experimental evidence in *Arabidopsis thaliana*, we investigate the utilities of single-cell RNA sequencing (scRNA-seq) data—a recently arisen omics data that has been used in several other fields—in predicting MPGs using five machine learning (ML) algorithms that support multi-label tasks. Compared with traditional bulk RNA-seq data, scRNA-seq data lead to different but comparable co-expression networks among MPGs within metabolic classes, and significantly higher prediction accuracy of MPGs into classes. Prediction accuracy for individual metabolic classes is not associated with the co-expression network tightness, but correlated with the number of MPGs within each class, indicating that including more benchmark genes in the future will improve the MPG prediction. Splitting the RNA-seq data into genetic background/condition or tissue-specific subsets can improve the gene co-expression network tightness and MPG prediction accuracy for some classes; scRNA-seq-based models still outperform bulk RNA-seq-based models for most classes when corresponding subsets are used. In addition, deep learning approaches outperform classical machine learning approaches; approaches implemented in an ensembled workflow AutoGluon tend to have severe overfitting issues potentially due to the relative scarcity of benchmark MPGs within classes. Our results demonstrate the superiority of scRNA-seq data over bulk RNA-seq data in predicting MPGs into metabolic classes, and propose that scRNA-seq data should be included in the future to advance the identification of plant MPGs.

Introduction

As sessile organisms, plants protect themselves from herbivores and stresses mainly via producing plant natural products (PNPs)[1]. Some PNPs, such as anthocyanin[2], paclitaxel[3], salicylic acid[4,5] and pyrethroids[6], have also been widely used in human nutrition, medicine, cosmetics, agriculture and other fields[7–10]. The content of these natural products in plants is generally quite low, and the extraction of them is usually inefficient and environmentally unfriendly[11]. Thus, synthesizing PNPs in engineered microbial cells or plant cells has become a hotspot in synthetic biology, which requires the elucidation of metabolic pathway genes (MPGs) that encode enzymes catalyzing the biosynthesis of these PNPs. However, only about 0.1% of metabolic pathways of the estimated ~1 million metabolites synthesized by plants have been revealed[12]. Before the genome era, the identification of plant MPGs relied heavily on traditional genetic and/or biochemical approaches, which are time- and labor-consuming. For example, the identification of the genes participating in the biosynthesis pathway of avenacin spanned about 22 years, from the isolation of saponin-deficient (*sad*) mutants[13] to the identification of the 12-gene cluster[14].

With the development of omics technologies, more and more metabolic pathways have been deciphered by researchers via approaches such as genome-wide association studies (GWAS)[15,16] and association analysis between transcriptome and metabolome[17,18]. Although these approaches have improved the efficiency and accuracy of MPG identification, they generally target one or a few pathways with relatively low throughput, but require large-scale genotyping or RNA-sequencing, in company with metabolomic profiling. The development of computational tools, such as those for identifying plant metabolic gene clusters[12,19,20] or annotating enzyme genes[12], and machine learning (ML) approaches utilizing one or multiple types of omics data[21,22], has made it possible to predict plant MPGs on a larger scale. In one of our previous studies, we investigated the utility of transcriptome data in predicting MPGs in tomato using ML approaches, based on the hypothesis that genes in the same metabolic pathways may be transcripted similarly to cooperatively synthesize the target metabolites[22]. Although the ML approaches outperformed naive methods that solely compared expression similarities among genes in tomato, the global prediction accuracy was still not good enough (the median optimal F1 was 0.4 for 85 pathways) even though exhaustive combinations of the ways how transcriptome data were used and types of ML algorithms were evaluated. We reasoned that the failure in predicting tomato MPGs using

transcriptome data may be partially due to the potential shortcomings of the traditional bulk RNA sequencing (bulk RNA-seq) data used in the study: the RNA sequencing is generally processed for a mixed pool of multiple types of cells within the sampled tissues. Considering that the biosynthesis of many PNPs are cell type-specific[23], and that single-cell RNA sequencing (scRNA-seq) data have been used to explore the spatial organization of PNPs and the transcripts of contributing MPGs[23,24], we hypothesize that the prediction of plant MPGs may be potentially improved by using scRNA-seq data compared with bulk RNA-seq data.

Here, we compare the utilities of scRNA-seq vs. bulk RNA-seq data in predicting MPGs in *Arabidopsis thaliana*, for which there are the most comprehensive MPGs with experimental evidence (hereafter referred to as benchmark MPGs) and scRNA-seq data among other plant species. We first compare the degrees of co-expression levels for benchmark MPGs within the same metabolic classes using these two types of RNA-seq data. Next, we establish ML models for predicting MPGs into metabolic classes using these two data, employing five ML algorithms that support multi-class and multi-label classification. Finally, we apply our models to enzyme genes that lack experimental evidence, and compare the predicted metabolic classes of these genes with the annotated classes of them in the Plant Metabolic Network (PMN) database. Examples are also given to demonstrate the usefulness of our models in MPG prediction and the superiority of scRNA-seq data over bulk RNA-seq data.

Results

Different co-expression networks of benchmark metabolic pathway genes based on two types of RNA-seq data

Both scRNA-seq and bulk RNA-seq data were collected from Arabidopsis plants—either wild type (WT) or non-WT—grown under normal or stress conditions, with a total of 21 scRNA-seq datasets comprising 259 cell types (**Table S1**) and 19 bulk RNA-seq datasets encompassing 34 samples (**Table S2**) from same tissues, same genotypes and under same conditions (sample types see **Table S3,4**). We obtained the annotations for Arabidopsis MPGs from the PMN database[12], and only retained 1,129 benchmark MPGs that have experimental evidence and belong to 636 metabolic pathways (**Table S5**). Considering that the majority of these 636 metabolic pathways (75.47%) only contained ≤ 5 benchmark MPGs (**Fig. 1A**), we grouped them

into 13 metabolic classes according to a previous paper[25] (**Table S5**) to facilitate the following analysis. This classification resulted in an average number of MPGs at 121 across 13 classes (a MPG can belong to multiple classes). Additionally, 34 genes from 37 pathways that do not belong to any of the 13 classes were assigned to the category "Others" (**Fig. 1B**, **Table S5**).

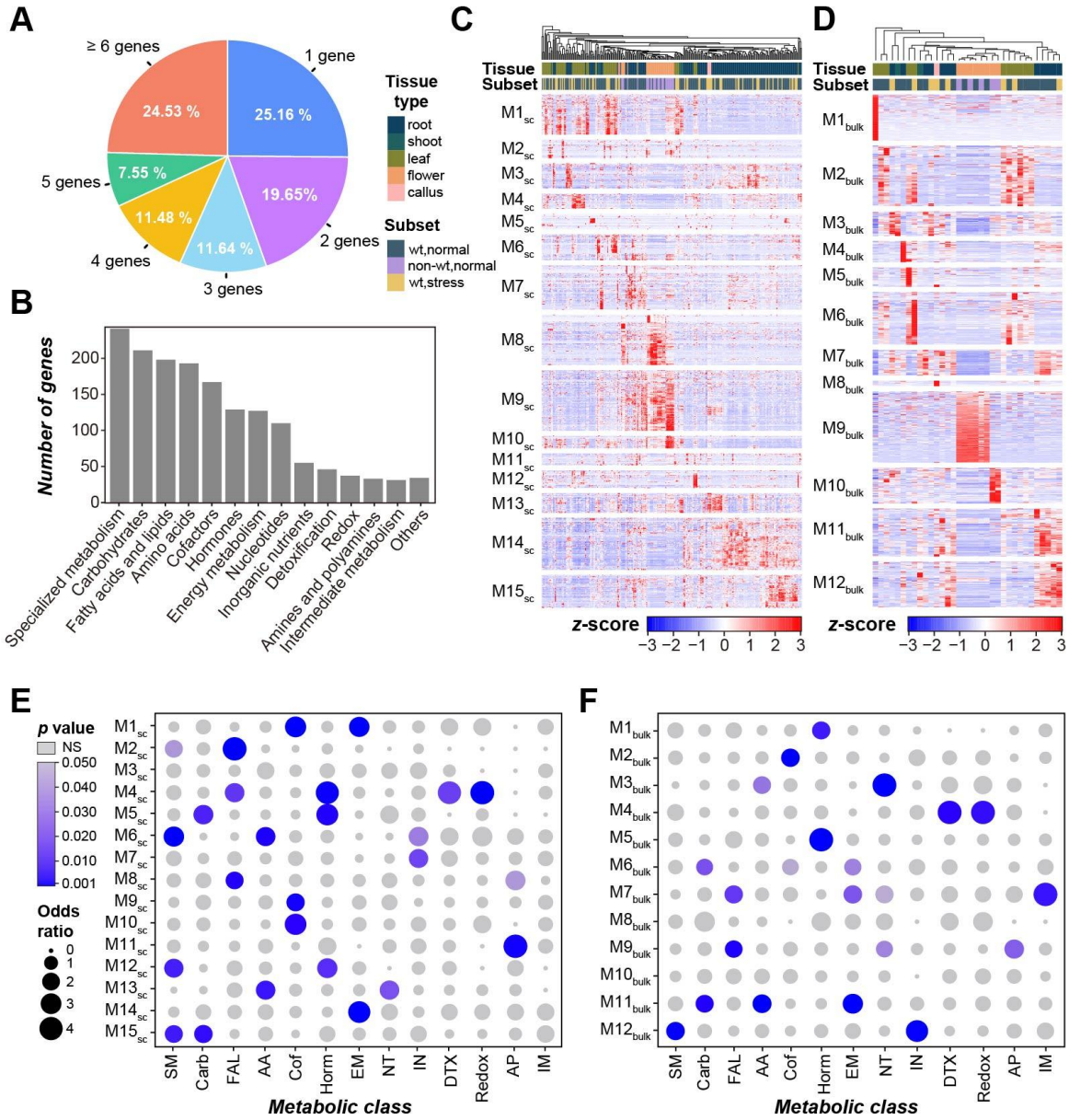


Fig. 1. Different expression patterns for genes within metabolic classes revealed using single-cell RNA-seq (scRNA-seq) and bulk RNA-seq data. (**A**) A pie plot showing the percentages of 636 metabolic pathways that have a certain number of benchmark metabolic pathway genes (MPGs): blue, 1 MPG; purple, 2; cyan, 3; orange, 4; green, 5; red, ≥ 6 MPGs. (**B**) A bar chart showing the numbers of MPGs belonging to 13 metabolic classes and the "Others" category that contains all the other pathways

not belonging to any classes. (**C,D**) Heatmaps showing the expression patterns of 1,129 MPGs, which were grouped into 15 and 12 co-expression modules ($M1_{sc}$–$M15_{sc}$ for **C**, $M1_{bulk}$–$M12_{bulk}$ for **D**) inferred by WGCNA using the scRNA-seq and bulk RNA-seq data, respectively. Color scale in the heatmaps: the standardized expression values (transcripts per kilobase million) using *z*-score normalization; red: high expression, blue: low expression; all values > 3 were assigned a value of 3. Tissue types and subset types that scRNA-seq cell types or bulk RNA-seq samples belong to are indicated above the heatmaps with colors. For tissue types: dark green, root; green, shoot; olive, leaf; orange, flower; pink, callus. For subset types: blue, wild type (WT) plants grown under normal conditions; purple, non-WT under normal conditions; yellow, WT under stress conditions. (**E,F**) Bubble plots showing enrichment of MPGs from 13 different metabolic classes (columns) within 15 and 12 co-expression modules (rows) using scRNA-seq (**E**) and bulk RNA-seq (**F**) data, respectively. Bubble size: odds ratio from Fisher's exact test, with all odds ratio > 4 being assigned a value of 4; color scale: *p*-values for over-representative cases (with odds ratio > 1); all *p*-values ≥ 0.05 were shaded gray. SM: specialized metabolism; Carb: carbohydrates; FAL: fatty acids and lipids; AA: amino acids; Cof: cofactors; Horm: hormones; EM: energy metabolism; NT: nucleotides; IN: inorganic nutrients; DTX: detoxification; AP: amines and polyamines; IM: intermediate metabolism.

