

1 **MEANtools: multi-omics integration towards metabolite** 2 **anticipation and biosynthetic pathway prediction**

3

4 Kumar Saurabh Singh ^{1, 3, 5, 6 #}, Hernando Suarez Duran ¹, Elena Del Pup ¹, Olga
5 Zafra-Delgado ², Saskia C.M. Van Wees ³, Justin J.J. van der Hooft ^{1, 4 #}, Marnix H.
6 Medema ^{1 #}

7 ¹ Bioinformatics Group, Wageningen University, Wageningen, the Netherlands

8 ² Departamento de Genética Molecular de Plantas, Centro Nacional de Biotecnología–Consejo Superior de Investigaciones
9 Científicas, Campus Universidad Autónoma, 28049 Madrid, Spain

10 ³ Plant-Microbe Interactions, Institute of Environmental Biology, Utrecht University, the Netherlands

11 ⁴ Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa

12 ⁵ Plant Functional Genomics, Brightlands Future Farming Institute, Maastricht University, the Netherlands

13 ⁶ Faculty of Environment, Science and Economy, University of Exeter, TR10 9FE Penryn Cornwall UK

14 # Corresponding authors, kumarsaurabh.singh@maastrichtuniversity.nl, justin.vanderhooft@wur.nl, marnix.medema@wur.nl

15

16 **Abstract**

17 During evolution, plants have developed the ability to produce a vast array of
18 specialized metabolites, which play crucial roles in helping plants adapt to different
19 environmental niches. However, their biosynthetic pathways remain largely elusive. In
20 the past decades, increasing numbers of plant biosynthetic pathways have been
21 elucidated based on approaches utilizing genomics, transcriptomics, and
22 metabolomics. These efforts, however, are limited by the fact that they typically adopt
23 a target-based approach, requiring prior knowledge. Here, we present MEANtools, a
24 systematic and unsupervised computational integrative omics workflow to predict
25 candidate metabolic pathways *de novo* by leveraging knowledge of general reaction
26 rules and metabolic structures stored in public databases. In our approach, possible
27 connections between metabolites and transcripts that show correlated abundance
28 across samples are identified using reaction rules linked to the transcript-encoded
29 enzyme families. MEANtools thus assesses whether these reactions can connect
30 transcript-correlated mass features within a candidate metabolic pathway. We validate
31 MEANtools using a paired transcriptomic-metabolomic dataset recently generated to
32 reconstruct the falcarindiol biosynthetic pathway in tomato. MEANtools correctly

33 anticipated five out of seven steps of the characterized pathway and also identified
34 other candidate pathways involved in specialized metabolism, which demonstrates its
35 potential for hypothesis generation. Altogether, MEANtools represents a significant
36 advancement to integrate multi-omics data for the elucidation of biochemical pathways
37 in plants and beyond.

38

39

40 **Introduction**

41

42 Plants have long been recognized for their ability to produce a variety of chemical
43 compounds, known as specialized metabolites (SM). It is estimated that a total of over
44 200,000 plant SMs have been reported so far that can be classified into distinct
45 metabolite classes, mainly terpenoids, alkaloids, phenolics, sulphur-containing
46 compounds, and fatty-acid derivatives ¹. Additionally, metabolomics has revealed an
47 extensive plant ‘dark matter’, in the sense that a major proportion of metabolites are
48 yet structurally unknown ². Also, the functions of most plant SMs are largely
49 unexplored, but they are generally regarded as crucial for fitness and survival ^{3–7}.
50 Humans have harnessed these chemical compounds in various areas, including
51 traditional medicines, pharmaceuticals, cosmetics, and agricultural products. The
52 biosynthesis of SMs, however, often hinges on external triggers and follows specific
53 metabolic pathways, which are largely unknown ⁸. This poses a substantial challenge
54 in obtaining, cultivating, and extracting these compounds in quantities suitable for
55 research or commercial production. This lack of knowledge has driven interest in
56 developing new methodologies to predict and identify new metabolic products as well
57 as the enzymes that catalyze their biosynthesis.

58

59 In the past decades, along with cost reductions, substantial progress in the generation
60 of high-throughput omics datasets has resulted in increasing numbers of high-quality
61 genome assemblies, transcriptome, metabolome, and enzyme reaction datasets ⁹.
62 Moreover, advances in synthetic biology allow the validation of *in silico* analyses *in*
63 *vivo*, increasing the rate at which novel SMs and the associated enzymes can be
64 characterized ¹⁰. This has amplified the discovery and characterization of biosynthetic
65 pathways in plants. Reconstructing biosynthetic pathways computationally requires
66 details about genes that encode enzymes catalyzing reactions, as well as the

67 metabolites involved in these processes. Tools such as plantiSMASH ¹¹, PhytoClust
68 ¹² and PlantClusterFinder ¹³ are instrumental in identifying gene clusters that are likely
69 to encode enzymes associated with SM pathways. Yet, many SM pathways in plants
70 do not have their genes chromosomally clustered. Additionally, co-expression
71 analyses can be employed to predict functional associations between genes based on
72 their expression patterns ^{14,15}. In general, individual omics-based investigations, such
73 as genomics, transcriptomics, or metabolomics, have played pivotal roles in
74 delineating specific metabolic pathways and their correlated metabolic products ^{16–27}.
75 Nevertheless, despite these advancements, the intricate genetic makeup and
76 functional diversity of plant biosynthetic pathways continue to present a formidable
77 challenge. Specifically, a key limitation to current transcriptome- and metabolome-
78 based pathway discovery strategies is that they require prior knowledge on a
79 compound or enzyme that can be used as ‘bait’ ²⁸ to identify other compounds and/or
80 enzymes involved in the same pathway. Yet, prior such knowledge may not always be
81 available.

82 A promising solution to this limitation may be found in the integrative analysis of
83 genomic, transcriptomic, and metabolomics data. Due to the intricate, cooperative
84 interplay of genes and metabolites in SM biosynthesis, implementing multi-omic
85 approaches ensures a comprehensive perspective on the entire process. Indeed, the
86 inclusion of multiple omics layers has facilitated the discovery of several biosynthetic
87 pathways ^{29–35}. Multi-omics integration strategies can be broadly separated into four
88 categories: conceptual, statistical, model, and pathway-based. Each strategy presents
89 distinct challenges, and all have been reviewed in detail before, with multiple examples
90 of successful usage ^{36,37}. Such integrative omics technologies ⁹ provide new
91 opportunities for systematic, unsupervised multi-omics approaches for untargeted or
92 *de novo* discovery of pathways involved in the biosynthesis of SMs.

93

94 Here, we introduce MEANtools, a computational pipeline that combines statistical- and
95 reaction-rules-based integration strategies. MEANtools implements a mutual rank-
96 based ¹⁵ correlation approach to capture mass features that are highly correlated with
97 biosynthetic genes. Our pipeline makes use of general reaction rules and metabolite
98 structures, stored in public databases like RetroRules ³⁸ and LOTUS ³⁹, to predict
99 putative reactions that either constitute intermediate steps or complete biosynthetic
100 pathways. The workflow enables users to explore the biosynthetic potential associated

101 with identified mass features and formulate specific hypotheses about potential
102 pathways associated with the corresponding metabolites.

103

104 **Results**

105

106 *MEANtools integrates omics data to link transcripts to metabolites*

107 MEANtools integrates mass features from metabolomics data and transcripts from
108 transcriptomics data to predict possible metabolic reactions and thus generates
109 hypotheses that can be prioritized for experimental validation (Figure 1a). Reaching
110 the prediction stage involves several independent steps, including formatting and
111 annotating the input data, thereby ensuring the data is ready for subsequent
112 meaningful analysis. MEANtools then leverages RetroRules³⁸, a retrosynthesis-
113 oriented database of enzymatic reactions annotated with known and predicted protein
114 domains and enzymes linked to each reaction, to assess whether observed chemical
115 differences between metabolites (inferred from observed mass shifts) can logically be
116 explained by reactions that are known to be catalyzed by transcript-associated protein
117 families (Figure 1b). To identify putative structure annotations for metabolite features,
118 MEANtools matches their masses to LOTUS³⁹, a comprehensive well annotated
119 resource of Natural Products, taking into account possible adducts (Figure 1c).
120 MEANtools correlates the expression of genes with co-abundant metabolites across
121 samples in paired transcriptomics and metabolomics experiments, ideally spanning a
122 range of different conditions, tissues and timepoints. Although the correlation
123 approach has aided the characterization of diverse metabolic processes in plants by
124 reducing the dimensionality of the problem and thus generating a small set of testable
125 hypotheses, it is known to result in a high number of false positive metabolite-transcript
126 associations when used in isolation. As illustrated in Figure 1d, we use a mutual rank-
127 based correlation method that maximizes highly correlated metabolite-transcript
128 associations.

129

130 MEANtools then integrates all this information to identify sets of transcript-metabolite
131 pairs that are both highly correlated in abundance and then highlight cases where the
132 metabolites are logically connected by catalytic activities associated with these same
133 transcripts. Thus, MEANtools generates a reaction network where each node is a
134 mass signature within the metabolome, or an unmeasured *ghost mass signature*⁴⁰. In

135 this network, nodes are linked by directed edges representing enzymatic reactions that
136 can be catalyzed by at least one of the enzyme families encoded by the genes
137 correlated to one of the two mass signatures the reaction links. This network
138 representation of the data allows users to explore the biosynthetic potential of any
139 molecular structure and generate concrete hypotheses about possible pathways
140 leading up to (or from) a given metabolite, which can be tested in the laboratory.
141 Results are displayed in a variety of formats for users to interact with, describing
142 predicted metabolic pathways along with the metabolites, enzymes and reactions that
143 are potentially involved in them. Altogether, MEANtools serves as a strong basis for
144 the development of methodologies to explore ways in which paired genomic,
145 transcriptomic, and metabolomic data can be used to analyze biosynthetic diversity.

146

147 *RetroRules and LOTUS database integration*

148

149 In the above process, strongly correlated mass feature-transcript pairs are examined
150 using the general reaction rules obtained from the RetroRules. All enzymatic reactions
151 in the RetroRules database are cross-referenced with the MetaNetX ⁴¹, a repository of
152 metabolic networks that MEANtools uses to identify the mass differences (shifts in the
153 masses) between the substrates and products of known enzymatic reactions (Figure
154 1e). MEANtools then annotates all reactions with an associated mass shift. This step
155 needs to be executed only once, either during the initial retrieval of the database or
156 when it is updated. As a next step, users can manually annotate a subset of mass
157 signatures (mass-to-charge ratios of the measured ions) in the metabolomic dataset
158 with metabolite structures (Figure 1f & g). Alternatively, MEANtools can assign
159 potential structure matches by identifying adducts in the metabolome and querying the
160 LOTUS database for matching metabolites based on molecular weight (Figure 1c).

161

162 To determine the significance of the presence of experimentally characterized
163 biosynthetic reactions in the RetroRules database, we tested the presence of selected
164 biosynthetic reactions from the Singh et al., review Figure 1 ⁹ (Supplementary File 1).
165 Among 187 experimentally characterized biochemical reactions, 134 were found in the
166 RetroRules database and 53 were missing. The presence of 72% of selected reactions
167 in the RetroRules database is significantly higher (χ^2 -statistic: 35.10; DF=1; $p < 0.001$)

168 than expected under the null hypothesis of equal probability. This indicates that
169 RetroRules database has a good coverage of experimentally characterized
170 biosynthetic reactions, enhancing its reliability for further pathway analysis.
171 Additionally, for the same set of experimentally characterized reactions, we
172 investigated the presence of structures for both the substrates and the products, from
173 the list of experimentally characterized biosynthetic reactions, in the LOTUS database
174 (Supplementary File 1). Compared to the total 374 structures from the selected
175 reactions, 132 structures were found in the database with a significance of $p < 0.001$
176 (χ^2 -statistic: 32.353; DF=1), highlighting substantial structural overlap.

