

Document downloaded from:

<http://hdl.handle.net/10251/83646>

This paper must be cited as:

Folch Fortuny, A.; Marques, R.; Isidro, IA.; Oliveira, R.; Ferrer, A. (2016). Principal elementary mode analysis (PEMA). *Molecular BioSystems*. 12(3):737-746.
doi:10.1039/c5mb00828j.



The final publication is available at

<http://doi.org/10.1039/c5mb00828j>

Copyright Royal Society of Chemistry

Additional Information

Principal elementary modes analysis (PEMA)

Abel Folch-Fortuny^{1,*}, Rodolfo Marques², Inês A. Isidro², Rui Oliveira² and Alberto Ferrer¹

¹Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, 46022 València, Spain.

²REQUIMTE/CQFB, Chemistry Department, FCT/ Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Abstract

Principal component analysis (PCA) has been widely applied in fluxomics to compress data into a few latent structures in order to simplify the identification of metabolic patterns. These latent structures lack a direct biological interpretation due to the intrinsic constraints associated to a PCA model. Here we introduce a new method that significantly improves the interpretability of the principal components with a direct link to metabolic pathways. This method, called Principal elementary modes analysis (PEMA), establishes a bridge between a PCA-like model, aimed at explaining the maximum variance in flux data, and the set of elementary modes (EMs) of a metabolic network. It provides an easy way to identify metabolic patterns in large fluxomics data sets in terms of the simplest pathways of the organism metabolism. The results using a real metabolic model of *Escherichia coli* show the ability of PEMA to identify the EMs that generated the different simulated flux distributions. Actual flux data of *E. coli* and *Pichia pastoris* cultures confirm the results observed in the simulated study, providing a biologically meaningful model to explain flux data of both organisms in terms of the EMs activation. The PEMA toolbox is freely available for non-commercial purposes on <http://mseg.webs.upv.es>.

* E-mail address: abfolfor@upv.es

25 1 Introduction

26 Principal component analysis (PCA) is one of the most applied statistical methods in
27 Systems Biology. Its ability to compress large amounts of data, combining different kinds
28 of variables, allows distinguishing between biologically relevant information and noise.
29 This information is contained in a set of new variables built by PCA, the so-called principal
30 components (PCs). In the context of fluxomics, PCA has been widely applied¹⁻³ with two
31 main goals: (i) identify which parts of the metabolism retain the main variability in flux
32 data and (ii) relate them to the behaviour of the organism, *e.g.* substrates consumption and
33 protein production. This way, the PCs identify subsets of reactions based on the correlation
34 structure of the flux data.

35 However, in the context of fluxomics PCA has some limitations. It is difficult to drive the
36 PCs into a biologically meaningful solution, since PCA is a hard modelling method. For
37 example, the main active pathways in a metabolic network could not be orthogonal, so PCA
38 would be unable to describe them accurately in their PCs. To overcome these problems
39 Multivariate Curve Resolution - Alternating Least Squares algorithm⁴ (MCR-ALS) has
40 been proposed to improve the biological interpretation of the components⁵. This method
41 permits the incorporation of constraints, such as non-negativity and selectivity, when
42 building the components. Finally, as with PCA, different sets of reactions or pathways
43 emerge as the driving forces guiding the fluxes in the metabolic network.

44 Here we propose a new method to improve the interpretability of the components extracted
45 by PCA and MCR-ALS, using the topology of the network to obtain the biologically
46 relevant pathways in the model. This method is called Principal elementary modes analysis
47 (PEMA). Its main advantage, over the previous methods, is that instead of building
48 artificial components based solely on the correlation structure of the data (and some *a*
49 *priori* knowledge in the case of MCR-ALS), the components are selected from the
50 complete set of elementary modes (EMs) of the metabolic network. The EMs are the
51 simplest representations of pathways across a metabolic network. Basically, each EM
52 connects substrates with end-products concatenating reactions in a thermodynamically
53 feasible way. The EMs analysis of a metabolic network allows extracting meaningful

54 information of a fluxome data set, since a given metabolic state can be represented as a
55 linear combination of a specific subset of EMs. The PEMA algorithm is designed to
56 identify the most relevant set of active EMs in flux data, using a strategy akin to PCA in
57 dimensionality reduction.