To compare how well genes within different metabolic classes can be distinguished using scRNA-seq and bulk RNA-seq data, we established co-expression networks for benchmark MPGs using the weighted gene co-expression network analysis (WGCNA) (**Methods**). These 1,129 MPGs were clustered into 15 co-expression modules by scRNA-seq data ($M1_{sc}$–$M15_{sc}$, **Fig. 1C**, **Table S6**) and 12 modules by bulk RNA-seq data ($M1_{bulk}$–$M12_{bulk}$, **Fig. 1D**, **Table S7**), respectively. MPGs from each individual metabolic class were enriched in 1~4 co-expression modules, except for intermediate metabolites (IM) when scRNA-seq were used (**Fig. 1E,F**), indicating that MPGs from different classes can be potentially distinguished by using these two types of RNA-seq data.

However, the gene components were mostly different between M_{sc} and M_{bulk} modules, even for those that were overrepresented by MPGs from the same metabolic classes and had the same tissue-specific gene expression profiles. For example, MPGs from specialized metabolism (SM) were enriched in $M2_{sc}$, $M6_{sc}$, $M12_{sc}$ and $M15_{sc}$ (containing 16, 27, 17 and 28 SM MPGs, respectively), and $M12_{bulk}$ (containing 39 SM MPGs, **Fig. 1E,F**, **Table S6,7**). There were 2 and 10 SM MPGs shared between $M6_{sc}$ and $M12_{bulk}$, and between $M15_{sc}$ and $M12_{bulk}$, respectively, but no overlapping SM MPGs between $M2_{sc}$ or $M12_{sc}$ and $M12_{bulk}$ (**Fig. S1A**). MPGs in $M15_{sc}$

133 and $M12_{bulk}$ were preferentially expressed in root-related cell types or tissues; MPGs in $M6_{sc}$
134 tended to be expressed in broad types of tissues; MPGs in $M2_{sc}$ were preferentially expressed in
135 leaf and flower-related cell types, while MPGs in $M12_{sc}$ tended to be expressed in leaf-related
136 cell-types (**Fig. 1C,D**). Another example is the hormones class, MPGs from which were
137 overrepresented in $M4_{sc}$ (containing 12 hormones MPGs), $M5_{sc}$ (11), $M12_{sc}$ (11), $M1_{bulk}$ (22) and
138 $M5_{bulk}$ (20). MPGs in $M4_{sc}$, $M12_{sc}$, $M1_{bulk}$ and $M5_{bulk}$ tended to be preferentially expressed in leaf
139 tissues/cell types, while those in $M5_{sc}$ tended to be expressed in root-related cell types (**Fig.**
140 **1C,D**). All hormones MPGs in $M4_{sc}$ were also grouped in $M5_{bulk}$, but only 4 and 2 hormone
141 MPGs were shared between $M5_{sc}$ and $M1_{bulk}$, and between $M12_{sc}$ and $M5_{bulk}$, respectively (**Fig.**
142 **S1B**). These results illustrate the differences in co-expression networks among MPGs revealed
143 by these two types of data.

144 Tighter co-expression networks among MPGs within metabolic classes 145 revealed by tissue-specific scRNA-seq data

146 To better assess the tightness of the gene co-expression networks for individual metabolic
147 classes, we calculated the clustering coefficients (*C* values) for MPGs within 13 metabolic
148 classes using these two types of RNA-seq data (see **Methods**). The *C* value is normally used to
149 measure the degree to which nodes in a graph tend to cluster together[26], and has been widely
150 used in several real-world networks, particularly in social networks[27]; the higher the *C* value, the
151 tighter the network. To illustrate how well the *C* value reflects the tightness of co-expression
152 network within metabolic classes (**Table S8**), we further calculated the *z*-score of the observed *C*
153 value in 10,000 background *C* values where the metabolic class annotations of all the 1,129
154 MPGs were randomly shuffled (**Table S9**, see **Methods**). A *z*-score>1.645 (with a corresponding
155 *p*-value<0.05) indicates that MPGs in a class are significantly more tightly co-expressed than
156 random gene groups (upper inset in **Fig. 2A**), whereas a negative *z*-score indicates not tightly co-
157 expressed (lower inset in **Fig. 2A**).

158 We found that only the energy metabolism (EM), cofactors, and redox classes had *z*-
159 scores >1.645 both when two types of RNA-seq data were used (**Fig. 2A**, **Fig. S2**); another two
160 (nucleotides [NT] and IM) and two (detoxification [DTX] and amino acids [AA]) classes had *z*-
161 scores >1.645 only when scRNA-seq or bulk RNA-seq data were used, respectively. All the
162 remaining six classes had *z*-scores <1.645 regardless of the types of RNA-seq data used. This is

inconsistent with enrichment of MPGs in co-expressed modules (**Fig. 1E,F**). Considering that some PNPs may be synthesized in specific tissues/cell types or under particular conditions[28], we hypothesize that co-expression signals of MPGs synthesizing these natural products may be diluted by combining RNA-seq data from all the samples.

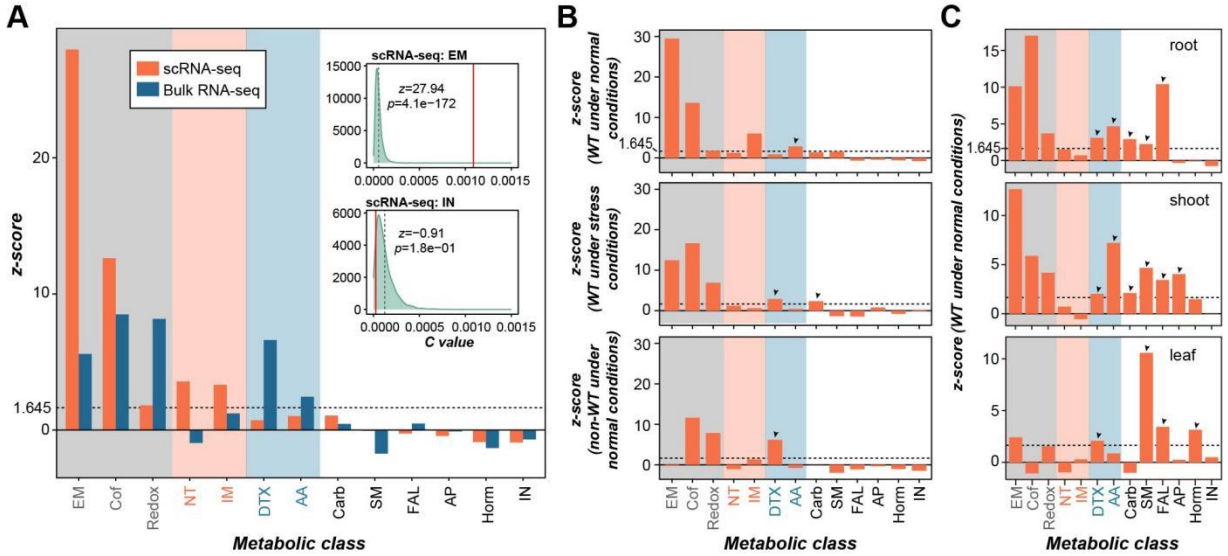


Fig. 2. Clustering coefficients (*C* values) for genes within 13 metabolic classes calculated using two types of RNA-seq data. (**A**) The *z*-scores of *C* values for metabolic pathway genes (MPGs) within 13 metabolic classes compared with background values (based on 10,000 simulations where the metabolic classes of all MPGs were shuffled randomly), when scRNA-seq (orange) and bulk RNA-seq (blue) data were used. The dashed horizontal line indicates a *z*-score of 1.645, corresponding to a *p*-value of 0.05. Colored backgrounds: metabolic classes with significantly higher *C* values than background values when both types of RNA-seq data were used (light grey); only scRNA-seq (light pink) and bulk RNA-seq data (light blue) were used. The two inset density plots show the distribution of 10,000 background *C* values for genes in the energy metabolism (EM, upper) and inorganic nutrients (IN, lower) classes as examples when scRNA-seq data were used. Red solid vertical line: the observed *C* value for genes within EM or IN class; black dashed vertical line: the average *C* value of 10,000 background values. *z*: *z*-score of the observed *C* value in the background distribution; *p*: the corresponding *p*-value of the *z*-score. (**B**) *Z*-scores of *C* values for MPGs within 13 metabolic classes when three subsets of scRNA-seq data (upper: WT plants grown under normal conditions; middle: WT plants under stress conditions; lower: non-WT under normal conditions) were used. (**C**) *Z*-scores of *C* values for MPGs within 13 metabolic classes when three tissue-specific subsets of WT-normal scRNA-seq data (upper: root; middle: shoot; lower: leaf) were used. Black arrow heads: additional metabolic classes that had significantly higher *C* values than background when subsets of scRNA-seq data were used.

To test this, we first split the scRNA-seq data into three subsets according to the genetic background and condition combinations (**Table S3**): WT under normal growth conditions (WT-normal); WT under stresses (WT-stress); non-WT under normal conditions (non-WT-normal). Bulk RNA-seq data were not split further due to the relatively small sample size (34). When the three scRNA-seq subsets were used, one or two additional metabolic classes had *z*-scores >1.645, but the total number of classes above the threshold remained unchanged or became less compared with when the entire scRNA-seq data were used (**Fig. 2B**, **Fig. S3**). We further split WT-normal scRNA-seq data into three tissue-specific subsets: root, shoot and leaf (the flower-specific subset was not considered due to the small number of cell types [n=6]). Five, six, and four additional metabolic classes had *z*-scores >1.645 when root, shoot and leaf-specific scRNA-seq data were used, respectively **(Fig. 2C, Fig. S4**). Particularly, DTX, SM, and fatty acids and lipids (FAL) classes had *z*-scores >1.645 when each of these three subsets was used. In contrast, *z*-scores of NT, inorganic nutrients (IN) and IM classes decreased when subsets were used compared with those using entire scRNA-seq data. These results are consistent with the relatively broad distribution of PNPs from NT, IN and IM classes, and tissue-specific distribution of PNPs from classes such as DTX, SM and hormones, and highlight the need to explore different RNA-seq data when examining co-expression tightness among MPGs for different metabolic classes.