177
178 RetroRules is populated with ~43,000 reactions annotated with enzymes that are
179 predicted to be associated with all reactions. Most of these annotated enzyme-reaction
180 associations, however, are the result of propagating the annotation of characterized
181 reactions to other reactions with the same enzyme commission (EC) number and they
182 therefore of various reliability and require verification. To increase confidence in the
183 enzymatic annotations, we cross referenced each reaction in RetroRules to the
184 manually curated reaction databases Rhea ⁴² and KEGG ⁴³. We refined reaction-
185 enzyme associations supported by experimental evidence and then propagated these
186 annotations through KEGG-orthology groups (Methods). This way, we generated
187 three datasets namely, *strict*, *medium*, and *loose*, differing in the coverage of chemical
188 space and confidence in the enzymatic annotations. This was done to remove the
189 most generic Pfam annotations. *Loose* dataset contains 2,704,948 reaction rules-
190 enzyme associations expanded from the RetroRules database by cross-referencing
191 with the Rhea and KEGG-orthology database (Supplementary Figure 1). *Medium*
192 dataset contains 429,267 entries consist of experimentally validated entries together
193 with the ECDomainMiner predictions. Finally, the *strict* dataset contains 67,501
194 experimentally validated entries (Supplementary Figure 1). These datasets are
195 specifically developed for enzyme function prediction and are especially relevant when
196 specificity is preferred over sensitivity. All three datasets come with taxonomic origin
197 annotations. Users can therefore not only select the datasets between *loose*, *medium*,
198 *strict*, but also use the taxonomy of the samples (Supplementary Figure 2) for further
199 refinement of their analyses based on the species-specificity of Pfams.

200

201

202 *Reconstruction of the falcarindiol pathway in tomato*

203

204 To assess the performance of MEANtools in predicting metabolic pathways, we used
205 data derived from a recently published paired omics dataset. Specifically, we assessed
206 whether MEANtools would be able to reconstruct the falcarindiol pathway in tomato
207 using the dataset published by Jeon *et al.* in 2020 ³² in the study that originally
208 elucidated this pathway. MEANtools correctly anticipated five out of seven
209 transformations of intermediate metabolites in the falcarindiol pathway, along with the
210 enzymes that catalyze the reactions. The initial untargeted metabolomics and
211 transcriptomics data comprised 11266 mass features and 20576 transcripts. To
212 narrow down the counts and select the most informative mass features and transcripts,
213 we performed differential abundance analysis of mass features and differential
214 expression analysis of transcripts across samples and time-points. After selecting
215 features and transcripts based on a corrected p-value and log fold change threshold
216 of 0.01 and 2, respectively, 1230 mass features and 7590 transcripts remained.
217 Correlation analysis (step 1), with a minimum absolute *Pearson* correlation coefficient
218 of 0.1, further refined the count of informative mass features and transcripts. Four
219 networks (N) were created with different decay rates (DR). The number of transcripts
220 and mass features assigned to functional clusters in N1 (DR=5) were 2912 (38.4% of
221 input genes) and 232 (18.9% of input mass features) respectively. Similarly, for N2
222 (DR=10) the count was 5488 (72.3%) and 236 (19.2%). For N3 (DR=25) and N4
223 (DR=50) the count was 6491 (85.5%) / 238 (19.3%), and 6420 (84.6%) / 238 (19.3%)
224 respectively. MEANtools also returns a p-value for every transcript-mass feature
225 correlation. This p-value is based on the hypothesis test whether the true correlation
226 between the two datasets is zero. The distribution of the p-values resulting from the
227 correlation step (Supplementary Figure 3) is heavily skewed towards the right and
228 significantly (Kolomogorov-Smirnov statistics=0.987; p-value= \sim 0.0) deviates from
229 what would be expected under the null hypothesis of no significant effects, showing a
230 subset of transcripts and mass features that are significantly associated and reflecting
231 real biological interactions.

232

233 In the functional clusters (FCs), we first looked for biosynthetic genes (based on
234 classification using plantiSMASH profile hidden Markov models) predicted to be

235 involved in SM pathways, specifically for falcarindiol-related genes³². Our analysis
236 revealed a single FC in N2 encompassing three out of the four biosynthetic genes from
237 this cluster (Figure 2c). This FC, containing all three key biosynthetic genes related to
238 the falcarindiol pathway, also included a CYP450 gene suspected to be involved in the
239 modification of dehydrocrepenynic acid—one of the pathway intermediates within the
240 pathway³². Other FCs that harbored mass-features present in Figure 2c were merged
241 and taken further to the pathway prediction step of MEANtools. By using only
242 experimentally validated enzyme-reaction associations (*strict* settings), MEANtools
243 anticipated the second step of the falcarindiol biosynthesis pathway as proposed by
244 Jeon et al. (crepenynic acid -> dehydrocrepenynic acid), seen in Figure 3b. For this
245 step, MEANtools predicted Solyc12g100250.1, which shows strong correlation (0.744;
246 p-value 1.106E-10 (Figure 2D; Supplementary File 2) that Jeon et al. identified as a
247 major desaturase in the falcarindiol pathway that was linked to this reaction using
248 transient expression³². MEANtools also anticipated steps five and six of the pathways
249 proposed by Jeon et al., (i.e., octadecene diynoic acid -> octadecadiene diynoic acid
250 -> metabolite_6 -> metabolite_7), as seen in Figure 3a, and provided candidate genes
251 encoding enzymes with a protein domain that has been characterized as able to
252 perform each reaction. To further explore the predictive power of MEANtools, we
253 repeated the analysis with *medium* and *loose* settings. As we moved from strict to
254 medium and then to *loose* settings, we observed an increase in enzyme associations
255 due to the inclusion of less specific Pfam annotations (Supplementary Figure 6).
256 Distribution of the correlation coefficients of all mass feature-transcript associations
257 for the falcarindiol pathway can be seen in Supplementary Figure 4. A table with all
258 the predictions is available in Supplementary File 2.

259
260 *Identification of Functional Clusters encompassing other tomato metabolic pathways*

261
262 Within the Jeon *et al.*, dataset³², a wider investigation unveiled multiple FCs housing
263 biosynthetic genes primarily from three distinct metabolic pathways: the hydroxy
264 cinnamic acid amide (HCAA) pathway⁴⁴ the α -tomatine pathway⁴⁵, and the
265 chlorogenic acid pathway⁴⁶.

266
267 We identified two FCs containing biosynthetic genes associated with the synthesis of
268 *p*-coumaroyl-CoA from phenylalanine, a process catalyzed by PAL and 4CL, as well

269 as the subsequent biosynthesis of *p*-coumaroyltyramine, a reaction mediated by THT
270 (Figure 4). Interestingly, all metabolites within these two functional clusters were
271 putatively annotated within the superclass of phenylpropanoids and polyketides
272 (Supplementary File 1). Figure 4c depicts the FC containing PAL (Solyc10g086180;
273 node with a pink border) and 4CL (Solyc03g117870; node with orange border), and
274 metabolites involved in the conversion of phenylalanine to *p*-coumaroyl-CoA. This FC
275 also contains other co-expressed genes along with PAL and 4CL. Additionally, the
276 correlation analysis performed by MEANtools revealed another FC (Figure 4f) related
277 to the production of *p*-coumaroyltyramine catalyzed by THT (Solyc08g068790; node
278 with a red border). According to the expression heatmaps depicted in Figure 4a and
279 b, while PAL and 4CL showed constitutive expression patterns across both mock and
280 treated samples, THT exhibited significant differential expression (p -value < 0.05 and
281 $\log_{2}FC = 4.7$) in samples treated with fungal pathogens compared to mock-treated
282 ones. Both the metabolite and genes present in the THT FC (Figure 4f) show
283 overlapping abundance and expression patterns (highlighted with black solid bar in
284 the heatmaps of Figure 4a-b).

285

286 In another biosynthetic pathway, namely the α -tomatine pathway, we observed the
287 presence of genes distributed across multiple FCs (Supplementary Figure 5). This
288 pathway involves nine specific biosynthetic genes responsible for converting
289 cholesterol into α -tomatine, and these genes have been extensively characterized in
290 tomato ⁴⁵. MEANtools captured all biosynthetic genes involved in the glycoalkaloid
291 metabolism (GAME) group, including GAME1, GAME4, GAME6, GAME7, GAME11,
292 GAME12, GAME17, and GAME18, in nine different FCs (Supplementary figure 9).
293 Furthermore, GAME9, an APETALA2/Ethylene response factor, related to regulator of
294 the steroidal glycoalkaloid pathway in tomato, was also captured within one of the 9
295 FCs. Additionally, we found biosynthetic genes involved in the synthesis of precursors
296 for the α -tomatine pathway, such as SQS (Squalene Synthase), TTS1 (β -Amyrin
297 Synthase), TTS2 (β -Amyrin Synthase), and SSR2 (Sterol Side Chain Reductase 2),
298 present in multiple instances throughout the network. We used coexpression network
299 to merge FCs, resulting in coexpression edges between biosynthetic genes from the
300 α -tomatine pathway GAME12 transaminase, and 2-oxoglutarate-dependent
301 dioxygenase GAME11, and GAME17 (UDP-glucosyltransferase) and GAME1 (UDP-
302 galactosyltransferase) (Supplementary figure 8). Additionally, MEANtools successfully

303 pinpointed another crucial biosynthetic gene associated with the chlorogenic acid
304 biosynthetic pathway, known as HQT (Hydroxycinnamoyl-CoA quinate:
305 hydroxycinnamoyl transferase). HQT plays a pivotal role in facilitating the
306 transformation of quinic acid into caffeoyl quinic acid, which represents another
307 specialized metabolite within the phenylpropanoid pathway.

308

309 *MEANtools facilitates prioritization of reaction steps using reaction likelihood scores*
310 MEANtools generates reaction-likelihood scores based on substrate-enzyme
311 association, for each anticipated reaction (Figure 5). To obtain the score, the likelihood
312 of each atom in the substrate is calculated for being a site-of-metabolism using the
313 GNN-SOM⁴⁷ method. This results in an array of likelihoods for each atom in the
314 substrate. Later, using ReactionDecoder⁴⁸, reaction centers and bond cleavages are
315 predicted between each substrate and product. MEANtools makes use of this
316 information to extract likelihood scores only for atoms that are involved in reaction
317 centers and bond cleavages. The maximum value of likelihood score within the
318 reaction center represents the reaction likelihood score. Figure 5c shows the
319 distribution of likelihood scores for experimentally characterized enzyme-substrate
320 pairs, referred to as *Known* in Supplementary file 1, and randomly assembled enzyme-
321 substrate pairs as *Random*. The likelihood scores differ significantly (Mann-Whitney U
322 statistic: 2573.0, P-value: 4.4e-07) between Known and Random pairs, with median
323 and mean values of 0.86 and 0.70 for Known pairs, and 0.29 and 0.39 for Random
324 pairs, respectively.

325

326 **Discussion**

327

328 MEANtools can generate testable hypotheses on metabolic pathways with little to no
329 prior knowledge, by integrating metabolomics and transcriptomics data. This method
330 effectively automates the identification of key Pfam domains required for a specific
331 reaction and allows users to tune the reaction-Pfam domain associations according to
332 their level of confidence or based on the taxa of origin. To do so, MEANtools queries
333 RetroRules, a retrosynthesis-oriented enzymatic reactions database, showing that
334 tools and methods within the retrosynthetic biology and synthetic pathway design
335 fields have considerable application potential for metabolic pathway prediction and
336 potentially SM discovery.