58 Some methods have been proposed in the literature to select a set of representative or active
59 EMs. One such attempt is the concept of the α -spectrum⁶, which involves a linear
60 optimization to determine how the extreme pathways (a systemically independent subset of
61 EMs) contribute to a given steady-state flux distribution. This algorithm allows the
62 determination of maximum and minimum possible weightings for each extreme pathway. A
63 different approach involves the quadratic decomposition of a single steady-state flux into a
64 set of EMs⁷. In this algorithm, a particular set of EMs is chosen, based on the minimization
65 of the weighting vector length. A reinterpretation of this methodology was also proposed
66 by projecting the flux space into the yield space⁸, thus restricting the search for active EMs
67 in a bounded convex space. The PEMA algorithm is quite different from the previous
68 approaches. On the one hand, since PEMA is considering the whole set of EMs, instead of
69 only the extreme pathways, the flux data can be interpreted with fewer pathways. On the
70 other hand, PEMA finds the common set of active EMs in several flux distributions,
71 reducing substantially the number of pathways needed to explain a complete flux data set.

72

73 **2 Methods**

74 *2.1 Principal Component Analysis (PCA)*

75 PCA is a multivariate projection method aimed at finding the underlying patterns of data
76 that represent their main features⁹. The projection is achieved defining new variables, the
77 so-called principal components (PCs), which are built as linear combinations of the original
78 variables, exploiting the correlations among them. The PCA model equation is:

$$\mathbf{X}=\mathbf{T}\cdot\mathbf{P}^T+\mathbf{F} \quad (1)$$

79 where \mathbf{X} is the original data set, \mathbf{T} is the score matrix, containing the new uncorrelated

80 variables (PCs), \mathbf{P} is the loading matrix, which contains the coefficients for the linear
 81 combinations of the original variables, and \mathbf{F} is the error matrix. The number of
 82 components extracted by PCA is usually assessed taking into account the eigenvalues of the
 83 decomposition and the cumulative explained variance of the components¹⁰.

84 *2.2 Elementary modes*

85 The concept of elementary mode (EM) is key for the analysis of metabolic networks. The
 86 set of EMs arises from the stoichiometric matrix, and each EM is defined as a minimal set
 87 of cellular reactions able to operate at steady-state, with each reaction weighted by the
 88 relative flux they need to carry for the mode to function¹¹. The EMs are usually organized
 89 in a data matrix, \mathbf{EM} , having the EMs by columns, the reactions in the metabolic network
 90 by rows, and the relative fluxes in its entries.

91 The set of EMs is obtained from convex analysis¹² and it is unique for a given metabolic
 92 network. Since this set represents a convex basis, any particular steady-state flux
 93 distribution can be obtained as a non-negative linear combination of EMs. Current
 94 algorithms for the computation of EMs face a common problem when dealing with highly
 95 interconnected metabolic networks¹³. In such cases, the combinatorial explosion of the
 96 number of EMs renders the analysis of large networks difficult. Very recently, two new
 97 methods^{14,15} have been proposed to compute the EMs in large networks in a fast and
 98 efficient way.

99 *2.3 Principal elementary modes analysis (PEMA)*

100 PEMA is proposed with the aim of improving the interpretability of the PCA results. This
 101 way PEMA uses the set of EMs as the candidates for the PCs. Let \mathbf{X} be a flux data set with
 102 N observations or experiments and K fluxes. The PEMA model is as follows:

$$\mathbf{X} = \mathbf{\Lambda} \cdot \mathbf{PEM}^T + \mathbf{F} \quad (2)$$

103 where \mathbf{PEM} is the $K \times E$ principal elementary modes matrix, formed by a subset of E EMs
 104 from the entire \mathbf{EM} matrix; $\mathbf{\Lambda}$ is the $N \times E$ weightings matrix; and \mathbf{F} is the $N \times K$ residual
 105 matrix. It is worth noting that the values in $\mathbf{\Lambda}$ are forced to be positive, since from a

106 network-based point of view, each possible steady-state flux distribution can be expressed
 107 as a non-negative combination of EMs¹⁶.