Higher prediction accuracy for MPGs using scRNA-seq data than using bulk RNA-seq data

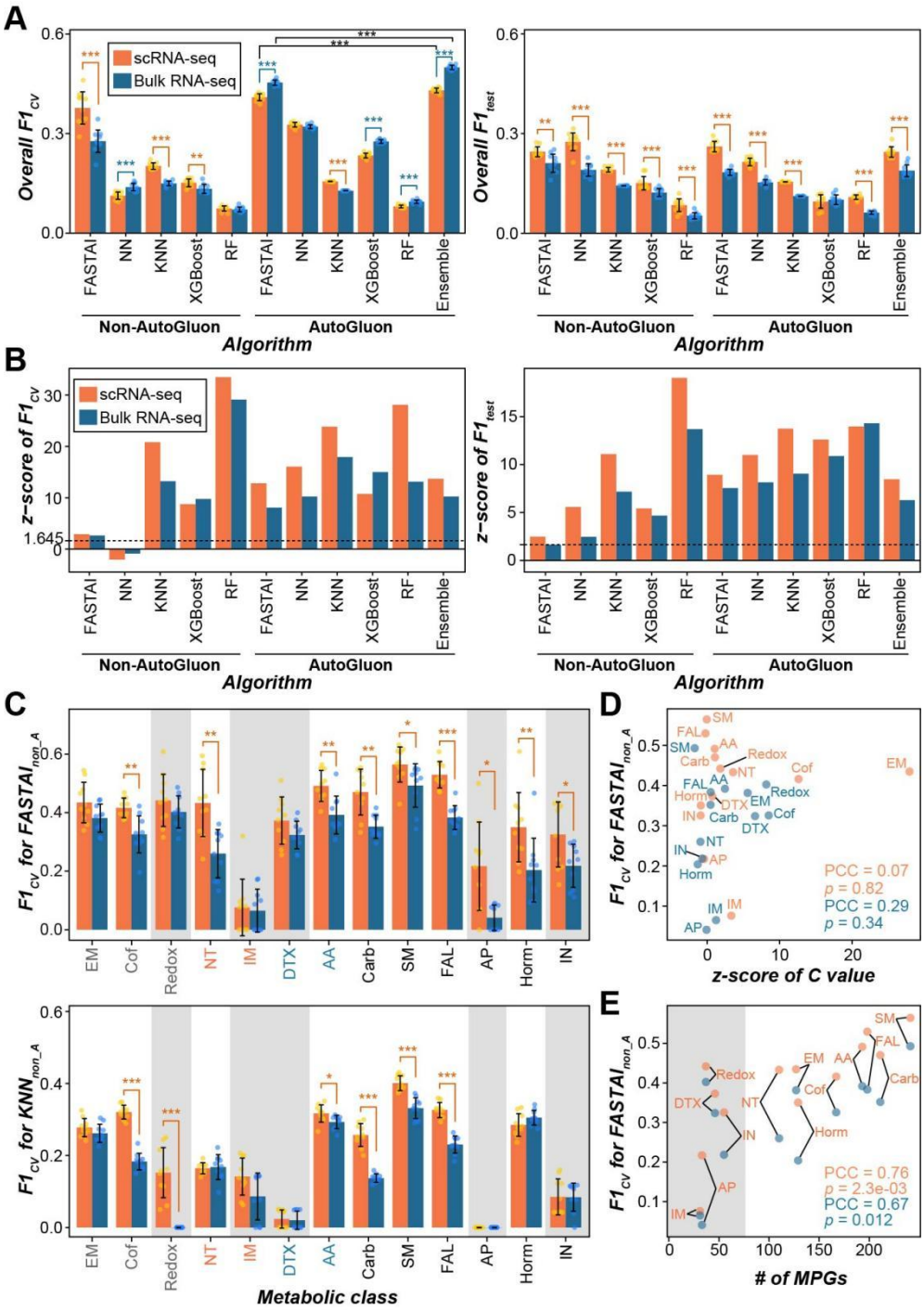
Thus far, we showed that benchmark MPGs within metabolic classes had different co-expression properties when scRNA-seq and bulk RNA-seq data were used. Next, we asked to what degree benchmark MPGs can be correctly predicted into metabolic classes using these two types of RNA-seq data. Considering that some MPGs belong to multiple metabolic classes (**Table S5**), we built ML models using five algorithms that support multi-class and multi-label classification tasks, including three classical ML algorithms: K-nearest neighbors[29] (KNN), eXtreme Gradient Boosting[30] (XGBoost), Random Forest[31] (RF), and two deep learning algorithms: FASTAI[32] and neural network[33] (NN) implemented in Pytorch[34]. See **Methods** for brief introduction of these five algorithms. In addition, we built prediction models using their corresponding algorithm in AutoGluon[35] (hereafter referred to as $algorithm_A$ to distinguish them

from those outside AutoGluon which will be referred to as $algorithm_{non_A}$) and an ensemble model comprising all these five algorithms ($ensemble_A$) as well. AutoGluon is a Python library that automates ML workflows including steps such as data preprocessing and hyperparameter tuning, and has been applied in multiple fields with promising performances, including plant SM gene prediction[36]. Twenty percent of benchmark MPGs within each class were first held out as the test set, and the remaining 80% MPGs were used to train the models using a five-fold cross-validation scheme (see **Methods**). F1 scores on the cross-validation ($F1_{CV}$) and test ($F1_{test}$) sets for models with the optimal combination of hyperparameters were both reported, where the F1 score was the harmonic mean of precision and recall of the prediction (see **Methods**).

When examining overall $F1_{CV}$ scores for MPGs across all the 14 classes (13 metabolic classes plus the "Others" class), we found that scRNA-seq-based models outperformed bulk RNA-seq-based models when algorithms $FASTAI_{non_A}$, KNN_{non_A}, $XGBoost_{non_A}$ and KNN_A were used, but were inferior to bulk RNA-seq-based models when NN_{non_A}, $FASTAI_A$, $XGBoost_A$, RF_A and $ensemble_A$ were used (left panel in **Fig. 3A**). In contrast, when overall $F1_{test}$ scores were examined, scRNA-seq-based models significantly outperformed bulk RNA-seq-based models for all but $XGBoost_{non_A}$ algorithms (right panel in **Fig. 3A**). $FASTAI_A$, NN_A and $XGBoost_A$ models had much higher overall $F1_{CV}$ scores than their corresponding $algorithm_{non_A}$ models, however, they tended to have severe overfitting issues: overall $F1_{test}$ scores were much smaller than their corresponding $F1_{CV}$ scores (**Fig. 3A**). This is potentially due to the relatively few genes within each metabolic class, whereas a typical AutoGluon task has hundreds and thousands of instances to learn from. The $ensemble_A$ models utilizing all the five $algorithms_A$ had improved $F1_{CV}$ but no improvement in $F1_{test}$ and with the overfitting issue unsolved. Among $algorithms_{non_A}$, $FASTAI_{non_A}$ (with a less severe overfitting issue) models performed best in terms of both $F1_{CV}$ and $F1_{test}$ (only slightly lower than that of NN_{non_A} model when scRNA-seq data were used), followed by NN_{non_A} (with an underfitting issue: $F1_{CV} < F1_{test}$), KNN_{non_A}, $XGBoost_{non_A}$ and RF_{non_A} models successively; the latter three classical algorithms tended to have no obvious overfitting issues.

To assess how well these models performed compared with random guesses, we established 1,000 additional models (100 for $ensemble_A$ models) using each of these $algorithms_{non_A}$ and $algorithms_A$, where the metabolic classes of all the benchmark MPGs were randomly shuffled, and calculated the *z*-scores of observed F1 scores in the distribution of these values from 1,000

(or 100) random models. Models built using true data generally had *z*-scores above 1.645, except for $F1_{CV}$ scores of NN_{non_A} models (with negative *z*-scores both when two types of RNA-seq data were used) and $F1_{test}$ score of $FASTAI_{non_A}$ model (1.640) when bulk RNA-seq data were used (**Fig. 3B, Fig. S5, Table S10**). This indicates significantly better prediction performance of these models than random guesses and the usefulness of both RNA-seq data in MPG prediction.



253 **Fig. 3. Prediction accuracy of MPGs for models built using scRNA-seq and bulk RNA-seq data**
254 **and different algorithms**. (**A**) Overall F1 scores on the five-fold cross-validation ($F1_{CV}$) and those on the
255 test set ($F1_{test}$) for all 13 metabolic classes and the "Others" class when scRNA-seq (orange) and bulk
256 RNA-seq (blue) data were used (n=10 replicate runs). (**B**) *Z*-scores of $F1_{CV}$ and $F1_{test}$ compared with
257 1,000 random guesses (100 for $ensemble_A$ models) where metabolic classes of all MPGs were shuffled
258 randomly. (**C**) $F1_{CV}$ scores for individual metabolic classes for $FASTAI_{non_A}$ (upper) and KNN_{non_A} (lower)
259 models. Light gray background: metabolic classes with < 55 MPGs. (D) Correlation between *z*-scores of
260 *C* value and the $F1_{CV}$ for metabolic classes in $FASTAI_{non_A}$ models, when scRNA-seq (orange) and bulk
261 RNA-seq (blue) data were used. PCC: Pearson correlation coefficient; *p*: *p*-value corresponding to the
262 PCC. (**E**) Correlation between the numbers of MPGs and the $F1_{CV}$ for metabolic classes in $FASTAI_{non_A}$
263 models, when scRNA-seq (orange) and bulk RNA-seq (blue) data were used. Error bar: standard
264 deviation of results from 10 replicate runs. Asterisk(s) indicate(s) significant levels for the two-sided
265 Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01; ***, *p*-value < 0.001.

266 Considering that MPGs within different classes behaved differently in the co-expression
267 analysis as shown in the above sections, we also calculated F1 scores for MPGs within each class.
268 Consistent with results of overall F1 scores, scRNA-seq-based models achieved higher $F1_{CV}$ and
269 $F1_{test}$ scores for most metabolic classes than the corresponding bulk RNA-seq-based models
270 when $algorithms_{non_A}$ were used (except for NN_{non_A} models); $FASTAI_A$, NN_A, $XGBoost_A$ and
271 $ensemble_A$ models had higher $F1_{CV}$ than their $algorithm_{non_A}$ counterparts, but with overfitting
272 issues; $FASTAI_{non_A}$ models outperformed other $algorithm_{non_A}$ models but with slight overfitting
273 issues; KNN_{non_A} models performed best among three classical algorithms (**Fig. 3C**, **Fig. S6**).
274 Thus, in the following sections, we will mainly focus on results of $FASTAI_{non_A}$ and KNN_{non_A}
275 models, unless otherwise specified. When examining the *z*-scores of F1 scores, which were
276 supposed to better reflect the prediction performance of models built with different factors than
277 F1 scores *per se*, we found that most of these 13 classes had *z*-scores >1.645 for both $F1_{CV}$ and
278 $F1_{test}$ regardless of RNA-seq data and algorithms used (**Fig. S7–15, Table S11–18**), further
279 suggesting the usefulness of both RNA-seq data in MPG prediction. We found no correlation
280 between the tightness of co-expression networks within classes and the prediction performance
281 (**Fig 3D**). Instead, there was positive correlation between prediction performance with the
282 number of MPGs within each class both when scRNA-seq and bulk RNA-seq data were used
283 (**Fig. 3E**). For example, MPGs in SM were generally predicted with the highest accuracy than
284 those in other classes, consistent with the largest number of MPGs in SM. Taken together, these

results demonstrate the advantage of scRNA-seq in predicting MPGs over bulk RNA-seq data and suggest that with the increase of knowledge about plant MPGs in the future, the prediction accuracy for these classes could be further improved.

Improvement in prediction accuracy by splitting the RNA-seq data

In the above section, we showed that splitting scRNA-seq data into subsets or tissue-specific data led to higher *C* values for certain metabolic classes (**Fig. 2B,C**, **Fig. S2–4**) than the entire scRNA-seq data. We next asked whether splitting the RNA-seq data could also lead to better prediction for MPGs. For comparison, bulk RNA-seq data were split into WT-normal, WT-stress and non-WT-normal subsets as well, but the WT-normal bulk RNA-seq data were not further split into tissue-specific subsets due to the small data volume. We found no improvement in $F1_{CV}$ and $F1_{test}$ were found for any classes in $FASTAI_{non_A}$ models (**Fig. 4A**, **Fig. S16A**). In KNN_{non_A} models, there was significant improvement in $F1_{CV}$ for redox and DTX when tissue-specific scRNA-seq data were used and $F1_{CV}$ for cofactors, redox, carbohydrates and amines and polyamines (AP) when subsets of bulk RNA-seq data were used; the $F1_{test}$ was significantly improved for more classes especially when subsets or tissue-specific scRNA-seq data were used (**Fig. 4B**, **Fig. S16C**). For other $algorithm_{non_A}$ models, splitting these two types of RNA-seq data led to improved prediction performance for some classes as well (**Fig. S16B**, **Fig. S17**). When comparing the $FASTAI_{non_A}$ models built using subsets of two types of RNA-seq data, scRNA-seq-based models still outperformed bulk RNA-seq-based models for most classes (**Fig. 4C**, **Fig. S18A**). For KNN_{non_A} models, scRNA-seq-based models had higher $F1_{CV}$ and $F1_{test}$ for most classes when subsets were used than corresponding bulk RNA-seq-based models, with a few exceptions (**Fig. 4D**, **Fig. S18C**). These results highlight the need to explore different scRNA-seq data when predicting MPGs into different metabolic classes, and further suggest the advantages of scRNA-seq data over bulk RNA-seq data in MPG prediction.