337

338 Metabolomics and transcriptomics datasets are typically used as CSV-formatted pre-
339 processed tables featuring mass-feature abundances and transcript expressions,
340 respectively. Integrating such datasets solely through Pearson-based correlations
341 often results in many false-positive associations. Additionally, determining an optimal
342 threshold for eliminating weak correlations poses significant challenges. The use of
343 mutual-rank statistics has proven effective for constructing global gene co-expression
344 networks, as demonstrated by Wisecaver *et al* ⁴⁹. Leveraging this approach, we
345 utilized the mutual rank-based method to develop a correlation-based global gene-
346 metabolite network. This network highlights strongly correlated genes and
347 metabolites. Ideally, individual FCs should advance to the next stage of pathway
348 prediction. However, the FCs size may sometimes be insufficient for forming a
349 complete biosynthetic pathway. The FCs size proved stable across the treatment
350 combinations in Jeon *et al.* (2020) dataset ³² (Supplementary Figure 7). Given that
351 genes and metabolites in plant biosynthetic pathways tend to overlap, FCs are also
352 overlapping in nature. MEANtools provides a script (*merge_clusters.py*) to merge
353 multiple FCs that share common mass features. Mass features that exhibit distinct
354 abundance patterns across samples are then grouped into separate clusters following
355 this merging process. This step is crucial for ensuring enough mass features and
356 transcripts remain to either fully or partially reconstruct a biosynthetic pathway.
357 Changing the size of FCs is also possible using the ClusterONE ⁵⁰ inbuilt parameter.
358 However, this also changes the clustering pattern of mass features and transcripts.
359 Additionally, the current method to provide significance to each FC could also be
360 improved, as this was originally developed for co-expression datasets.

361

362 The RetroRules database is publicly available as SQLite database and can be used
363 directly with MEANtools. The three reaction rules datasets resulting from RetroRules,
364 *loose*, *medium*, and *strict* are available in a single CSV file in the GitHub repository.
365 MEANtools includes these three different datasets as an input parameter (*strict*,
366 *medium*, and *loose*, respectively) to allow the user to constrain the predictions for
367 specific purposes and find the right balance between sensitivity and specificity,
368 considering the tradeoff between enzymatic annotation confidence and diversity of the
369 resulting set of enzymatic reactions. In the *strict* dataset, which is a smaller subset of
370 reaction-rule-enzyme associations, the number of resulting candidate genes in the

371 final reaction anticipations was reduced due to its more specific Pfam annotations. On
372 the contrary, reaction anticipations with the *loose* set were associated to unrelated
373 Pfams (Supplementary Figure 7), such as AminoTran_1_2 or Glyco_transf_20,
374 responsible for transferring amino and sugar groups respectively⁵¹. Such unrelated
375 Pfams were not found in the *strict* rule dataset, as shown in the hydroxylation of
376 octadecadiene-diyonic acid into metabolite 6 and its subsequent hydroxylation into
377 metabolite 7. These spurious links, coming from RetroRules, have been kept in the
378 *loose* dataset after applying the Pfam cutoff of 6, which highlights the importance of
379 using the strict rules dataset when specificity is preferred over sensitivity. Most
380 importantly, both predictions correctly predict the enzyme associated to the conversion
381 of octadecadiene- diyonic acid into metabolite 6, as reported by Jeon *et al.*, (2020)³².
382 Transient expression of this enzyme in *Nicotiana benthamiana* (Solyc10g100250) was
383 experimentally associated to depletion of crepenynic acid and the production of two
384 new metabolites³². According to the observed LC-MS profile, one of the metabolites
385 was putatively identified as octadecadiene- diyonic acid, making plausible the role of
386 Solyc10g100250 in its conversion to metabolite 6. By cross-checking these reaction-
387 enzyme association datasets with sets of correlated enzyme-coding genes and
388 metabolites, MEANtools effectively filters the set of possible mass shift-reaction
389 associations based on the available -omics evidence.

390

391 In the reconstruction of PAL and THT biosynthetic pathways, the reconstruction of
392 reaction steps using the second step of MEANtools was hindered due to two main
393 factors. Firstly, the conversion of phenylalanine to p-coumaroyl-CoA involves
394 stereoisomers, which are not captured by the mass spectrometric data. Secondly, the
395 conversion to p-coumaroyltyramine requires two substrates, tyrosine, and p-
396 coumaroyl-CoA, whereas RetroRules-based rules are designed for single-substrate
397 reactions only. Although RetroRules contains a rule for the stereomeric conversion of
398 phenylalanine to p-coumaroyl-CoA, MEANtools filters such rules involving stereomeric
399 structures to avoid complexity.

400

401 The initial construction of this substructure map occurs once, either during the initial
402 retrieval of the RetroRules database or during updates (step 1 Figure 1b). MEANtools
403 uses this substructure map to generate pathway predictions. To this end, in step 2
404 (Figure 1c), it predicts possible metabolites and their corresponding molecular

405 structures for each mass feature by identifying possible adducts and querying the
406 LOTUS database, or a user-defined metabolite database that can be supplied in CSV
407 format. The LOTUS database was converted to an SQLite format to be compatible
408 with MEANtools, and it was made available in the GitHub repository. In step 3 (Figure
409 1e-g), MEANtools exclusively queries reactions that would yield metabolites with mass
410 features that can be mapped within the metabolome or as ghost mass signatures.
411 Collectively, this strategic approach enables MEANtools to efficiently utilize computing
412 resources when generating *in-silico* molecules. Because of step 3, MEANtools
413 produces a sequence of subsequent reactions, along with predicted products for all
414 pairs of mass signatures, correlated enzyme-coding genes, and references to
415 characterized reactions and enzymes that served as the rules for predicting these
416 reactions. Step 3 can be iterated multiple times, as desired by the user, enabling the
417 generation of pathway predictions extending beyond a single enzymatic reaction away
418 from the initial query molecule.

419
420 Because of MEANtools' flexible and modular design, there is room for improvement in
421 many of its processing steps. Annotating mass signatures with predicted structures
422 can be improved by allowing to load MS/MS data and use mass spectral library and
423 networking-based annotation approaches⁵² to increase accuracy and allow validation,
424 in a similar way as done by MetWork⁴⁰. Gene-metabolite clusters can further be
425 improved by a more elaborate co-expression and/or molecular network analysis.
426 Converting predicted reaction networks into directed acyclic graphs (DAGs) is
427 currently used to study and present unsupervised predictions, but more complex
428 manipulations of the network may allow for predictions better tailored for the user, such
429 as prioritizing specific reactions or molecular substructures, for example by integrating
430 MS2LDA analyses⁵³. We also note that further curating the reaction-Pfam domain
431 associations or allowing the user better control over them by allowing customized
432 reaction-rule databases could improve the method as well: some enzyme domains
433 may be linked to large numbers of reactions, likely leading to false positives when the
434 objective is to predict biosynthesis pathways, but these enzyme domains could be
435 useful when exploring the biosynthetic potential of a structure when designing a
436 synthetic pathway. Finally, the reaction likelihood scores can also be improved by
437 adopting or developing precise methods for reaction site predictions.

438

439 Altogether, we present a novel computational method to predict metabolic pathways
440 guided by multi-omics evidence, allowing researchers to conveniently generate
441 testable and easy-to-browse hypotheses. Furthermore, we anticipate that our work
442 provides the basis for future work to expand the numbers of ways in which paired
443 genomic, transcriptomic and metabolomic data can be used to link natural product
444 chemistry to biosynthesis genes and producers, and to analyze biosynthetic diversity
445 in nature.

446

447 **Methods**

448

449 *Correlation-based integration generates testable associations*

450 Global reconstruction of co-expression modules in gene expression data has been
451 shown to be a powerful method to identify groups of genes involved in the same
452 metabolic pathway when querying for modules with genes that encode biosynthetic
453 enzymes ⁴⁹. In MEANtools, instead of generating co-expression modules using
454 transcriptomics dataset, functional clusters (FC) are generated for different network
455 sizes by integrating transcriptomics and metabolomics data (Figure 1d). Inspired from
456 the work of Wisecaver *et al.* ⁴⁹, correlation values between mass features and
457 transcripts are first converted to mutual ranks (MR) ⁵⁴, which are then subjected to an
458 exponential decay function that converts continuous MR values to numbers between
459 0 and 1 and referred here as *edge weights*. Both node types and edge weights are
460 further subjected to clustering using ClusterONE ⁵⁰, which results in multiple
461 overlapping FCs. Each FC represents a significant association of mass abundance
462 and transcripts expression patterns across samples. In a network view, mass feature
463 and transcripts represent two unique node types connected by edge weights.
464 MEANtools allows users to visualize the expression of each FC in the form of
465 heatmaps with transcripts sorted in three categories according to the protein domains
466 they encode, following the same categorization used by plantiSMASH: scaffold-
467 generating enzymes, tailoring enzymes, and the remaining genes ¹¹.

468

469 *Rescaling input data using Median Absolute Deviation (MAD)*

470

471 MEANtools employs the Median Absolute Deviation (MAD) for data rescaling. MAD
472 calculates the median of all values within the dataset, which represents the 50th

473 percentile. It then determines the absolute difference between each value and the
474 calculated median and ensures that the differences are expressed as positive values,
475 regardless of whether they are greater or less than the median. Finally, computing the
476 median of these absolute differences yields the MAD (equation 1).

477

$$478 \quad \text{MAD} = \text{median}(|X_i - \text{median}(X)|)$$

479 (1)

480 where X_i refers to the i_{th} row element present in the data matrices

481

482 *Computation of mutual rank and edge weights*

483

484 Pairwise correlations of transcripts and mass features are converted to mutual ranks
485 (MR; calculated as a geometric mean of the rank of *Pearson* correlation coefficient
486 (PCC) of transcript A to mass feature A and of the PCC rank of mass feature A to
487 transcript A. This MR statistic is calculated for every transcript-mass feature pair. Since
488 the MR value can vary between 1 and $n-1$, where n represents the total number of
489 features in either the transcriptomic or metabolomic dataset, we transform the MR
490 scores into *edge weights*, ranging between 0 and 1, using an exponential decay
491 function⁴⁹. By default, MEANtools computes edge weights by using four different rates
492 of decay (5, 10, 25, 50) resulting in five different networks of varying sizes (equation
493 2). The modified exponential decay function is:

494

$$495 \quad N_{i \rightarrow j} : e^{-(MR-1)i \rightarrow j}$$

496 (2)

497 where $i \rightarrow j$ refers to multiple decay rates and $N_{i \rightarrow j}$ represents a combined network
498 generated using $i \rightarrow j$ decay rates. MR denotes the estimated mutual rank between
499 genes and metabolites. Gene-metabolite pairs that show lower edge score than 0.01
500 are excluded in the $N_{i \rightarrow j}$ networks.

501

502 MEANtools then employs the graph-clustering method ClusterONE⁵⁰, which identifies
503 overlapping clusters of transcripts and mass features. Clustered transcripts and mass
504 features can assemble into a biologically significant sub-network, which we refer to as
505 a *functional cluster* (FC). Such clusters represent a higher-level organization of the

506 transcriptome and metabolome. The average number of FCs per network decreases
507 with increasing network size. For each FC, ClusterONE assigns a p-value derived from
508 the comparison between edges within the FC and those that radiate out of the FC. The
509 resulting network, stemming from various decay rates, is then stored within an SQLite
510 database.