108 In PEMA algorithm, the PEMs are chosen from the complete set of EMs in a step-wise
 109 fashion. The weightings associated to the PEMs are obtained by solving Equation 2:

$$\Lambda = \mathbf{X} \cdot \mathbf{PEM} \cdot (\mathbf{PEM}^T \cdot \mathbf{PEM})^{-1} \quad (3)$$

110 Unlike the loadings in PCA, the PEMs are not orthonormal, so Equation 3 usually requires
 111 the computation of the pseudo-inverse of $\mathbf{PEM}^T \cdot \mathbf{PEM}$.

112 The first step of PEMA consists of calculating the weightings for each EM. So, initially,
 113 \mathbf{PEM} and Λ are column vectors. Then the explained variance by each EM is obtained as
 114 follows¹⁰:

$$EV = 100\% \cdot (\|\mathbf{X}\|^2 - \|\mathbf{F}\|^2) / \|\mathbf{X}\|^2 \quad (4)$$

115 The EMs are sorted by EV , and the EM explaining most variance becomes the first PEM,
 116 with its associated Λ values. Afterwards, the variance explained jointly by the first PEM
 117 and each of the rest of EMs is calculated, and the pairs of EMs are sorted again by EV . The
 118 EM explaining more variance (jointly with the first PEM) becomes the second PEM, with
 119 their corresponding new Λ values. This procedure is iterated until reaching the maximum
 120 number of EMs. Since the weightings are recalculated for the 1st- i th PEMs when the
 121 $(i+1)$ th PEM is computed, the amount of variance explained by the current set of PEMs is
 122 maximum.

123 When the PEMs are extracted step-wise, selecting the EMs explaining most variance at
 124 each step, the greedy solution is obtained. This is the usual procedure in PCA. The loadings
 125 are built in such a way that they explain as much variance in data as possible, and
 126 additionally, the resulting loadings are orthonormal. However, with PEMA, the EMs are
 127 not orthonormal (neither orthogonal). Therefore, the greedy solution may not be the best
 128 subset of EMs for explaining the data, since the choice of the first PEM influences the
 129 variance in data that the following PEMs could explain.

130 Two tuning parameters are introduced in the algorithm to cope with the previous problem.
 131 The greedy selection of the EMs is improved using a relaxation parameter R . This
 132 parameter makes the algorithm considers the best R EMs for the current PEM, and based on
 133 the variance explained extracting more PEMs, the best EM from the set of R is selected.
 134 This relaxation step can be done for several consecutive selections of PEMs. The branch
 135 point number, B , marks up to which PEM the relaxed selection is performed. Figure 1
 136 shows an example of how the tuning parameters affect the selection of EMs. For instance,
 137 with $R=3$ and $B=2$, if one PEM is selected in the PEMA model it will be EM_1 , since it is the
 138 EM explaining most variance; if two PEMs are selected it is possible that EM_1 and any of
 139 its 2nd PEM candidates (EM_6 , EM_{11} , or EM_{19}) explain less variance than, for example,
 140 EM_4 and EM_8 , so these last two will be the EMs selected in the PEMA model with two
 141 PEMs, and so on. The greedy approach accumulates the selected PEMs, but with $R>1$ the
 142 EMs may change completely from one PEM to the next one, in order to explain more
 143 variance with a fixed number of PEMs.

144 The number of PEMs evaluations, *i.e.* the number of times that the algorithm solves
 145 Equation 4 for all EMs, can be calculated using R and B . Let M be the maximum number of
 146 PEMs to be extracted by PEMA. Then, the number of evaluations, A , has the following
 147 expression:

$$A = \sum_{i=1}^M R^i + (M - B) \cdot R^B \quad (5)$$

148 where A grows exponentially with the number of branch points B . This way, the
 149 computation time required for each possible pair (R, B) can be estimated using Equation 5
 150 and the computation time of the greedy approach ($R=B=1$ and $A_{greedy}=M$).