Furthermore, considering that scRNA-seq and bulk RNA-seq data led to different co-expression networks mentioned above, we investigated whether combining these two data would improve the prediction accuracy for MPGs. However, no significant improvement in prediction accuracy was found when combining these two data compared with single data-based models, no matter which algorithms were used (**Fig. S19**).

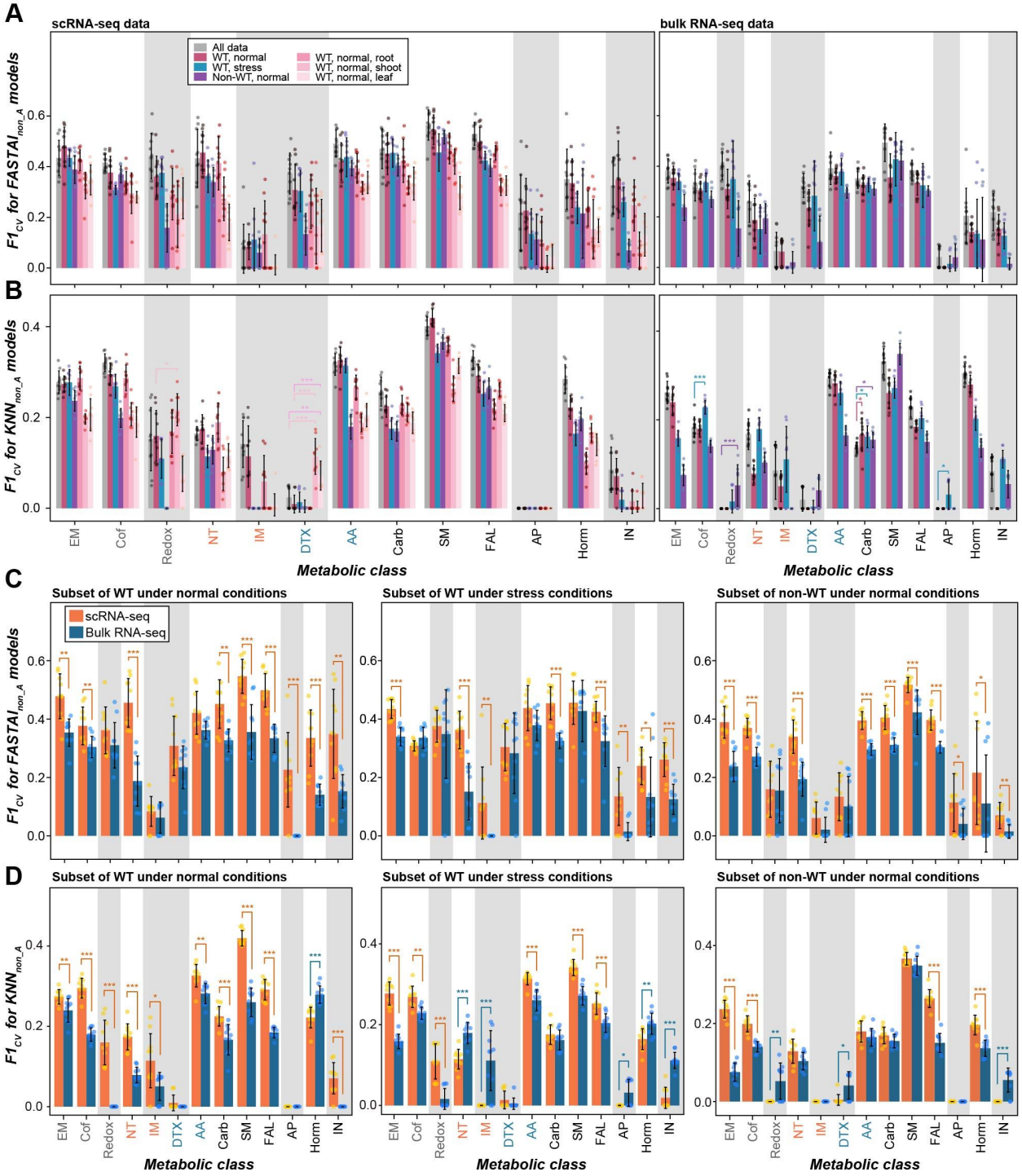


Fig. 4. Prediction accuracy of MPGs for $FASTAI_{non_A}$ and KNN_{non_A} models built using the entire and subsets of RNA-seq data. (**A,B**) $F1_{CV}$ scores for individual metabolic classes for $FASTAI_{non_A}$ (**A**) and KNN_{non_A} (**B**) models built using the entire sets and three subsets of scRNA-seq data, as well as three tissue-specific subsets of WT-normal scRNA-seq data (left panel), and the corresponding entire and three subsets of bulk RNA-seq data (right). Grey: models built using the entire

321 scRNA-seq or bulk RNA-seq data; rose, WT-normal subset; blue, WT-stress; purple, non-WT-normal;
322 pink, root-specific WT-normal scRNA-seq subsets; bright pink, shoot-specific WT-normal scRNA-seq
323 subsets; light pink, leaf-specific WT-normal scRNA-seq subsets. Asterisk(s) indicate(s) significant levels
324 (only showed when the subset-based models significantly outperformed corresponding models built using
325 the entire RNA-seq data) for the two-sided Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01;
326 ***, *p*-value < 0.001. (**C,D**) Comparison of $F1_{CV}$ between $FASTAI_{non_A}$ (**C**) and KNN_{non_A} (**D**) models
327 built using subsets of scRNA-seq (orange) and bulk RNA-seq data (blue). Light gray background:
328 metabolic classes with < 55 MPGs. Error bar: standard deviation of results from 10 replicate runs.

329 Prediction of metabolic classes for enzyme genes without experimental 330 evidence

331 To evaluate how our models performed for new genes, we applied our models to 9,441
332 enzyme genes that have not been assigned to any metabolic pathways supported by experimental
333 evidence in the PMN database (hereafter referred to as unknown genes). Only $FASTAI_{non_A}$ and
334 KNN_{non_A} models were used due to their relatively better performance compared with other
335 models. Using $FASTAI_{non_A}$ models, 9,296 (98.46%) and 9,202 (97.47%) unknown genes were
336 predicted into at least one of the 13 metabolic classes when scRNA-seq and bulk RNA-seq data
337 were used, respectively (**Fig. 5A, Table S19**), whereas 6,164 (65.3%) and 5,588 (59.2%)
338 unknown genes were predicted by scRNA-seq-based and bulk RNA-seq-based KNN_{non_A} models,
339 respectively (**Fig. S20A, Table S19**). The number of unknown genes that were predicted into
340 each metabolic class was positively correlated with the number of benchmark MPGs within the
341 class (**Fig. 5A, Fig. S20A**). This result suggests two potential scenarios that are not mutually
342 exclusive: 1) the current pre-knowledge about MPGs are relatively proportionally distributed
343 among metabolic classes, namely, the number of benchmarks genes in each class is correlated
344 with the number of real enzyme genes in Arabidopsis; 2) the data structure of predicted instances
345 are influenced by the data structure of the training instances.

346 Next, we examined whether the predicted metabolic classes for unknown genes are
347 consistent between models built using two types of RNA-seq data and with the annotation from
348 the PMN database (**Table S5, S19**). Using $FASTAI_{non_A}$ models, 266 unknown genes were
349 predicted into exactly the same metabolic classes using scRNA-seq and bulk RNA-seq data, and
350 11 out of them were annotated in the same classes by PMN (strict criteria, upper panel in **Fig.
351 5B**); the numbers were 5,778 and 548 for a loose criteria where a gene had at least one common

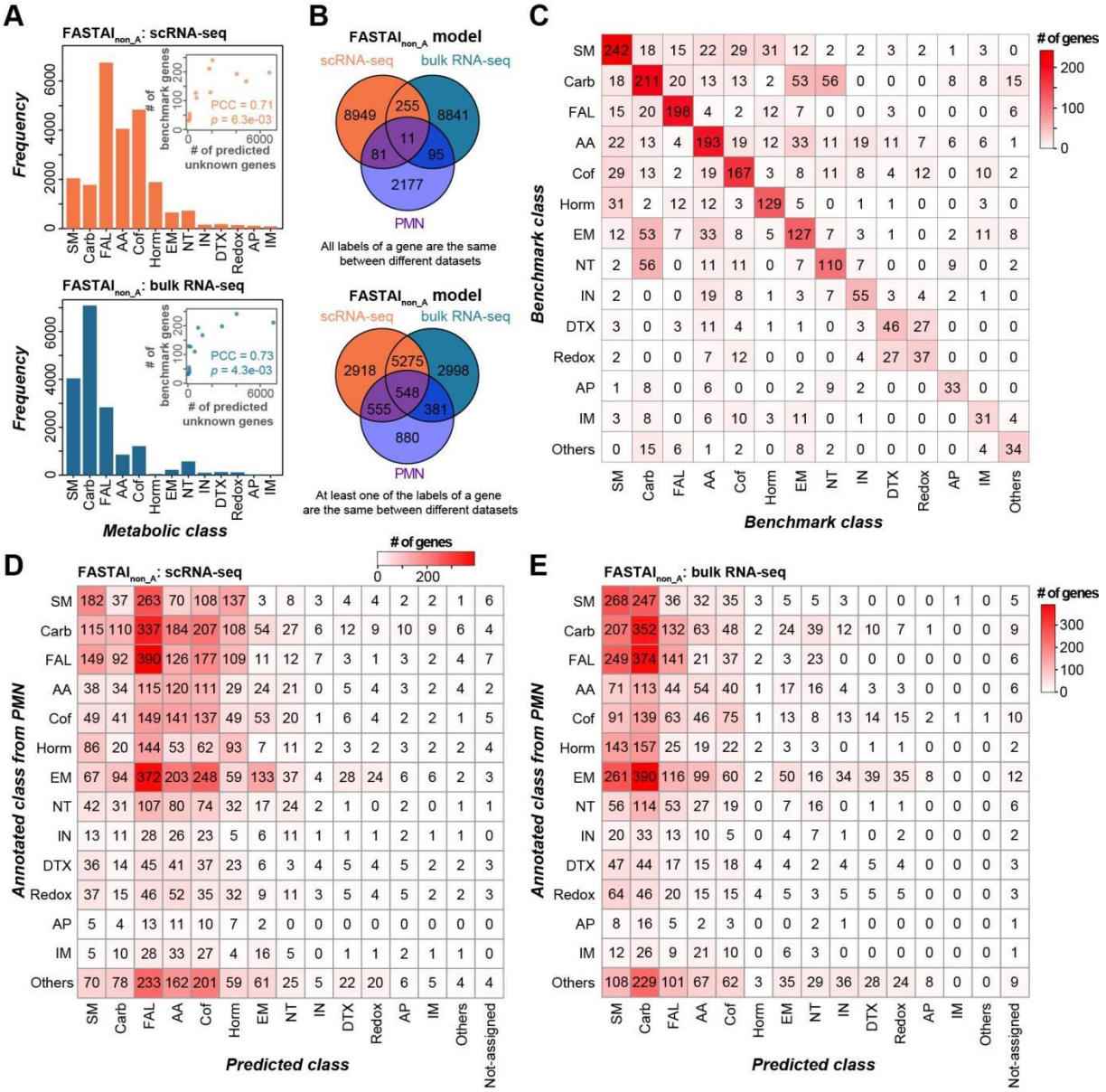


Fig. 5. Prediction of metabolic classes for unknown genes by FASTAI$_{non_A}$ models. (**A**) Count of unknown genes that were predicted into 13 metabolic classes by FASTAI$_{non_A}$ models built using scRNA-seq (upper panel) and bulk RNA-seq (lower) data. Inset scatter plot shows the correlation between the number of benchmark MPGs and the number of predicted unknown genes within classes when using scRNA-seq or bulk RNA-seq data. (**B**) Venn diagrams showing the overlap of unknown genes that were predicted into metabolic classes by FASTAI$_{non_A}$ models built using scRNA-seq (orange) and bulk RNA-seq (blue) data, and those that were annotated as within metabolic classes in the PMN database (purple). In the upper and lower panels, overlapping regions show the number of genes with exactly the same classes or at least one common class between datasets, respectively. (**C**) A heatmap showing the numbers of benchmark MPGs that were annotated within each class (diagonal cells) or several classes (non-