511

512 *Mass-shifts associated to reactions serve as templates for pathway predictions*

513 MEANtools leverages the established relationships between reactions and their
514 associated mass shifts to scan the input metabolome. It assigns molecular structures
515 to each mass feature of the metabolome by mapping them with a list of adducts and
516 then querying LOTUS database (downloaded on 10/10/2023). LOTUS database was
517 converted to an SQLite format to be compatible with MEANtools and made available
518 in the GitHub repository. It identifies pairs of mass features with discernible differences
519 in mass-charge ratios that can be logically explained by known reactions. Within this
520 process, one mass feature is annotated as a potential substrate, while the other is
521 marked as a product. It is worth noting that a given mass shift might be assigned to
522 more than one reaction, and many reactions are bidirectional in nature. Consequently,
523 any pair of mass features can be associated with multiple reactions, considering both
524 forward and reverse directions. Additionally, recognizing that not all metabolites within
525 a metabolic pathway may reach detectable levels in the (measured) metabolome,
526 MEANtools optionally generates 'ghost mass signatures.' These ghost signatures
527 serve as virtual, unmeasured intermediates between any two metabolites that possess
528 measured mass signatures. This concept, recently introduced in MetWork ⁴⁰ in the
529 construction of metabolic networks based on MS/MS spectra, is also applied here.
530 Notably, although the ghost mass feature is provided as an option to use for all
531 reactions, it automatically gets switched on when MEANtools fails to assign mass
532 features either as substrates or products. By incorporating information on reaction-
533 mass-shift associations, MEANtools constructs a comprehensive reaction network.
534 This network comprises mass signatures connected by annotated reactions and forms
535 the foundational framework for the subsequent prediction of metabolic pathways.

536

537 *Prediction of metabolic pathways*

538 MEANtools leverages the reaction network to facilitate the generation of pathway
539 predictions. Initially, it predicts potential metabolites along with their corresponding

540 molecular structures for each mass signature. Subsequently, MEANtools employs the
541 RDKit ^{55,56} Python package (v 2019.03.2.0) to computationally generate *in silico*
542 structures resulting from each reaction-associated substrate.

543

544 Given the substantial number of reactions cataloged in RetroRules, generating all
545 product molecules for the metabolite structures predicted in a metabolome by querying
546 every reaction can be time-consuming and computationally intensive. From each
547 reaction, new metabolites emerge, leading to a large number of molecular structures.
548 To expedite this process, MEANtools relies on -omics evidence, specifically the
549 reaction-substrate-enzyme pairs under the confinement of FCs, to guide the
550 generation of *in-silico* molecules.

551

552 Further acceleration is achieved by targeting specific substructures within each
553 metabolite structure, employing a divide-and-conquer strategy (Figure 6). For each
554 metabolite structure, MEANtools initiates by verifying the presence of specific atoms,
555 such as N or C. Upon success, the next step involves querying reactions that pertain
556 only to simple substructures, like N=N and C=C. If both atoms are present, MEANtools
557 extends its search to reactions centered on substructures like C=N and C-N. In
558 subsequent rounds, MEANtools explores more complex substructures based on the
559 substructures identified in prior steps. For instance, metabolites with the C=N
560 substructure are exclusively queried for reactions centered on the C=N-C
561 substructure. This iterative process continues until no further successful queries are
562 obtained for a given metabolite.

563

564 *Easy-to-browse MEANtools output*

565 MEANtools generates user-friendly visualizations and supplementary data in the form
566 of easy-to-browse tables. MEANtools stores these tables in an SQLite database. It
567 also offers python-based utility scripts to retrieve and visualize FCs within the MR-
568 based correlation network.

569

570 MEANtools analyzes the reaction network created in the preceding stages to predict
571 candidate metabolic pathways aligned with the user's interests. To accomplish this,
572 the NetworkX ⁵⁶ Python package (v2.4) is utilized. MEANtools constructs a distinct
573 subnetwork for each of the initial metabolites provided by the user. These subnetworks

574 are transformed into directed acyclic graphs (DAGs) by identifying any cycles within
575 the network, representing potential reversible reactions. Only links capable of
576 advancing the reaction away from the initial metabolite are retained. In instances
577 where cycles occur among metabolites at the same reaction distance from the initial
578 metabolite, the edge featuring the weakest enzyme-metabolite correlation is
579 eliminated. This approach yields multiple DAGs rooted at the initial metabolites, each
580 offering the potential for candidate metabolic pathways. The longest reaction path in
581 each subnetwork, commencing from the initial metabolite, is identified to predict these
582 pathways. This process is repeated to generate a DAG for each initial metabolite at
583 the termination of the reaction, yielding two pathway predictions for each input
584 structure. MEANtools then delivers the complete reaction network and all DAGs in the
585 form of CSV tables, facilitating seamless import and exploration within Cytoscape.
586 Furthermore, pathway predictions are presented as SVG image files, providing
587 comprehensive details regarding the involved metabolites, reactions, genes, and their
588 respective correlations. To enhance user exploration, MEANtools offers an option to
589 generate SVG files for each molecular structure predicted in earlier stages. This lets
590 users pinpoint and prioritize structures or reactions of interest. MEANtools can
591 construct DAGs and pathway predictions rooted at any user-selected molecule.

592

593 **Data availability**

594 Raw paired-transcriptomics and -metabolomics data for the case study was taken from
595 NCBI BioProject: PRJNA509154 and EBI's MetaboLights: MTBLS1039 respectively
596 ³². Pre-processed file of the metabolomics data is available at
597 https://github.com/sattely-lab/falcarindiol_pathway_metabolomics. All the input files
598 used in the case study can be found in the 'data' folder in the MEANtools GitHub
599 repository <https://github.com/kumarsaurabh20/meantools>.

600

601 **Code availability**

602 MEANtools is open source and is freely available on its GitHub page
603 (<https://github.com/kumarsaurabh20/meantools>), under the permissive MIT license.
604 The MEANtools documentation and tutorial with the demo data is available on GitHub
605 at <https://meantools.readthedocs.io/en/latest/>.

606

607 **Acknowledgements**

608 This work was funded by the Netherlands Organization for Scientific Research (NWO)
609 under the Groot grant [OCENW.GROOT.2019.063]. We thank Jennifer Wisecaver
610 (Purdue University, USA) for her valuable insights into co-expression analysis utilizing
611 mutual-rank statistics.

612

613 **Competing interests**

614 JJJvdH is currently member of the Scientific Advisory Board of NAICONS Srl., Milano,
615 Italy, and consults for Corteva Agriscience, Indianapolis, IN, USA. M.H.M. is a member
616 of the scientific advisory board of Hexagon Bio. The other authors declare to have no
617 competing interests.

618 **Figure legend**

619

620 **Figure 1:** MEANtools predicts metabolic pathways by integrating transcriptomic,
621 metabolomic, and genomic data. a) Mass signature or mass feature profiles are
622 collected using standard metabolomic data processing pipelines. The feature table
623 has rows as unique features and columns are divided into multiple components, like
624 m/z values, retention times, and mass abundance values across samples. Similarly,
625 the transcript expression matrix is collected using a standard RNA-seq data
626 processing pipeline. In the expression matrix, rows represent different transcripts and
627 columns have normalized count data across samples. b) The RetroRules database is
628 formatted by cross-referencing it with the MetaNetX database for its substrate and
629 related mono-isotopic masses. Based on these masses, mass transition values are
630 calculated for all reactions. c) Feature IDs and m/z values of mass signatures are
631 mapped against a list of user-defined adducts table. By default, MEANtools provides
632 a list of 48 adducts from both positive and negative mode operations. All m/z values
633 are accounted with the adducts masses and PPM value and mapped against the
634 LOTUS database. This mapping results in the putative annotation of each feature ID
635 with specific structures from the LOTUS database. d) Correlations are computed
636 between expression levels of transcripts and abundances of metabolites. e) The
637 protein families/domains encoded by the genes in the correlated transcript-metabolite
638 pairs are used to query RetroRules and identify which enzymatic reactions may be
639 associated with each transcript. f&g) MEANtools then integrate the results of previous
640 steps to identify cases in which metabolite pairs are correlated to a transcript that
641 encodes an enzyme capable of catalyzing a reaction that explains their mutual mass
642 difference. Finally, MEANtools maps the product of these reactions to other mass
643 signatures in the metabolome and repeats the procedure to generate pathway
644 predictions.

645

646 **Figure 2:** Identification of the Functional Cluster (FC) belonging to the falcarindiol
647 pathway. a) Network diagram illustrating the connections between transcripts and
648 metabolites within the falcarindiol FC, with pathway-related transcripts marked with an
649 asterisk. b) Heatmap displaying the expression levels of all genes within the
650 falcarindiol FC. c) Heatmap showing the abundance of mass-signatures associated

651 with the falcarindiol FC. d) Summary table presenting the correlations between
652 transcripts and metabolites from the falcarindiol FC

653

654 **Figure 3:** MEANtools reconstructs parts of the falcarindiol pathway as proposed by
655 Jeon et al., and the genes responsible for each enzymatic step. A) MEANtools predicts
656 the second step of falcarindiol biosynthesis in reverse (dehydrocrepenynic acid →
657 crepenynic acid). The transformation is annotated with the reaction rule used in the
658 transformation, diameter of reaction and RetroRules-based reaction IDs, enzyme
659 support, edge support and reaction likelihood. B) MEANtools predicts the third step of
660 falcarindiol biosynthesis in reverse starting from falcarindiol. Each transformation is
661 annotated with a reaction rule associated with that transformation. Additionally, the
662 reaction rule is annotated with the diameter of the reaction and reaction IDs from
663 RetroRules database. Each transformation in the second step of falcarindiol
664 biosynthesis is also annotated with enzyme support, edge support based on
665 correlation values and the reaction likelihood scores.

666

667 **Figure 4:** Detection of functional clusters (FCs) specific to the phenylalanine (PAL)
668 and p-coumaroyltyramine (THT) pathways. a) Network depicting the relationship
669 between transcripts and mass signatures within the PAL FC. b) Network illustrating
670 the interplay between transcripts and mass signatures within the THT FC. c) Heatmap
671 illustrating the expression levels of all transcripts within the PAL and THT FCs. d)
672 Heatmap displaying the abundance of all mass signatures present in the PAL and THT
673 FCs. e) Correlation matrix highlighting the correlations among transcripts and mass
674 signatures within the PAL FC. f) Correlation matrix displaying the relationships
675 between transcripts and mass signatures within the THT FC, including Mutual rank
676 and transformed edge weights.

677

678 **Figure 5:** Overview of the estimation of reaction likelihood scores. A) The
679 transformation of naringenin to 2-hydroxynaringenin requires a flavanone 2-
680 hydroxylase enzyme. B) To estimate the likelihood score of this reaction, the SMILES
681 ID and the enzyme EC number was used as an input to the GNN-SOM method. GNN-
682 SOM predicts likelihood scores of each atom in the molecule for being a site-of-
683 metabolism. As a next step, we take the SMILES ID of the substrate and the product
684 and use the ReactionDecoder tool to identify the reaction centers and possible bond

685 formation/cleavage site(s). Referring to the atom index of the atoms in the reaction
686 center, we select the highest likelihood score. This value represents the reaction
687 likelihood score of a reaction which is 0.94 for the transformation of naringenin to the
688 2-hydroxynaringenin. C) Distribution of reaction likelihood scores from experimentally
689 validated enzyme-substrate pairs (Known) and randomly assigned enzyme-substrates
690 pairs (Random).

691
692 **Figure 6:** MEANtools identifies reactions for a molecular structure according to a
693 divide-and-conquer strategy. For each metabolite, MEANtools first queries the
694 presence of key atoms and then continues to query, in rounds, increasingly complex
695 reactant substructures according to which substructures have already been identified.
696 For example, A) a set of metabolites is first queried for B) nitrogen and carbon atoms.
697 C) Metabolites that pass these criteria are then queried for more complex
698 substructures like C-N or C=C. D) In the following round, MEANtools queries
699 substructures with more complexity according to which substructures have already
700 been identified: in this manner, only metabolites with the N=C substructure is queried
701 for the N=C-N substructure.