363 diagonal cells). For example, from 242 benchmark MPGs that were annotated as in SM, 31 MPGs were
364 also annotated as in the hormones class. (**D**,**E**) Heatmaps showing the numbers of unknown genes that
365 were predicted as in a class (x-axis) by $FASTAI_{non_A}$ models built using scRNA-seq (**D**) and bulk RNA-
366 seq (**E**) data, and were also annotated as within the same or other classes (y-axis) by PMN database.
367 Values in the column "Not-assigned" indicate the numbers of unknown genes that were predicted into
368 neither any of the 13 metabolic classes nor the "Others" category. Cell colors in (**C,D,E**) indicate the
369 numbers of genes.

370 class between different datasets (lower panel in **Fig. 5B**). In comparison, using KNN_{non_A} models,
371 1,150/47 and 1,596/140 unknown genes shared common class memberships between two
372 predictions/among three datasets, when using strict or loose criteria, respectively (**Fig. S20B**).
373 Considering that a benchmark MPG can be annotated within more than one classes, and the
374 multiple class annotations for a gene were not randomly distributed (e.g., MPGs in DTX tended
375 to be also annotated as in redox, compared with other classes; some MPGs in SM also participate
376 in the biosynthesis of cofactors and hormones, **Fig. 5C**), we further investigated whether the
377 mismatches of classes between PMN annotations and model predictions have a similar bias
378 among classes (**Fig. 5D,E, Fig. S20C,D**). We found moderate but significant correlations
379 between the mismatch matrix and the multiple label matrix of benchmark MPGs: PCC = 0.393
380 (p=1.2e-08), 0.425 (5.1e-10), 0.429 (3.4e-10) and 0.432 (2.7e-10) for $FASTAI_{non_A/sc}$,
381 $FASTAI_{non_A/bulk}$, $KNN_{non_A/sc}$, and $KNN_{non_A/bulk}$, respectively. This result suggests that the 'mis-
382 prediction' of unknown MPGs of RNA-seq data-based models compared with PMN annotations
383 are potentially confounded by the multiple functions of these enzyme genes, at least to some
384 extent. The discrepancies between predictions and PMN annotations can be also explained by the
385 unbalanced numbers of benchmark MPGs from different classes used to train the models, and the
386 far-from-perfect prediction accuracy of our models built using transcriptomic data. The potential
387 low quality of PMN annotations for these unknown genes can be another rationale that should
388 not be excluded at the current stage.

389 Notably, *AT2G05990* and *AT3G45770* had PMN annotations as FAL genes (**Table S5**),
390 encoding enoyl-acyl carrier protein (enoyl-ACP) reductases that catalyze the reduction of enoyl-
391 CoA substrates in the fatty acid synthase system[37]. *AT2G05990* was correctly predicted by
392 $FASTAI_{non_A}$ models regardless of the RNA-seq data used, and by KNN_{non_A} model when using
393 scRNA-seq data, while *AT3G45770* was predicted to be within FAL by $FASTAI_{non_A}$ model

394 when using scRNA-seq data (**Table S19**). *AT4G22870*, a tandem duplicate of
395 *ANTHOCYANIDIN SYNTHASE* (*ANS*, *AT4G22880*, involved in pro-anthocyanin biosynthesis),
396 was annotated as a SM gene by PMN without experimental evidence (**Table S5**). It was
397 predicted to be within SM and other related classes by $FASTAI_{non_A/sc}$, $FASTAI_{non_A/bulk}$ and
398 $KNN_{non_A/sc}$ models, but not predicted to be within any metabolic classes by $KNN_{non_A/bulk}$ model
399 (**Table S19**). These results further demonstrate the utilities of our models in plant MPG
400 prediction, and the superiority of scRNA-seq data over bulk RNA-seq data, at least for some
401 metabolic classes.

402 # Discussion

403 Models built using two types of RNA-seq data had significantly higher prediction accuracy
404 for most metabolic classes than random guesses, suggesting the usefulness of both RNA-seq data
405 in predicting MPGs into metabolic classes. Generally, scRNA-seq-based models led to higher
406 performance in predicting MPGs into classes than bulk RNA-seq-based models, consistent with
407 our assumption that scRNA-seq data may provide more detailed expression profiles for genes
408 than bulk RNA-seq data. However, there are still several potential limits for scRNA data. For
409 example, the sequencing and analytical techniques for single-cell transcriptome at present still
410 have multiple limitations, such as the low capture efficiency for transcripts within cells,
411 especially for genes with low expression, resulting in inherently sparse and noisy nature of the
412 scRNA-seq data[38]; the analysis of scRNA-seq data is complex and challenging, and requires
413 prior knowledge about marker genes[39]. With the development and improvement of sequencing
414 and analytical techniques, scRNA-seq data is expected to show more promise in multiple fields,
415 including predicting plant MPGs. In addition, the biosynthesis of a PNP may occur sequentially
416 in a series of specific cells. For example, the transcription of 20 genes responsible for the
417 biosynthesis of monoterpenoid indole alkaloids occur in three types of cells (namely, internal
418 phloem-associated parenchyma cells, epidermal cells, and laticifer cells) in the *Catharanthus*
419 *roseus* leaves, and the intermediates produced during the biosynthesis process are transported
420 between different cell types[23]. If this phenomenon is common for PNPs, bulk RNA-seq data that
421 capture gene expression profiles for the entire tissues rather than individual cell types are
422 supposed to be superior to scRNA-seq data in predicting MPGs. Bioinformatic approaches that
423 take into account the temporal or spatial frameshift in expression when measuring the expression

correlation between genes could potentially make better use of the scRNA-seq data in predicting MPGs.

Our predictive models used the expression profiles of genes in scRNA-seq and bulk RNA-seq data, rather than the expression correlations among genes (e.g., *C* values of genes). In one of our previous studies[22], we showed that gene-to-pathway expression similarity led to better prediction for pathway genes than gene expression levels, thus models built using *C* values of genes may improve the prediction performance. However, a major challenge for using *C* values of genes as predictors in the models is that the *C* value of a gene is calculated by considering all the neighbors that are directly connected with the gene. It is a tricky problem to calculate the *C* values for genes in the validation and test subsets with neighbors from the training subset while controlling the risks of data leakage during the calculation.

In our study, we combined MPGs from different metabolic pathways that belong to the same metabolic class together due to the limited number of MPGs in individual pathways. However, MPGs in different metabolic pathways of a single class may have totally different expression profiles, especially those for specialized metabolites, which may be synthesized in specific tissues or cell types, or under specific conditions, such as biotic and abiotic stresses. For example, acyl sugars are specifically synthesized in the glandular trichomes of certain plants[40]; the rupture of a cassava leaf mesophyll cell caused by herbivore attack or animal chewing leads to the production of cyanide which has the bitter taste for mammals[41]. Co-expression analysis (e.g., *C*-value) for MPGs in those classes led to no better or even poorer results than randomized background (e.g, lower *z*-score of *C*-values for SM), whereas ML algorithms tended to have the capability to overcome this limitation by learning from expression profiles of known MPGs in the same metabolic class (e.g., higher prediction accuracy for SM than those for others). In addition, splitting the scRNA-seq data into tissue-specific subsets improved z-scores of *C*-values for most classes, but did not necessarily improve MPG predictions for those classes, further suggesting the disassociation between co-expression network tightness with prediction accuracy.

The finding that prediction performance for a given class was influenced by the number of MPGs within the class regardless of the types of RNA-seq data used (**Fig. 3E**) suggests an urgent need to expand our knowledge about plant MPGs to improve the prediction accuracy, which in turn needs assistance from other approaches (such as ML approaches in our study) beyond conventional forward genetics. One potential solution for this seemingly "endless loop" is to

455 experimentally evaluate part of the preliminary predictions (error-prone) and feed the validated
456 knowledge back to the ML models to boost the prediction performance. Another potential
457 solution would be the development of ML algorithms which can handle with giant amount of
458 features (e.g., expression and epigenetic profiles of genes across a large number of tissues, cell
459 types, stages, environments, and so on) while learning from a limited numbers of instances (i.e.,
460 genes). The AutoGluon algorithm has shown great performance in prediction tasks such as
461 OpenML AutoML and Kaggle Benchmark[35]. However, AutoGluon models had the risk of
462 overfitting in our study, potentially due to the scarcity of benchmark MPGs within each class to
463 learn from. This indicates that for some prediction tasks in biology with limited numbers of
464 instances (e.g., plant MPGs), the algorithms developed for big data (e.g., AutoGluon) that show
465 great promise in other fields (e.g., image recognition) should be used with caution.

466 In summary, the prediction of plant MPGs is a complex and challenging task, more efforts
467 should be put in the future to expand our knowledge about the contributing genes underlying
468 PNPs with higher efficiency and to accelerate the development of synthetic biology. Specifically,
469 scRNA-seq data should be considered in plant MPG identification.

470 Materials and Methods

471 RNA-seq data collection and preprocessing

472 A total of 21 scRNA-seq datasets of *A. thaliana* (**Table S1**), containing 259 cell types, were
473 downloaded from the scPlantDB database[42] (https://biobigdata.nju.edu.cn/scplantdb, accessed on
474 Nor 26th, 2024). These collected scRNA-seq datasets contained samples taken from root, shoot,
475 leaf, flower and callus tissues of WT plants grown under normal conditions, WT under biotic or
476 abiotic stress conditions, and non-WT under normal conditions (**Table S3**). For comparison, 19
477 bulk RNA-seq datasets that contained samples from the same types of tissues of plants grown
478 under same conditions (**Table S2,S4**), containing in total 34 samples, were downloaded from the
479 Expression Atlas database (https://www.ebi.ac.uk/gxa/sc/home) and GEO (accessed on Dec 19th,
480 2024) database, or from the publication[43].

481 The RDS files downloaded from the scPlantDB database, containing scRNA-seq data with
482 cell type annotations, were processed using the Seurat (v4.3.0) package[44] in R (v4.2.0). Only
483 protein-coding genes were included, with missing values in the resulting expression matrix filled
484 with zeros. Specifically, the "NormalizeData" was used to normalize gene expression levels

among different cells, where read count of each gene was divided by the total counts within individual cells, then multiplied by a scale factor of 10,000 and log1p-transformed; "AverageExpression" was used to average gene expression levels across 259 cell types, where it first applied the expm1 transformation (exponentiation minus one) before averaging across cells of the same type (**Supplementary Data 1**).

For the preprocessing of bulk RNA-seq data, the gene expression levels (transcripts per million, TPM) were first averaged across biological replicates, and then were normalized using the "normalize.quantiles" function in R among different samples, where the expression values within individual samples were adjusted to ensure a consistent distribution and to eliminate technical variations between samples (**Supplementary Data 2**).