702
703 Supplementary Figure 1: Venn diagram of the content of three datasets, namely loose,
704 medium and strict. *Loose* dataset contains reaction rules-enzyme associations from
705 the RetroRules database, cross-referenced with the Rhea and KEGG-orthology
706 database. *Medium* dataset contains experimentally validated entries together with the
707 ECDomainMiner predictions. The *strict* dataset contains only experimentally validated
708 entries.

709
710 Supplementary Figure 2: Distribution of taxonomic groups in the reaction-enzyme
711 loose dataset. X- and Y-axis represent categories of the taxonomic group and their
712 counts respectively.

713
714 Supplementary Figure 3: Distribution of p-values from the correlation analysis between
715 the processed transcriptomics and metabolomics datasets from Jeon *et al.*, 2020.

716

717 Supplementary Figure 4: Distribution of correlation coefficients from the correlation
718 analysis between the processed transcriptomics and metabolomics datasets from
719 Jeon *et al.*, 2020³².

720

721 Supplementary Figure 5: Functional clusters (FC) encompassing genes from the
722 alpha-tomatine pathway of *Solanum lycopersicum*.

723

724 Supplementary Figure 6: Prediction of intermediate steps of faltarindiol pathway from
725 MEANtools using *strict* (A) and *loose* (B) datasets. Number of enzyme associations is
726 reduced while using *strict* dataset due to its more specific Pfam annotations. Blue
727 arrow shows non-specific enzyme associations predicted with *loose* dataset.

728 Supplementary Figure 7: Distribution of FC node size across decay rates and
729 treatments. For each combination of treatment from the Jeon *et al.* (2020) dataset³²,
730 fungal effectors, bacterial effectors, and all, the size of functional clusters (FCs)
731 generated by MEANtools is calculated and shown in a boxplot across four decay rates.
732 The FCs size is stable across treatment combinations, with a slight tendency for
733 smaller FCs when all treatments are considered. FCs size increases with decay rate.

734

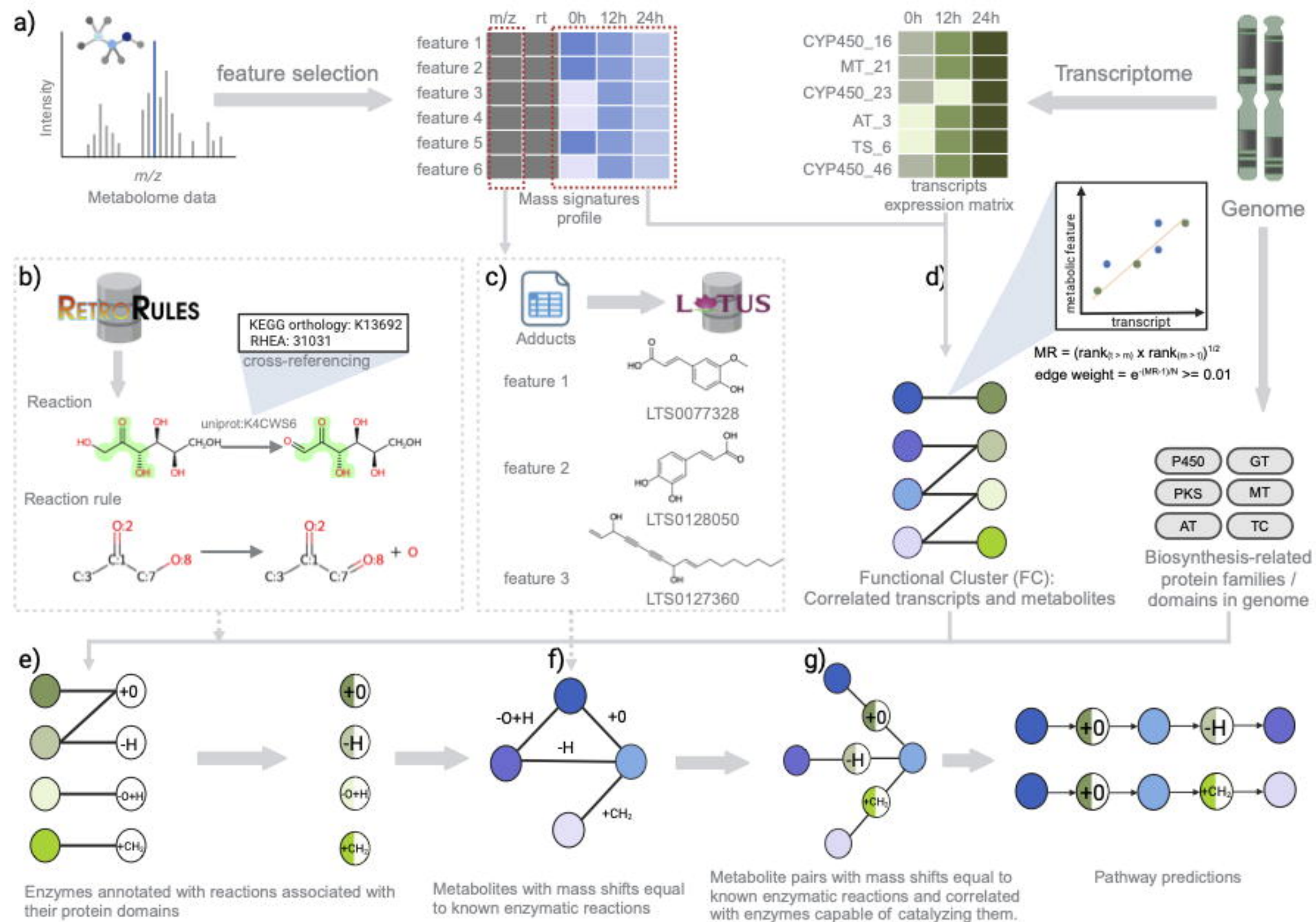
735 Supplementary Figure 8: Use of coexpression edges to merge and prioritize FCs.
736 Green edges represent coexpression networks and blue edges represent gene-
737 metabolite networks of FCs. Colored nodes represent biosynthetic genes annotated
738 from seven tomato pathways. Coexpression was detected across all treatment
739 dimensions. Coexpression networks and FCs were created with a mutual rank metric
740 and ClusterONE clustering with a decay rate of 10. FCs from the α -tomatine pathway
741 are connected thanks to coexpression edges between the genes GAME12
742 transaminase (Solyc12g006470), and 2-oxoglutarate-dependent dioxygenase
743 GAME11 (Solyc07g043420), and GAME17 (UDP-glucosyltransferase)
744 (Solyc07g043480) and GAME1 (UDP-galactosyltransferase) (Solyc07g043490). Two
745 genes associated with the biosynthesis of 4-coumarate CoA ligase (4CL) were also
746 connected by coexpression edges in the hydroxy cinnamic acid amide (HCAA)
747 pathway. Merging FCs via coexpression edges between biosynthetic genes was
748 robust across decay rates 10 and 25, with only the connections belonging to the HCAA
749 pathways displayed in decay rate 5.