Metabolic pathway gene annotation collection

We first downloaded the metabolic pathway gene annotation from the PMN database[12] (https://plantcyc.org/, accessed on Sep 9th, 2022, PlantCyc v15.1.0, which contained 3,508 MPGs annotated in 635 metabolic pathways. To obtain benchmark MPGs, we wrote a python script to crawl the PMN webpages of individual MPGs (accessed on Dec 19th, 2022) and kept MPGs with experimental evidence, such as those inferred from direct assay, mutant phenotype, etc. This resulted in a total of 1,131 benchmark MPGs, which belong to 636 metabolic pathways (**Table S5**). Two genes (*AT3G25830* and *ATMG01360*) were removed from the following analysis (thus retaining 1,129 benchmark MPGs) due to the lack of expression information in the scRNA-seq data for them. In particular, *AT3G25830* belonged to oleoresin monoterpene volatiles biosynthesis pathway (PWY-5423) and monoterpene biosynthesis pathway (PWY-3041), while *ATMG01360* was part of aerobic respiration I (cytochrome c) pathway (PWY-3781). Additional potential enzyme genes of *A. thaliana* were obtained from The Arabidopsis Information Resource (TAIR) databases (https://www.arabidopsis.org/), or annotated using the Ensemble Enzyme Prediction Pipeline (E2P2, v3.1) software. In summary, a total of 10,572 potential enzyme-encoding genes were annotated, including 1,131 benchmark MPGs (**Table S5**).

Weighted gene co-expression network analyses

Gene co-expression network analysis was conducted using the WGCNA (v1.72.1) package[45] in R. During the network analysis, the matrix of Pearson correlation coefficient (PCC) values for gene pairs was transformed into an adjacency matrix to achieve a scale-free network structure,

515 with the parameter 'networkType' being set to 'signed', where the PCC values were processed
516 using the equation:

517 $$a_{i,j}=\left(\frac{1+cor(i,j)}{2}\right)^{\beta} \quad (1)$$

518 where, $a_{i,j}$ represents the weight of the edge between two genes i and j; $cor(i, j)$ represents the
519 PCC between the two genes; β is the best soft threshold power estimated by the software
520 WGCNA using the function "pickSoftThreshold", which was estimated at 12 for scRNA-seq and
521 14 for bulk RNA-seq data. The "cutreeDynamic" function was employed to divide the network
522 into 15 co-expression modules for scRNA-seq data and 12 modules for bulk RNA-seq (**Table**
523 **S6,7**), respectively, using parameters as minModuleSize=30, mergeCutHeight=0.1 and
524 deepSplit=4 for both data.

525 ## Clustering coefficient calculation

526 To measure the tightness of gene co-expression within metabolic classes, we calculated the
527 clustering coefficient (C value) for each class. The gene expression values in two RNA-seq data
528 first plus one and were log-transformed with base 10, and then processed through the WGCNA
529 pipeline to obtain the gene co-expression networks. Specifically, for the leaf subset of scRNA-
530 seq WT-normal data when the pipeline failed to determine the optimal soft threshold, we
531 manually specified the soft threshold to 12 that corresponded to the point where the scale-free fit
532 index (signed R^2) first reached a peak, to ensure the process continues. In the networks, the
533 nodes were benchmark MPGs, edges between nodes were co-expression relationships between
534 MPGs with the thickness of edges (i.e., the strength of the connection) being the transformed
535 PCC values. The networks were next processed using the NetworkX (v3.1) package[46] in Python
536 to calculate the local C values for individual genes, with the equation[26]:

537 $$c_u=\frac{1}{deg(u)(deg(u)-1)}\sum_{vw}(\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3} \quad (2)$$

538 where, for a node (i.e., a gene) u with two adjacent nodes v and w as an example, c_u represents
539 the local C value of the node u; $deg(u)$ is the degree of the node u, representing the number of
540 neighboring nodes connected to the node u; w_{uv}, w_{uw}, and w_{vw} represent the weights of the edges
541 between nodes u, v, and w; the weights are normalized by being divided by the maximum weight
542 in the network, e.g., $\hat{w}_{uv}=w_{uv}/max(w)$; $(\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3}$ indicates the geometric mean of the
543 normalized weights between nodes u, v, and w; $\sum_{vw}$ represents the summation over all pairs of

544 neighboring nodes (i.e., v,w) that are directly connected to node u. The local C value ranges from
545 0 to 1. The average local C value of all genes within a metabolic class (the global C value,
546 hereafter referred to as C value for short) was used to represent the compactness of the class in
547 terms of gene expression (**Table S8**).

548 Since the C values can be influenced by the number of nodes within a network[45], to
549 understand what C values would indicate a tight network and to make the C values among
550 different metabolic classes comparable, we also estimated the C values for 10,000 random
551 simulations (**Table S9**) where the class memberships of all the benchmark MPGs were randomly
552 shuffled to simulate the out-of-order relationships among these MPGs. Then a z-score of the
553 observed C values in the distribution of 10,000 background C values were calculated using the
554 equation:

555 $$z=\frac{x-\mu}{\sigma} \qquad (3)$$

556 where, z represents the z-score value, x represents the observed C value, μ represents the mean of
557 10,000 background C values, and σ represents the standard deviation of 10,000 background C
558 values.

559 Machine learning model building

560 Five algorithms were used to build ML models: three traditional algorithms,
561 KNeighborsClassifier (K-Nearest Neighbors, KNN), XGBClassifier (eXtreme Gradient Boosting,
562 XGBoost), and RandomForestClassifier (RF), were called from scikit-learn[47] (v1.2.2,
563 implemented in Python v3.9.16); two deep learning algorithms, FASTAI and neural network
564 (NN) implemented in Pytorch. Specifically, KNN makes predictions of classes for a given gene
565 based on proximity of this gene with genes in the training set; XGBoost is an optimized
566 distributed gradient boosting library that provides parallel tree boosting; RF builds a number of
567 classification decision trees using a number of sub-samples of the dataset and then aggregates the
568 trees into a single ensemble model; FASTAI leverages deep learning architectures by providing
569 automated training strategies and data preprocessing pipelines; NN classifies genes through
570 multiple layers of neurons that transform inputs into output probabilities. In addition, their
571 corresponding algorithms implemented in the AutoGluon library (v0.8.2, Python) were used to
572 build models as well, since AutoGluon automates machine learning pipelines including the
573 hyperparameter tuning step and has been applied in multiple fields with promising performance.

Out of the benchmark MPGs within each class, 20% were used as the test set to evaluate the performance of the models. The remaining 80% of benchmark MPGs within each class were further split into five folds; MPGs in four folds were used to train the model and MPGs in the remaining fifth fold were used to validate the model performance; this training-validation step was conducted five times to make sure that MPGs in all five folds were used as in the validation subset once. Caution should be taken to avoid that genes that belong to multiple classes are held out as in the test set for one class but remain as in the training set for another class, and the same for the training/validation splits.

Hyperparameter tuning was conducted for KNN, XGBoost and RF models (this step is automatically embedded within AutoGluon) to select the optimal combination of parameters using GridSearch. Specifically, the hyperparameter spaces for KNN models were {3, 4, 5, 6, 7} for the 'n_neighbors' (*k* values) and {'uniform', 'distance'} for the 'weights'; for XGBoost modes: {3, 5, 10} for 'max_depth', {0.01, 0.1, 0.5, 1} for 'learning_rate', {10, 50, 100, 500} for 'n_estimators' and {0.1, 5} for 'gamma'; for RF models: {3, 5, 10} for the 'max_depth', {0.1, 0.25, 0.5, 0.75, 'sqrt', 'log2', None} for the 'max_features' and {10, 50, 100, 500} for the 'n_estimators'. The hyperparameter spaces for both FASTAI and NN models were {[200, 100], [100, 50], [512, 256]} for the 'layer_options', {1e-3, 1e-2} for the 'lr_options' (learning rate), {5, 10} for the 'epoch_options', and {32, 64} for the 'bs_options' (batch size). The best combination of parameters for each model was determined based on the F1 scores on the cross-validation set ($F1_{CV}$), where F1 was calculated with the equations:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

$$precision = \frac{TP}{TP + FP} \quad (5)$$

$$recall = \frac{TP}{TP + FN} \quad (6)$$

where TP (true positive) represents the number of benchmark MPGs that are correctly predicted into a class; FP (false positive) represents the number of benchmark MPGs that are mis-predicted into a class; FN (false negative) represents the number of benchmark MPGs that are not predicted into the correct class. For the repeatability of our results, we built 10 replicate models with different cross-validation splits using their specific best hyperparameters, and these 10 replicate models were used to make predictions for MPGs in the test set, while the last one out of these 10 models was used to make predictions for the unknown genes.

Acknowledgments

We thank Xinhao Fan, Qiaowei Li, Kangli Li, Ruipu Chen, Zhuobiao Wang, for their helpful suggestions in preprocessing the scRNA-seq data. This work was supported by the National Natural Science Foundation of China (32370241 to PW) and Scientific Research Foundation for Principle Investigator, Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003 to PW).

Author contributions

PW conceived and designed this study. ZW collected the metabolic pathway gene annotations; JM collected and preprocessed the bulk RNA-seq and scRNA-seq data with help from LZ. ZW and JM wrote the machine learning pipelines. JM conducted all the other analyses with help from LZ, XW and ZW. JM and PW wrote the manuscript with inputs from all the other authors. All authors read and approved the final manuscript.

Data availability

All the data used in this study are provided in the supplementary materials, and all the scripts used are available on Github at: https://github.com/peipeiwang6/Manuscript/tree/main/2024_scRNA_in_pathway_prediction.

Supplementary figure legends

Fig. S1. Shared benchmark MPGs within co-expression modules inferred by scRNA-seq and bulk RNA-seq data. (**A**) A venn diagram showing the numbers of benchmark MPGs shared by $M2_{sc}$, $M6_{sc}$, $M12_{sc}$, $M15_{sc}$, and $M12_{bulk}$, for which MPGs from the specialized metabolism class were overrepresented. (**B**) A venn diagram showing the numbers of benchmark MPGs shared by $M4_{sc}$, $M5_{sc}$, $M12_{sc}$, $M1_{bulk}$, and $M5_{bulk}$, for which MPGs from the hormone class were overrepresented. Orange: scRNA-seq-based modules; blue: bulk RNA-seq-based modules.

Fig. S2. Clustering coefficients (*C* values) for MPGs within each metabolic class calculated using scRNA-seq (A) and bulk RNA-seq (B) data. The density plot shows the distribution of

630 background *C* values for MPGs within each metabolic class for 10,000 random simulations
631 where the metabolic class annotations for MPGs were randomly shuffled. The red solid line
632 denotes the observed *C* value for the co-expression network of genes within each class; the black
633 dashed line indicates the average *C* value of 10,000 random simulations. The *z* represents the *z*-
634 score of the observed *C* value in the background distribution; *p* represents the corresponding *p*-
635 value of the *z*-score.
636
637 **Fig. S3. *C* values for MPGs within each metabolic class calculated using subsets of scRNA-**
638 **seq data**. Subsets are wild type (WT) plants grown under normal conditions (**A**), WT under
639 stress conditions (**B**) and non-WT under normal conditions (**C**). The density plot shows the
640 distribution of background *C* values for MPGs within each metabolic class for 10,000 random
641 simulations where the metabolic class annotations for MPGs were randomly shuffled. The red
642 solid line denotes the observed *C* value for the co-expression network of genes within each class;
643 the black dashed line indicates the average *C* value of 10,000 random simulations. The *z*
644 represents the *z*-score of the observed *C* value in the background distribution; *p* represents the
645 corresponding *p*-value of the *z*-score.
646
647 **Fig. S4. *C* values for MPGs within each metabolic class calculated using tissue-specific**
648 **subsets of WT-normal scRNA-seq data**. Subsets are WT-normal-root (**A**), WT-normal-shoot
649 (**B**) and WT-normal-leaf (**C**). The density plot shows the distribution of background *C* values for
650 MPGs within each metabolic class for 10,000 random simulations where the metabolic class
651 annotations for MPGs were randomly shuffled. The red solid line denotes the observed *C* value;
652 the black dashed line indicates the average *C* value of 10,000 random simulations. The *z*
653 represents the *z*-score of the observed *C* value in the background distribution; *p* represents the
654 corresponding *p*-value of the *z*-score.
655
656 **Fig. S5. Overall F1 scores for predictive models when scRNA-seq (A) and bulk RNA-seq (B)**
657 **data were used.** The density plot shows the distribution of overall F1 scores for models where
658 the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red
659 solid line: the average $F1_{CV}$ or $F1_{test}$ score for models built using the true labels of MPGs; pink
660 rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1

661 score of 1,000 random guesses for single algorithm models, and 100 for ensemble$_A$ models. NN:
662 neural network; KNN: K-Nearest Neighbors; XGBoost: eXtreme Gradient Boosting; RF:
663 Random Forest; algorithm$_{non_A}$: algorithm executed outside of AutoGluon; algorithm$_A$: algorithm
664 implemented in AutoGluon. *z*: *z*-score of the average F1 score of 10 replicate runs in the
665 background distribution; *p*: the corresponding *p*-value of the *z*-score.
666
667 **Fig. S6. Prediction accuracy of MPGs within individual metabolic classes for predictively**
668 **models**. (**A**) F1$_{test}$ scores of 13 metabolic classes for FASTAI$_{non_A}$ and KNN$_{non_A}$ models, and
669 F1$_{CV}$ and F1$_{test}$ scores for individual metabolic classes for NN$_{non_A}$, XGBoost$_{non_A}$ and RF$_{non_A}$
670 models when scRNA-seq (orange) and bulk RNA-seq (blue) data were used. (**B**) F1$_{CV}$ and F1$_{test}$
671 scores for individual metabolic classes for FASTAI$_A$, NN$_A$, KNN$_A$, XGB$_A$, RF$_A$ and ensemble$_A$
672 models when scRNA-seq (orange) and bulk RNA-seq (blue) data were used. Light gray
673 background: metabolic classes with < 55 MPGs. Error bars indicate the standard deviation of
674 results from 10 replicate runs. Asterisk(s) indicate(s) significant levels for the two-sided
675 Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01; ***, *p*-value < 0.001.
676
677 **Fig. S7. F1$_{CV}$ and F1$_{test}$ scores of individual classes for FASTAI$_{non_A}$ (A, B), NN$_{non_A}$ (C, D)**
678 **and KNN$_{non_A}$ (E, F) models built using scRNA-seq data**. The density plot shows the
679 distribution of F1 scores for 1,000 models where the labels of benchmark MPGs in the training
680 set were randomly shuffled (random guesses). Red solid line: the average F1 score for models
681 built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate
682 runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average
683 F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the
684 *z*-score.
685
686 **Fig. S8. F1$_{CV}$ and F1$_{test}$ scores of individual classes for XGBoost$_{non_A}$ (A, B), RF$_{non_A}$ (C, D)**
687 **and FASTAI$_A$ (E, F) models built using scRNA-seq data**. The density plot shows the
688 distribution of F1 scores for models where the labels of benchmark MPGs in the training set
689 were randomly shuffled (random guesses). Red solid line: the average F1 score for models built
690 using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs;
691 black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1

692 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-
693 score.
694
695 **Fig. S9. $F1_{CV}$ and $F1_{test}$ scores of individual classes for NN_A (A, B), KNN_A (C, D), and**
696 **$XGBoost_A$ (E, F) models built using scRNA-seq data**. The density plot shows the distribution
697 of F1 scores for models where the labels of benchmark MPGs in the training set were randomly
698 shuffled (random guesses). Red solid line: the average F1 score for models built using the true
699 labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed
700 line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1 score of 10
701 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.
702
703 **Fig. S10. $F1_{CV}$ and $F1_{test}$ scores of individual classes for RF_A (A, B) and $ensemble_A$ (C, D)**
704 **models built using scRNA-seq data**. The density plot shows the distribution of F1 scores for
705 models where the labels of benchmark MPGs in the training set were randomly shuffled (random
706 guesses). Red solid line: the average F1 score for models built using the true labels of MPGs;
707 pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1
708 score of 1,000 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the
709 background distribution; *p*: the corresponding *p*-value of the *z*-score.
710
711 **Fig. S11. $F1_{CV}$ and $F1_{test}$ scores of individual classes for $FASTAI_{non_A}$ (A, B), NN_{non_A} (C, D),**
712 **and KNN_{non_A} (E, F) models built using bulk RNA-seq data**. The density plot shows the
713 distribution of F1 scores for models where the labels of benchmark MPGs in the training set
714 were randomly shuffled (random guesses). Red solid line: the average F1 score for models built
715 using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs;
716 black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1
717 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-
718 score.
719
720 **Fig. S12. $F1_{CV}$ and $F1_{test}$ scores of individual classes for $XGBoost_{non_A}$ (A, B), RF_{non_A} (C, D)**
721 **and $FASTAI_A$ (E, F) models built using bulk RNA-seq data**. The density plot shows the
722 distribution of F1 scores for models where the labels of benchmark MPGs in the training set

were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: z-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S13. $F1_{CV}$ and $F1_{test}$ scores of individual classes for NN_A (A, B), KNN_A (C, D), and $XGBoost_A$ (E, F) models built using bulk RNA-seq data. The density plot shows the distribution of F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S14. $F1_{CV}$ and $F1_{test}$ scores of individual classes for RF_A (A, B) and $ensemble_A$ (C, D) models built using bulk RNA-seq data. The density plot shows the distribution of F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S15. *Z*-scores of F1 scores for 13 metabolic classes compared with values of random guess. (**A**) *z*-scores of $F1_{CV}$ (upper panel) and $F1_{test}$ (lower) scores for 13 metabolic classes compared with values of random guess for 1,000 times for $FASTAI_{non_A}$, NN_{non_A}, KNN_{non_A}, $XGBoost_{non_A}$, and RF_{non_A} models using scRNA-seq (orange) and bulk RNA-seq (blue) data. (**B**) *z*-scores of $F1_{CV}$ (upper panel) and $F1_{test}$ (lower) scores for 13 metabolic classes compared with values of random guess for 1,000 times for $FASTAI_A$, NN_A, KNN_A, $XGBoost_A$, and RF_A models, and 100 times for $ensemble_A$ models. A *z*-score above the dashed line (*z*-score=1.645, *p*-value=0.05) indicates that the model's performance is significantly better than random guess.

754

755 **Fig. S16. Performance of models built with three algorithms using the entire and subsets of**
756 **scRNA-seq or bulk RNA-seq data**. $F1_{test}$ scores of 13 metabolic classes for $FASTAI_{non_A}$
757 models (**A**), $F1_{CV}$ (upper panel) and $F1_{test}$ (lower) scores of 13 metabolic classes for NN_{non_A}
758 models (**B**), and $F1_{test}$ scores KNN_{non_A} models (**C**). Grey: models built using entire RNA-seq
759 data; rose, WT-normal subset; blue, WT-stress; purple, non-WT-normal; pink, root-specific WT-
760 normal scRNA-seq subset; bright pink, shoot-specific WT-normal scRNA-seq subset; light pink,
761 leaf-specific WT-normal scRNA-seq subset. Light gray background: metabolic classes with < 55
762 MPGs. Error bars indicate the standard deviation of results from 10 replicate runs. Asterisk(s)
763 indicate(s) significant levels (only showed when the subset-based models significantly
764 outperformed corresponding models built using the entire RNA-seq data) for the two-sided
765 Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01; ***, *p*-value < 0.001.

766

767 **Fig. S17. Performance of models built with $XGBoost_{non_A}$ and RF_{non_A} algorithms when the**
768 **entire and subsets of scRNA-seq or bulk RNA-seq data were used.** $F1_{CV}$ and $F1_{test}$ scores for
769 individual metabolic classes for $XGBoost_{non_A}$ (**A,B**) and RF_{non_A} (**C,D**) models built using the
770 entire sets and three subsets of scRNA-seq data, as well as three tissue-specific subsets of WT-
771 normal scRNA-seq data (left panel), and the corresponding entire and three subsets of bulk
772 RNA-seq data (right). Grey: model built using the entire scRNA-seq or bulk RNA-seq data; rose,
773 WT-normal subsets; blue, WT-stress; purple, non-WT-normal; pink, root-specific WT-normal
774 scRNA-seq subset; bright pink, shoot-specific WT-normal scRNA-seq subset; light pink, leaf-
775 specific WT-normal scRNA-seq subset. Light gray background: metabolic classes with < 55
776 MPGs. Error bar: standard deviation of results from 10 replicate runs. Asterisk(s) indicate(s)
777 significant levels (only showed when the subset-based models significantly outperformed
778 corresponding models built using the entire RNA-seq data) for the two-sided Wilcoxon rank-sum
779 test: *, *p*-value < 0.05; **, *p*-value < 0.01; ***, *p*-value < 0.001.

780

781 **Fig. S18. Comparison of prediction performance between models built using subsets of**
782 **scRNA-seq and bulk RNA-seq data.** Subsets are WT-normal (left panel), WT-stress (middle)
783 and non-WT-normal (right). (**A,C**) $F1_{test}$ for individual metabolic classes in $FASTAI_{non_A}$ and
784 KNN_{non_A} models. (**B,D,E**) $F1_{CV}$ (upper panel) and $F1_{test}$ (lower) for individual metabolic classes

785 in NN_{non_A}, $XGBoost_{non_A}$, and RF_{non_A} models. Light gray background: metabolic classes with <
786 55 MPGs. Orange: scRNA-seq subsets; blue: bulk RNA-seq subsets. Error bar: standard
787 deviation of results from 10 replicate runs. Asterisk(s) indicate(s) significant levels for the two-
788 sided Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01; ***, *p*-value < 0.001.
789
790 **Fig. S19. Performance of models built with five non-AutoGluon algorithms using scRNA-**
791 **seq (orange), bulk RNA-seq (blue), and the combined data (gray).** (**A-E**) results for
792 $FASTAI_{non_A}$, NN_{non_A}, KNN_{non_A}, $XGBoost_{non_A}$ and RF_{non_A}, respectively. No combined data-
793 based models showed significant improvement for MPGs predictions over scRNA-seq or bulk
794 RNA-seq-based models. Left panel: $F1_{CV}$; right panel: $F1_{test}$. Light gray background: metabolic
795 classes with < 55 MPGs. Error bars indicate the standard deviation of results from 10 replicate
796 runs.
797
798 **Fig. S20. Prediction of metabolic classes for unknown genes by KNN_{non_A} models**. (**A**) Count
799 of unknown genes that were predicted into 13 metabolic classes by KNN_{non_A} models built using
800 scRNA-seq (upper panel) and bulk RNA-seq data (lower). Inset scatter plot shows the correlation
801 between the number of benchmark MPGs and the number of predicted unknown genes within
802 classes when using scRNA-seq or bulk RNA-seq data. (**B**) Venn diagrams showing the overlap
803 of unknown genes that were predicted into metabolic classes by KNN_{non_A} models built using
804 scRNA-seq data (orange) and bulk RNA-seq (blue), and those that were annotated as within
805 metabolic classes in the PMN database (purple). In the upper and lower panels, overlapping
806 regions show the number of genes with exactly the same classes or at least one common class
807 between datasets, respectively. (**C**,**D**) Heatmaps showing the numbers of unknown genes that
808 were predicted as in a class (x-axis) by KNN_{non_A} models built using scRNA-seq (**C**) and bulk
809 RNA-seq data (**D**), and were also annotated as within the same or other classes (y-axis) by PMN
810 database. Values in the column "Not-assigned" indicate the numbers of unknown genes that were
811 predicted into neither any of the 13 metabolic classes nor the "Others" category. Cell colors in
812 (**C,D**) indicate the numbers of genes.
813

814 # Supplementary tables

815 **Table S1.** Information of 21 single-cell RNA sequencing (scRNA-seq) datasets.