750 References

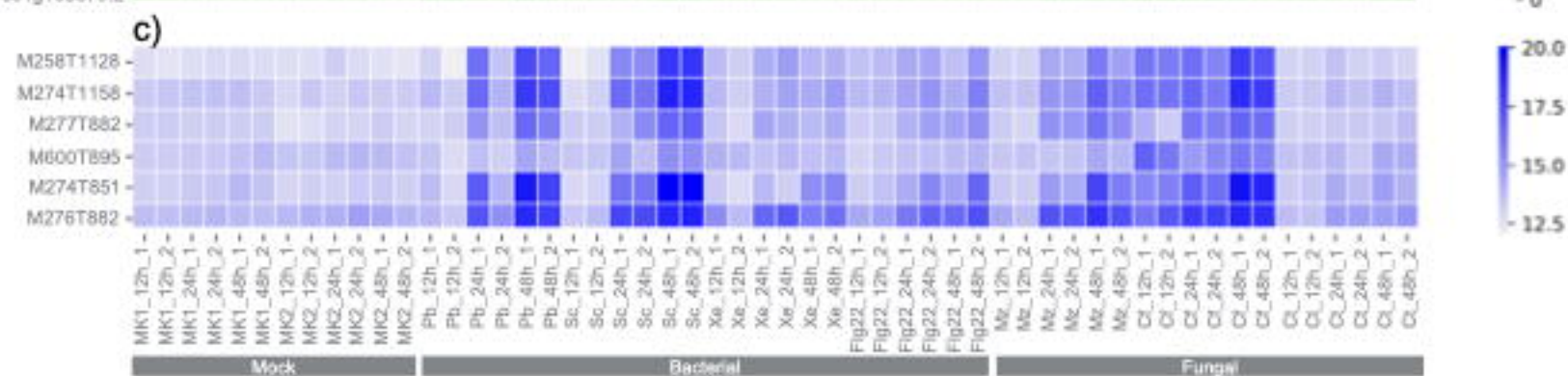
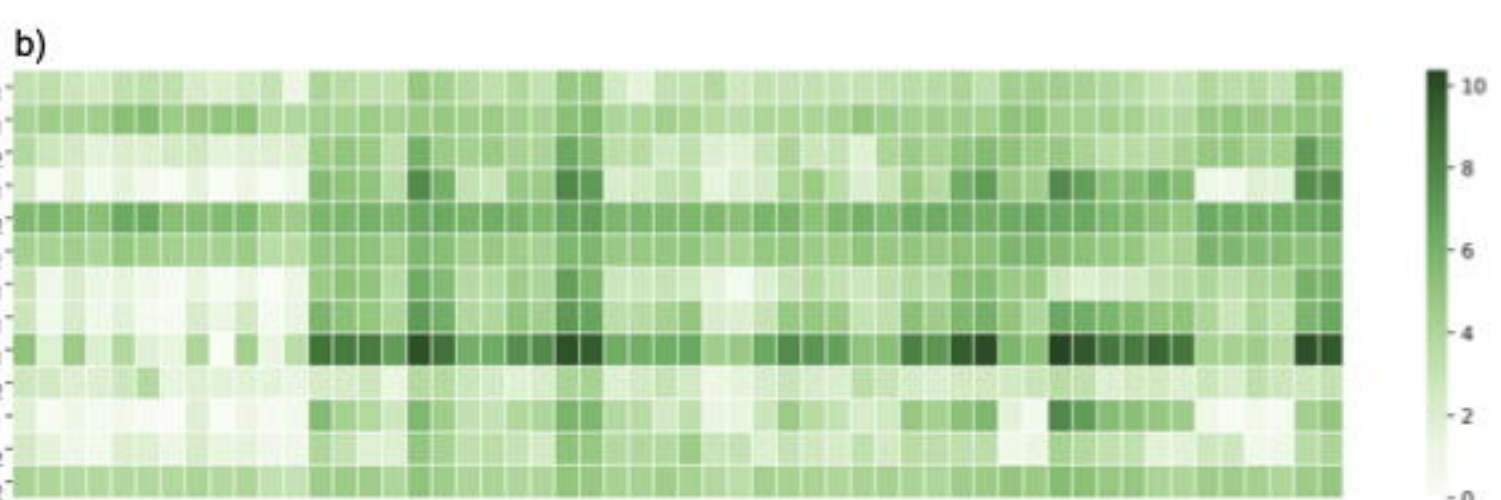
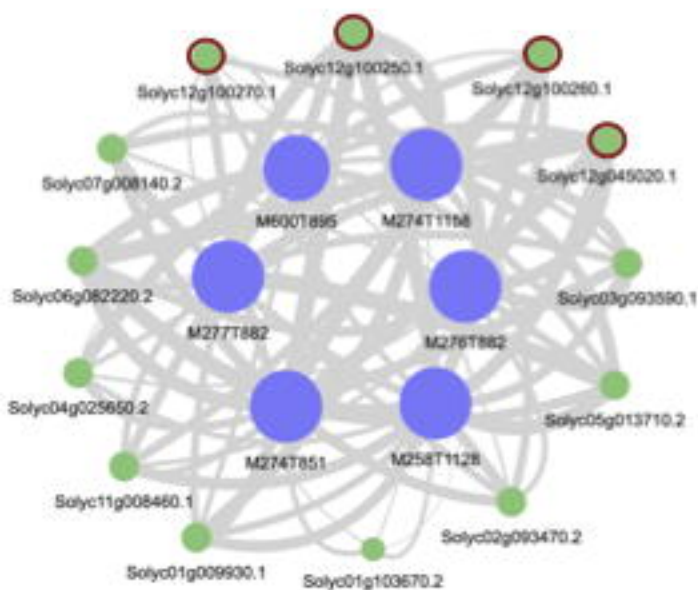
- 751 1. Osbourn, A. E. & Lanzotti, V. Plant-derived natural products: Synthesis,
752 function, and application. *Plant-derived Natural Products: Synthesis, Function,*
753 *and Application* 1–597 (2009) doi:10.1007/978-0-387-85498-4/COVER.
- 754 2. Da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in
755 metabolomics. *Proc Natl Acad Sci U S A* **112**, 12549–12550 (2015).
- 756 3. Aharoni, A., Jongsma, M. A. & Bouwmeester, H. J. Volatile science? Metabolic
757 engineering of terpenoids in plants. *Trends Plant Sci* **10**, 594–602 (2005).
- 758 4. Shen, S. *et al.* An *Oryza*-specific hydroxycinnamoyl tyramine gene cluster
759 contributes to enhanced disease resistance. *Sci Bull (Beijing)* **66**, 2369–2380
760 (2021).
- 761 5. Huang, A. C. *et al.* A specialized metabolic network selectively modulates
762 *Arabidopsis* root microbiota. *Science (1979)* **364**, (2019).
- 763 6. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of
764 specialized metabolism in plants. *Science (1979)* **344**, 510–513 (2014).
- 765 7. Erb, M. & Kliebenstein, D. J. Plant Secondary Metabolites as Defenses,
766 Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. *Plant*
767 *Physiol* **184**, 39–52 (2020).
- 768 8. Medema, M. H. & Osbourn, A. Computational genomic identification and
769 functional reconstitution of plant natural product biosynthetic pathways. *Nat*
770 *Prod Rep* **33**, 951–962 (2016).
- 771 9. Singh, K. S., van der Hooft, J. J. J., van Wees, S. C. M. & Medema, M. H.
772 Integrative omics approaches for biosynthetic pathway discovery in plants. *Nat*
773 *Prod Rep* **39**, 1876–1896 (2022).
- 774 10. Cravens, A., Payne, J. & Smolke, C. D. Synthetic biology strategies for
775 microbial biosynthesis of plant natural products. *Nature Communications* 2019
776 *10:1* **10**, 1–12 (2019).
- 777 11. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H.
778 plantiSMASH: automated identification, annotation and expression analysis of
779 plant biosynthetic gene clusters. *Nucleic Acids Res* **45**, W55–W63 (2017).
- 780 12. Töpfer, N., Fuchs, L. M. & Aharoni, A. The PhytoClust tool for metabolic gene
781 clusters discovery in plant genomes. *Nucleic Acids Res* **45**, 7049–7063 (2017).
- 782 13. Schläpfer, P. *et al.* Genome-Wide Prediction of Metabolic Enzymes, Pathways,
783 and Gene Clusters in Plants. *Plant Physiol* **173**, 2041–2059 (2017).
- 784 14. Li, C. *et al.* Single-cell multi-omics in the medicinal plant *Catharanthus roseus*.
785 *Nature Chemical Biology* 2023 *19:8* **19**, 1031–1041 (2023).
- 786 15. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for
787 Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**,
788 944–959 (2017).
- 789 16. Qi, X. *et al.* A gene cluster for secondary metabolism in oat: Implications for
790 the evolution of metabolic diversity in plants. *Proc Natl Acad Sci U S A* **101**,
791 8233–8238 (2004).
- 792 17. Field, B. & Osbourn, A. E. Metabolic diversification - Independent assembly of
793 operon-like gene clusters in different plants. *Science (1979)* **320**, 543–547
794 (2008).
- 795 18. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic
796 chromosomal regions. *Proc Natl Acad Sci U S A* **108**, 16116–16121 (2011).
- 797 19. Winzer, T. *et al.* A *Papaver somniferum* 10-gene cluster for synthesis of the
798 anticancer alkaloid noscapine. *Science* **336**, 1704–1708 (2012).

- 799 20. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is
800 mediated by clustered genes. *Science* **341**, 175–179 (2013).
- 801 21. King, A. J., Brown, G. D., Gilday, A. D., Larson, T. R. & Graham, I. A.
802 Production of bioactive diterpenoids in the euphorbiaceae depends on
803 evolutionarily conserved gene clusters. *Plant Cell* **26**, 3286–3298 (2014).
- 804 22. Shang, Y. *et al.* Biosynthesis, regulation, and domestication of bitterness in
805 cucumber. *Science* (1979) **346**, 1084–1088 (2014).
- 806 23. Huang, A. C. *et al.* A specialized metabolic network selectively modulates
807 *Arabidopsis* root microbiota. *Science* **364**, (2019).
- 808 24. Chen, X. *et al.* A pathogenesis-related 10 protein catalyzes the final step in
809 thebaine biosynthesis. *Nat Chem Biol* **14**, 738–743 (2018).
- 810 25. Shang, Y. *et al.* Biosynthesis, regulation, and domestication of bitterness in
811 cucumber. *Science* (1979) **346**, 1084–1088 (2014).
- 812 26. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the
813 biosynthetic pathway to the etoposide aglycone. *Science* **349**, 1224–1228
814 (2015).
- 815 27. Rajniak, J., Barco, B., Clay, N. K. & Sattely, E. S. A new cyanogenic
816 metabolite in *Arabidopsis* required for inducible pathogen defence. *Nature* **525**,
817 376–379 (2015).
- 818 28. Aoki, K., Ogata, Y. & Shibata, D. Approaches for Extracting Practical
819 Information from Gene Co-expression Networks in Plant Biology. *Plant Cell*
820 *Physiol* **48**, 381–390 (2007).
- 821 29. Nett, R. S., Dho, Y., Low, Y. Y. & Sattely, E. S. A metabolic regulon reveals
822 early and late acting enzymes in neuroactive *Lycopodium* alkaloid
823 biosynthesis. *Proc Natl Acad Sci U S A* **118**, (2021).
- 824 30. Fang, H. *et al.* A monocot-specific hydroxycinnamoylputrescine gene cluster
825 contributes to immunity and cell death in rice. *Sci Bull (Beijing)* **66**, 2381–2393
826 (2021).
- 827 31. Liu, Z. *et al.* Drivers of metabolic diversification: how dynamic genomic
828 neighbourhoods generate new biosynthetic pathways in the Brassicaceae.
829 *New Phytologist* **227**, 1109–1123 (2020).
- 830 32. Jeon, J. E. *et al.* A Pathogen-Responsive Gene Cluster for Highly Modified
831 Fatty Acids in Tomato. *Cell* **180**, 176-187.e19 (2020).
- 832 33. Hong, B. *et al.* Biosynthesis of strychnine. *Nature* 2022 607:7919 **607**, 617–
833 622 (2022).
- 834 34. Ding, Y. *et al.* Genetic elucidation of interconnected antibiotic pathways
835 mediating maize innate immunity. *Nat Plants* **6**, 1375–1388 (2020).
- 836 35. Nett, R. S., Lau, W. & Sattely, E. S. Discovery and engineering of colchicine
837 alkaloid biosynthesis. *Nature* **584**, 148–153 (2020).
- 838 36. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J. Transcriptomic and
839 metabolomic data integration. *Brief Bioinform* **17**, 891–901 (2016).
- 840 37. Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized
841 metabolism in medicinal plants. *Plant J* **90**, 764–787 (2017).
- 842 38. Duigou, T., Du Lac, M., Carbonell, P. & Faulon, J. L. RetroRules: a database
843 of reaction rules for engineering biology. *Nucleic Acids Res* **47**, D1229–D1235
844 (2019).
- 845 39. Rutz, A. *et al.* The LOTUS initiative for open knowledge management in
846 natural products research. *Elife* **11**, (2022).
- 847 40. Beauxis, Y. & Genta-Jouve, G. MetWork: a web server for natural products
848 anticipation. *Bioinformatics* **35**, 1795–1796 (2019).

- 849 41. Moretti, S., Tran, V. D. T., Mehl, F., Ibberson, M. & Pagni, M.
850 MetaNetX/MNXref: unified namespace for metabolites and biochemical
851 reactions in the context of metabolic models. *Nucleic Acids Res* **49**, D570–
852 D574 (2021).
- 853 42. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res*
854 **50**, D693 (2021).
- 855 43. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes.
856 *Nucleic Acids Res* **28**, 27 (2000).
- 857 44. Von Roepenack-Lahaye, E. *et al.* p-Coumaroylnoradrenaline, a Novel Plant
858 Metabolite Implicated in Tomato Defense against Pathogens. *Journal of*
859 *Biological Chemistry* **278**, 43373–43383 (2003).
- 860 45. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is
861 mediated by clustered genes. *Science (1979)* **341**, 175–179 (2013).
- 862 46. Niggeweg, R., Michael, A. J. & Martin, C. Engineering plants with increased
863 levels of the antioxidant chlorogenic acid. *Nat Biotechnol* **22**, 746–754 (2004).
- 864 47. Porokhin, V., Liu, L. P. & Hassoun, S. Using graph neural networks for site-of-
865 metabolism prediction and its applications to ranking promiscuous enzymatic
866 products. *Bioinformatics* **39**, (2023).
- 867 48. Rahman, S. A. *et al.* Reaction Decoder Tool (RDT): extracting features from
868 chemical reactions. *Bioinformatics* **32**, 2065–2066 (2016).
- 869 49. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for
870 Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**,
871 944–959 (2017).
- 872 50. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes
873 in protein-protein interaction networks. *Nature Methods* **2012 9:5 9**, 471–472
874 (2012).
- 875 51. Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J. & Imberty, A. Structures and
876 mechanisms of glycosyltransferases. *Glycobiology* **16**, 29R-37R (2006).
- 877 52. Beniddir, M. A. *et al.* Advances in decomposing complex metabolite mixtures
878 using substructure- and network-based computational metabolomics
879 approaches. *Nat Prod Rep* **38**, 1967–1993 (2021).
- 880 53. Van Der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers,
881 S. Topic modeling for untargeted substructure exploration in metabolomics.
882 *Proc Natl Acad Sci U S A* **113**, 13738–13743 (2016).
- 883 54. Obayashi, T. & Kinoshita, K. Rank of Correlation Coefficient as a Comparable
884 Measure for Biological Significance of Gene Coexpression. *DNA Research* **16**,
885 249–260 (2009).
- 886 55. RDKit. <https://www.rdkit.org/>.
- 887 56. Hagberg, A., Swart, P. J. & Schult, D. A. Exploring network structure,
888 dynamics, and function using NetworkX. Preprint at (2008).
- 889
- 890



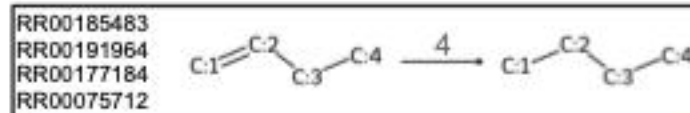
a) **Falcarindiol gene-metabolite network**



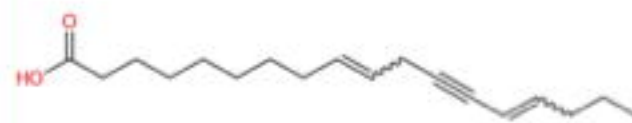
d)



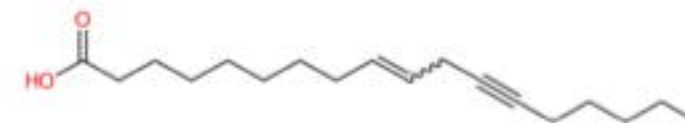
A



enzyme support = 1 | 5
edge support = 0.58 | 0.88
reaction score = 0.931



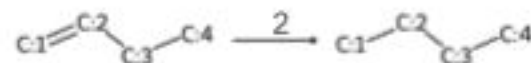
dehydro-crepenynic acid



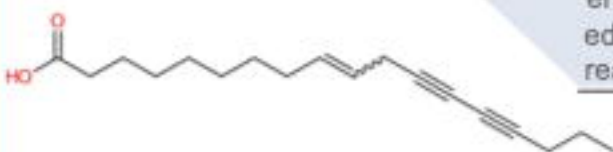
crepenynic acid

B

RR00185483

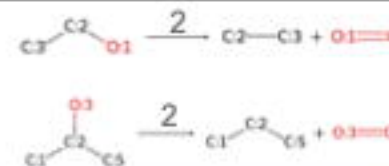


enzyme support = 1 | 0
edge support = 0.54 | 0
reaction score = 0.066

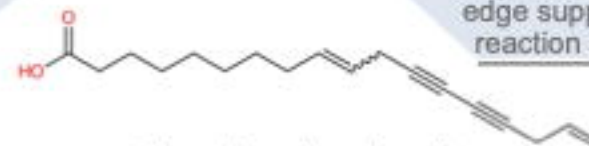


octadecene diyonic acid

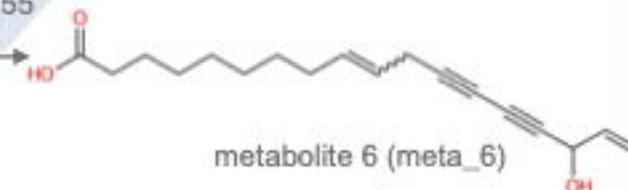
RR00055491
RR00104747
RR00077583
RR00109022
RR00167932
RR00062665
RR00125032
RR00062582



enzyme support = 28 | 15
edge support = 0.62 | 0.55
reaction score = 0.243

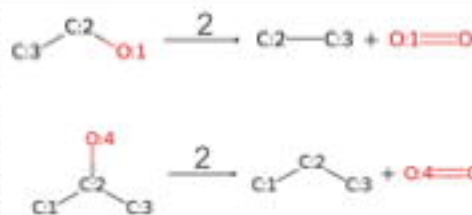


octadecadiene diyonic acid

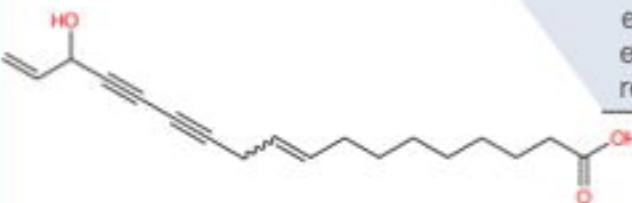


metabolite 6 (meta_6)

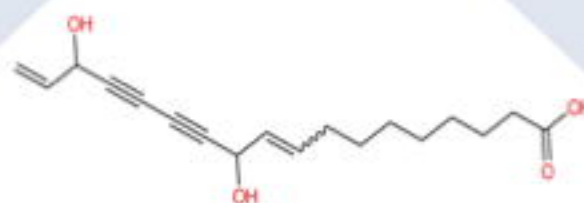
RR00136289
RR00125032
RR00196550
RR00131541
RR00062582
RR00141474
RR00060789
RR00055491
RR00109022
RR00104747



enzyme support = 8 | 17
edge support = 0.62 | 0.55
reaction score = 0.355

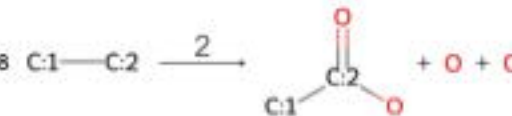


metabolite 6 (meta_6)

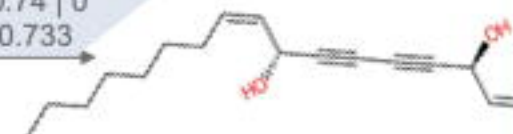


metabolite 7 (meta_7)

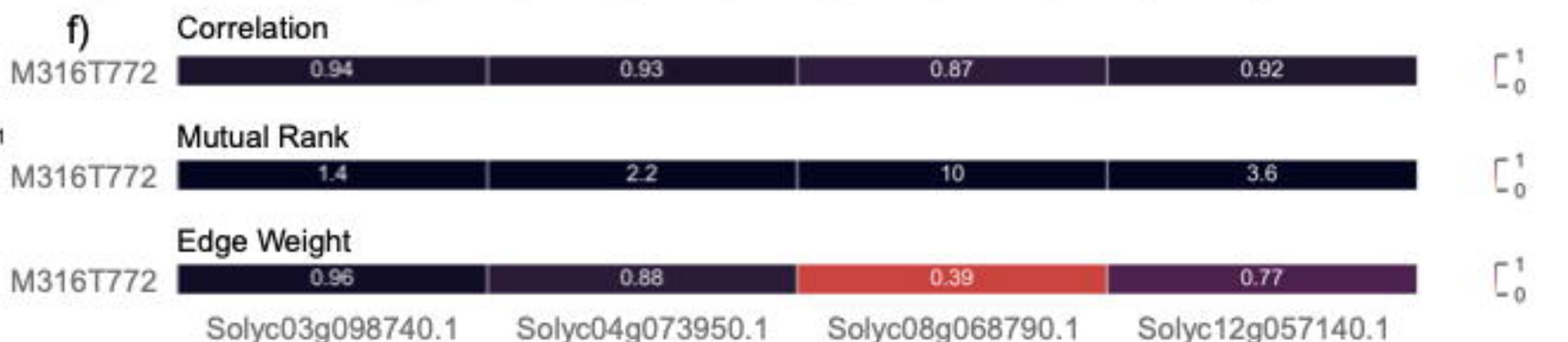
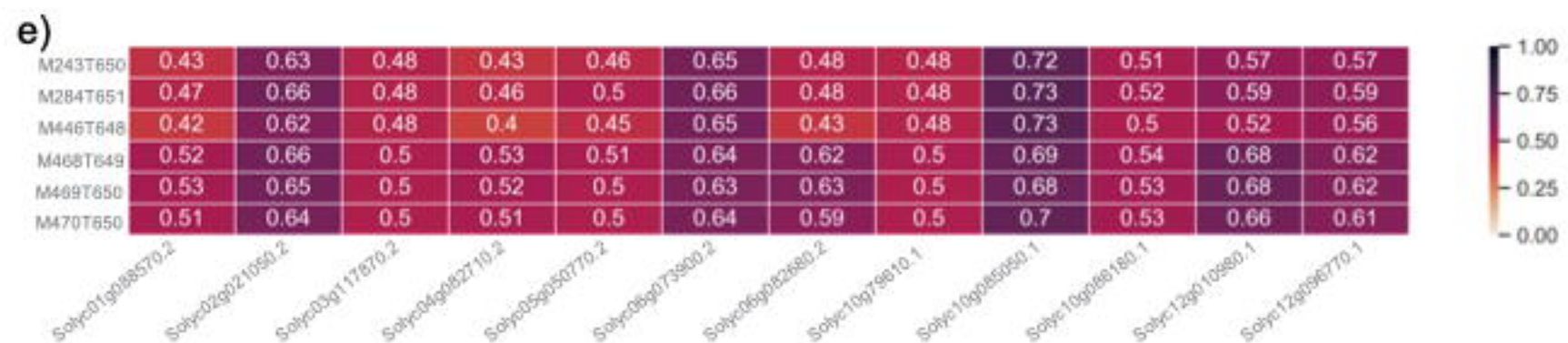
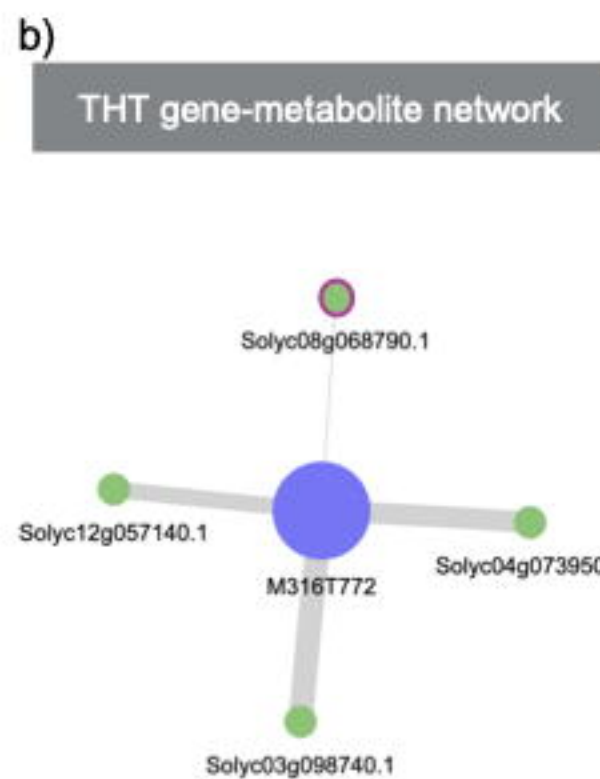
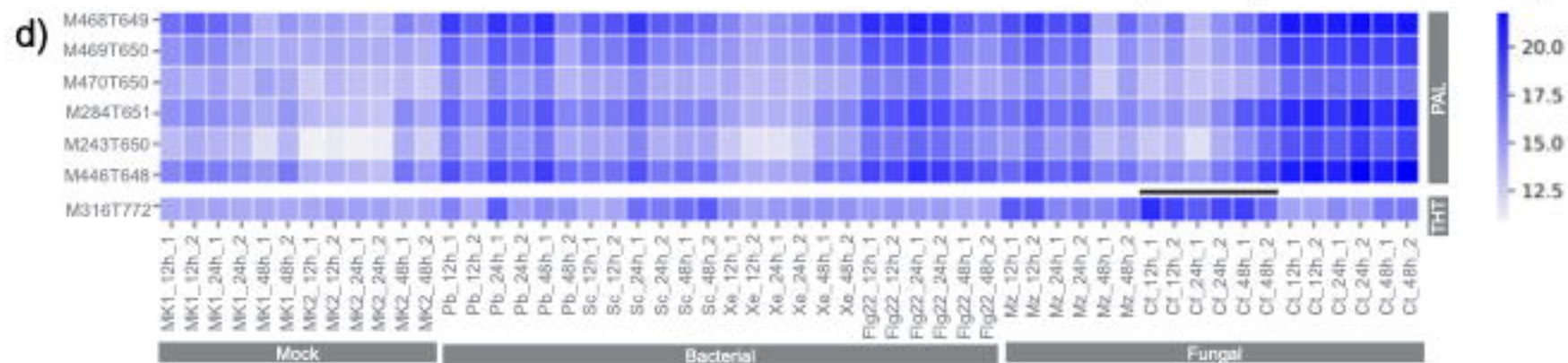
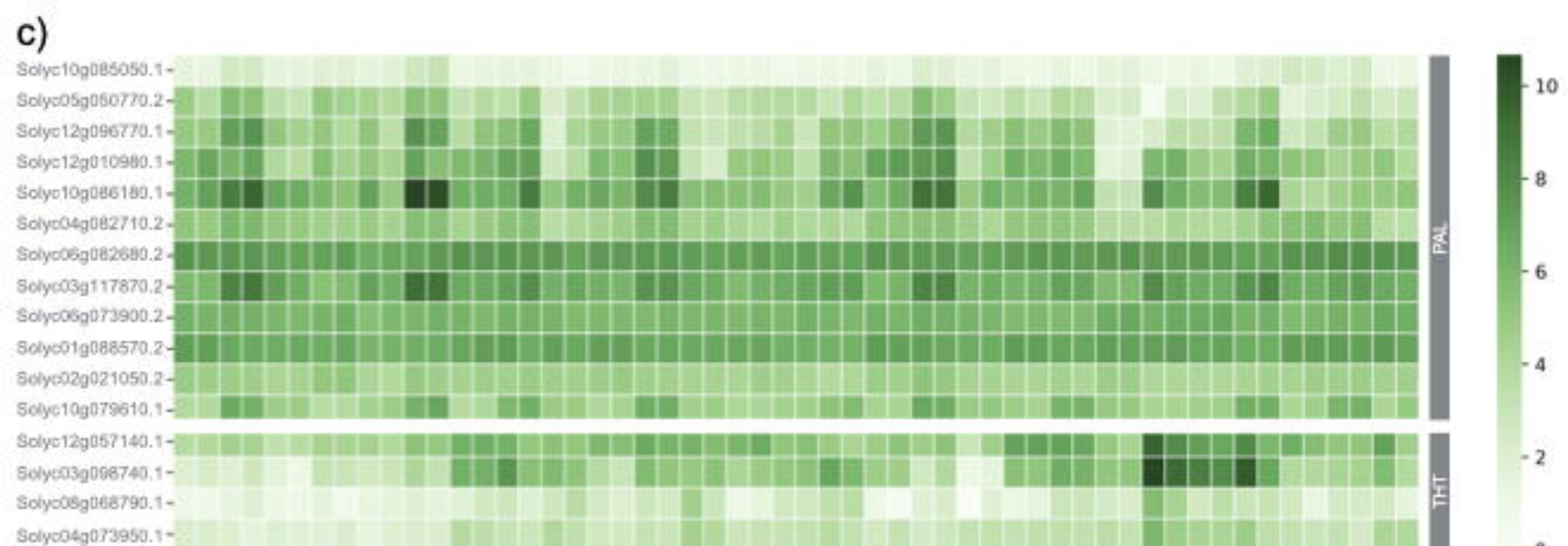
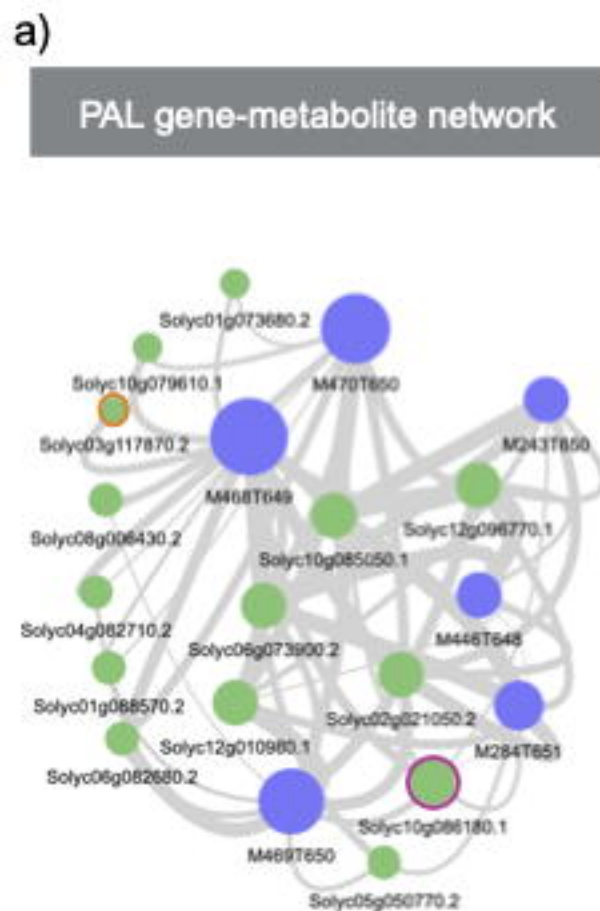
RR00149958