816 **Table S2.** Information of 19 bulk RNA-seq datasets.

817 **Table S3.** Types of tissues, genetic backgrounds and conditions that cell types in scRNA-seq
818 data belong to.

819 **Table S4.** Types of tissues, genetic backgrounds and conditions that samples in bulk RNA-seq
820 data belong to.

821 **Table S5.** 10,572 enzyme-encoding genes obtained from the PMN, TAIR databases, or
822 annotated with EC numbers using the E2P2 software.

823 **Table S6.** 15 gene co-expression modules determined by WGCNA using scRNA-seq data.

824 **Table S7.** 12 gene co-expression modules determined by WGCNA using bulk RNA-seq data.

825 **Table S8.** Clustering coefficients of metabolic pathway genes within each class based on two
826 types of RNA-seq data, three subsets of scRNA-seq data, and three tissue-specific WT-normal
827 scRNA-seq data.

828 **Table S9.** Clustering coefficients for 10,000 random simulations calculated using two types of
829 RNA-seq data, three subsets of scRNA-seq data, and three tissue-specific WT-normal scRNA-
830 seq data.

831 **Table S10.** Overall $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations (100 for $ensemble_A$) of
832 predictive models built using two RNA-seq data.

833 **Table S11.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
834 of $FASTAI_{non_A}$ and NN_{non_A} models when scRNA-seq data were used.

835 **Table S12.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
836 of KNN_{non_A}, $XGBoost_{non_A}$, and RF_{non_A} models when scRNA-seq data were used.

837 **Table S13.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
838 of $FASTAI_A$, NN_A, and KNN_A models when scRNA-seq data were used.

839 **Table S14.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
840 of $XGBoost_A$ and RF_A, and for 100 random simulations of $ensemble_A$ models when scRNA-seq
841 data were used.

842 **Table S15.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
843 of $FASTAI_{non_A}$ and NN_{non_A} models when bulk RNA-seq data were used.

844 **Table S16.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
845 of KNN_{non_A}, $XGBoost_{non_A}$, and RF_{non_A} models when bulk RNA-seq data were used.
846 **Table S17.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
847 of $FASTAI_A$, NN_A, and KNN_A models when bulk RNA-seq data were used.
848 **Table S18.** $F1_{CV}$ and $F1_{test}$ scores of individual metabolic classes for 1,000 random simulations
849 of $XGBoost_A$ and RF_A, and for 100 random simulations of $ensemble_A$ models when bulk RNA-
850 seq data were used.
851 **Table S19.** Prediction of 9,441 unknown genes into individual metabolic classes for
852 $FASTAI_{non_A}$ and KNN_{non_A} models when scRNA-seq or bulk RNA-seq data were used.

853 Supplementary data

854 **Supplementary Data 1.** Gene expression matrix of scRNA-seq data.
855 **Supplementary Data 2.** Gene expression matrix of bulk RNA-seq data.

856 Reference

857 1. Al-Khayri, J. M. *et al.* Plant secondary metabolites: the weapons for biotic stress
858 management. *Metabolites* **13**, 716 (2023).
859 2. Appelhagen, I. *et al.* Colour bio-factories: towards scale-up production of anthocyanins in
860 plant cell cultures. *Metab. Eng.* **48**, 218–232 (2018).
861 3. Yang, Y.-H., Mao, J.-W. & Tan, X.-L. Research progress on the source, production, and
862 anti-cancer mechanisms of paclitaxel. *Chin. J. Nat. Med.* **18**, 890–897 (2020).
863 4. Ahmadi, M. K., Fang, L., Moscatello, N. & Pfeifer, B. A. *E. coli* metabolic engineering for
864 gram scale production of a plant-based anti-inflammatory agent. *Metab. Eng.* **38**, 382–388
865 (2016).
866 5. Wang, Y. *et al.* Deciphering phenylalanine-derived salicylic acid biosynthesis in plants.
867 *Nature* **645**, 208–217 (2025).
868 6. Elshafie, H. S., Camele, I. & Mohamed, A. A. A comprehensive review on the biological,
869 agricultural and pharmaceutical properties of secondary metabolites based-plant origin. *Int. J.*
870 *Mol. Sci.* **24**, 3266 (2023).

871 7. Tiwari, R. & Rana, C. S. Plant secondary metabolites: a review. *Int. J. Eng. Res. Gen. Sci.* **3**,
872 661–670 (2015).
873 8. Fang, C., Fernie, A. R. & Luo, J. Exploring the diversity of plant metabolism. *Trends Plant*
874 *Sci.* **24**, 83–98 (2019).
875 9. Jucá, M. M. *et al.* Flavonoids: biological activities and therapeutic potential. *Nat. Prod. Res.*
876 **34**, 692–705 (2020).
877 10. Bhambhani, S., Kondhare, K. R. & Giri, A. P. Diversity in chemical structures and biological
878 properties of plant alkaloids. *Molecules* **26**, 3374 (2021).
879 11. Zhang, S. *et al.* Research advances in clinical applications, anticancer mechanism, total
880 chemical synthesis, semi-synthesis and biosynthesis of paclitaxel. *Molecules* **28**, 7517 (2023).
881 12. Schläpfer, P. *et al.* Genome-wide prediction of metabolic enzymes, pathways, and gene
882 clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).
883 13. Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. & Osbourn, A. E.
884 Compromised disease resistance in saponin-deficient plants. *Proc. Natl. Acad. Sci. U.S.A.* **96**,
885 12923–12928 (1999).
886 14. Li, Y. *et al.* Subtelomeric assembly of a multi-gene pathway for antimicrobial defense
887 compounds in cereals. *Nat. Commun.* **12**, 2563 (2021).
888 15. Shang, Y. *et al.* Biosynthesis, regulation, and domestication of bitterness in cucumber.
889 *Science* **346**, 1084–1088 (2014).
890 16. Dong, X. *et al.* Spatiotemporal distribution of phenolamides and the genetics of natural
891 variation of hydroxycinnamoyl spermidine in rice. *Mol. Plant* **8**, 111–121 (2015).
892 17. Magnard, J.-L. *et al.* Biosynthesis of monoterpene scent compounds in roses. *Science* **349**,
893 81–83 (2015).
894 18. Tohge, T. *et al.* Exploiting natural variation in tomato to define pathway structure and
895 metabolic regulation of fruit polyphenolics in the lycopersicum complex. *Mol. Plant* **13**,
896 1027–1046 (2020).
897 19. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH:
898 automated identification, annotation and expression analysis of plant biosynthetic gene
899 clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
900 20. Töpfer, N., Fuchs, L.-M. & Aharoni, A. The PhytoClust tool for metabolic gene clusters
901 discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).

902 21. Moore, B. M. *et al.* Robust predictions of specialized metabolism genes through machine
903 learning. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2344–2353 (2019).
904 22. Wang, P. *et al.* Optimising the use of gene expression data to predict plant metabolic
905 pathway memberships. *New Phytol.* **231**, 475–489 (2021).
906 23. Sun, S. *et al.* Single-cell RNA sequencing provides a high-resolution roadmap for
907 understanding the multicellular compartmentation of specialized metabolism. *Nat. Plants* **9**,
908 179–190 (2022).
909 24. Yu, C. *et al.* Integrated mass spectrometry imaging and single-cell transcriptome atlas
910 strategies provide novel insights into taxoid biosynthesis and transport in *Taxus mairei* stems.
911 *Plant J.* **115**, 1243–1260 (2023).
912 25. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized
913 metabolism in plants. *Science* **344**, 510–513 (2014).
914 26. Onnela, J.-P., Saramäki, J., Kertész, J. & Kaski, K. Intensity and coherence of motifs in
915 weighted complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **71**, 065103 (2005).
916 27. Opsahl, T. & Panzarasa, P. Clustering in weighted networks. *Soc. Networks* **31**, 155–163
917 (2009).
918 28. Daloso, D. D. M., Morais, E. G., Oliveira E Silva, K. F. & Williams, T. C. R. Cell-type-
919 specific metabolism in plants. *Plant J.* **114**, 1093–1114 (2023).
920 29. Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. KNN model-based approach in classification.
921 in *On The Move to Meaningful Internet Systems: CoopIS DOA and ODBASE* vol. 2888 986–
922 996 (Springer, Berlin, Germany, 2003).
923 30. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in *Proceedings of the*
924 *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*
925 785–794 (ACM, San Francisco California USA, 2016). doi:10.1145/2939672.2939785.
926 31. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
927 32. Howard, J. & Gugger, S. Fastai: A layered API for deep learning. *Information* **11**, 108
928 (2020).
929 33. Stevens, E., Antiga, L. & Viehmann, T. *Deep Learning with PyTorch: Build, Train, and*
930 *Tune Neural Networks Using Python Tools*. (Manning, 2020).
931 34. Paszke, A. *et al.* Pytorch: an imperative style, high-performance deep learning library. in
932 *Advances in neural information processing systems* (Vancouver, Canada, 2019).

933 35. Erickson, N. *et al.* AutoGluon-tabular: robust and accurate AutoML for structured data.
934 arXiv at https://doi.org/2003.06505 (2020).
935 36. Bai, W. *et al.* Machine learning assists prediction of genes responsible for plant specialized
936 metabolite biosynthesis by integrating multi-omics data. *BMC Genomics* **25**, 418 (2024).
937 37. Guan, X., Okazaki, Y., Zhang, R., Saito, K. & Nikolau, B. J. Dual-localized enzymatic
938 components constitute the fatty acid synthase systems in mitochondria and plastids. *Plant*
939 *Physiol.* **183**, 517–529 (2020).
940 38. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq.
941 *Nat. Methods* **18**, 723–732 (2021).
942 39. Ke, M., Elshenawy, B., Sheldon, H., Arora, A. & Buffa, F. M. Single cell RNA-sequencing:
943 a powerful yet still challenging technology to study cellular heterogeneity. *BioEssays* **44**,
944 2200084 (2022).
945 40. Schilmiller, A. L., Last, R. L. & Pichersky, E. Harnessing plant trichome biochemistry for
946 the production of useful compounds. *Plant J.* **54**, 702–711 (2008).
947 41. McMahon, J. M., White, W. L. & Sayre, R. T. Cyanogenesis in cassava (*Manihot esculenta*
948 crantz). *J. Exp. Bot.* **46**, 731–741 (1995).
949 42. He, Z. *et al.* scPlantDB: a comprehensive database for exploring cell types and markers of
950 plant cell atlases. *Nucleic Acids Res.* **52**, D1629–D1638 (2024).
951 43. Mergner, J. *et al.* Mass-spectrometry-based draft of the arabidopsis proteome. *Nature* **579**,
952 409–414 (2020).
953 44. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
954 45. Kaiser, M. Mean clustering coefficients: the role of isolated nodes and leafs on clustering
955 measures for small-world networks. *New J. Phys.* **10**, 083042 (2008).
956 46. Hagberg, A., Swart, P. J. & Schult, D. A. Exploring network structure, dynamics, and
957 function using networkx. in *Proceedings of the 7th Python in Science Conference* 11–15
958 (Los Alamos National Laboratory, Los Alamos, USA, 2008).
959 47. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–
960 2830 (2011).