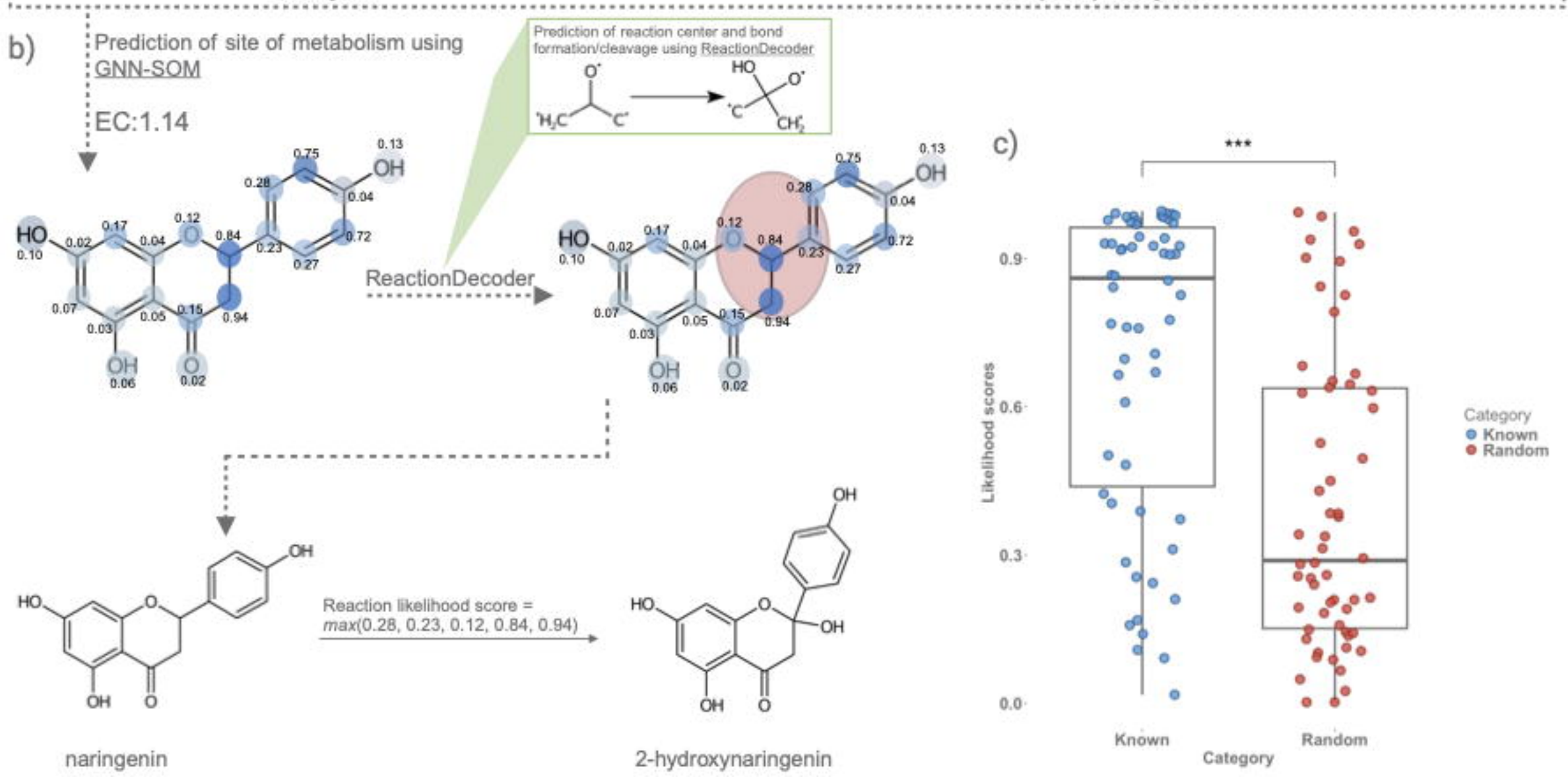
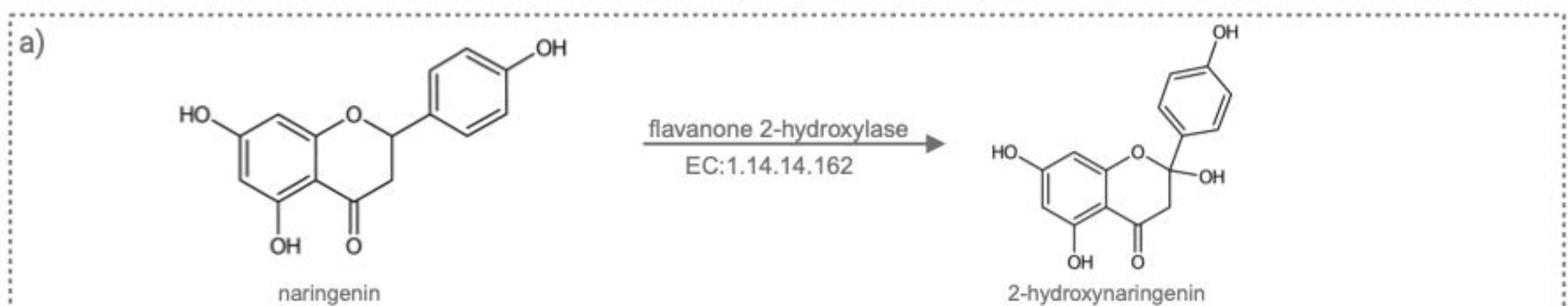


enzyme support = 1 | 0
edge support = 0.74 | 0
reaction score = 0.733



falcarindiol

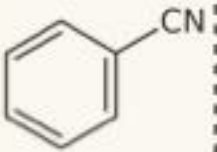
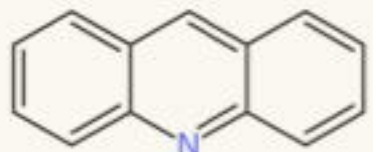
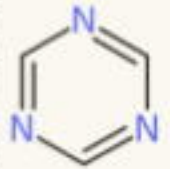
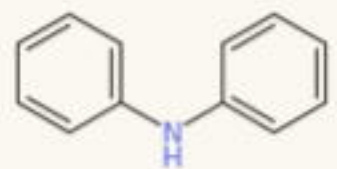
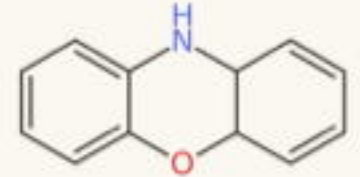





✓ Substructure in metabolite

✗ Substructure not in metabolite

○ Substructure untested

Input metabolites						
Reactant substructures						
N	✓	✓	✓	✓	✓	✓
C	✓	✓	✓	✓	✓	✓
C-C	✓	✓	✓	✓	✓	✓
C=C	✓	✓	✓	✓	✓	✓
N=N	✗	✗	✗	✗	✗	✗
N-C	✓	✓	✓	✓	✓	✓
N=C	✗	✓	✓	✗	✗	✗
C-N-C	✗	✗	✗	✓	✓	✓
C=N-C	○	✓	✓	○	○	○
N-C-N	✗	✗	✗	✗	✗	✗
N=C-N	○	✗	✓	○	○	○