151 PEMA is an heuristic approach to solve the problem *which EMs do reconstruct the flux*
 152 *data?* The mathematical formulation of this problem consists of minimizing the 2-norm of
 153 $\mathbf{X} \cdot \mathbf{A} \cdot \mathbf{PEM}^T$ subject to $\mathbf{PEM} \subseteq \mathbf{EM}$. The problem with this formulation is that it represents a
 154 mixed integer nonlinear programming problem, and since the number of fluxes and EMs
 155 may be extremely high, it is justified the application of an heuristic algorithm to find a

156 suboptimal solution to this problem. The proposed problem could be solved using genetic
157 algorithms, however, different models have to be fit in order to get solutions with different
158 number of PEMs. As well, the solution may change drastically depending on the initial
159 points and the genetic operator chosen. This kind of algorithms improve an objective
160 function, which can be the explained variance as in PEMA, but at some steps of the
161 algorithm the search within the feasible space is performed in a random fashion, while
162 PEMA focuses at each step in selecting the EMs explaining most variance. In this way, a
163 single run of PEMA presents several solutions with a different number of PEMs.

164 *2.4 Data preprocessing*

165 PCA aims at explaining the main variability in data using a few PCs. If the original
166 variables have strongly different means and/or variances, the PCs may focus on explaining
167 only the variables with the highest values and/or variances, disregarding the small variance
168 associated to the rest of variables.

169 PEMA has the same problem as PCA, so the flux data has to be preprocessed. While in
170 PCA it is relatively easy to scale and mean center the original data, in PEMA, since the
171 EMs are fixed, this is a subtle issue. To maintain the biological meaning of the EMs, if \mathbf{X}
172 is scaled column-wise by their standard deviations, the EM matrix has to be modified scaling
173 row-wise all the EMs by the same values. The scaling of the \mathbf{X} and \mathbf{EM} matrices gives,
174 initially, equal importance to all fluxes in the data, since their variances are equal to 1. This
175 preprocessing is always recommended, since the variance of external fluxes can be
176 exponentially greater than internal fluxes.

177 The mean centering of the PEMA model must not be done. When the data matrix \mathbf{X} is mean
178 centered, irreversible reactions would take negative fluxes thus the directionality of the
179 fluxes is lost. In this way, if \mathbf{X} is mean centered the PEMs are no longer able to fit the flux
180 data. One way to overcome the mean centering problem is fitting additional PEMA models
181 excluding the variables with the highest means. Once computed, the global and the local
182 models can be compared in terms of EMs activation and reaction usage, to assess whether
183 the global model is accounting for the fluxes with small values.

184 2.5 PEMA algorithm

185 The PEMA algorithm consists of the following steps:

- 186 1. Scale column-wise the original flux data \mathbf{X} by their standard deviations.
- 187 2. Scale row-wise the elementary modes matrix, \mathbf{EM} , using the standard deviations of the
188 original data set.
- 189 3. Choose the number of relaxations (R) and branch points (B).
- 190 4. Obtain the different PEMA models with 1 PEM, 2 PEMs, ..., M PEMs, solving Equation
191 2.
- 192 5. Select the number of EMs based on the aim of the study.
- 193 6. Recalculate the weightings Λ and the explained variance with the original flux data
194 (without scaling).

195 Practitioners should start with the greedy approach ($R=B=1$) and then, using the prediction
196 of the computation time, select different configurations to compare the models. To span the
197 different solutions that PEMA produces when changing the parameters, users are
198 encouraged to follow the configurations presented in section 3.1 (see also Table 1). For
199 large datasets, *e.g.* genome-scale networks with millions of EMs, the computation of the
200 greedy solution may take several hours. To avoid this long computation time, users can pre-
201 select a subset of relevant EMs prior to applying PEMA.

202 Also, the number of PEMs selected in each model, as in PCA, depends on the aim of the
203 study^{17,18}. In this way, the scree plot (see next section) may help to select the EMs
204 explaining most variance in the flux data.

205

206 3 Results

207 3.1 *Escherichia coli* simulated study

208 A simulated study is proposed here to validate the performance of PEMA. The study
209 consists of simulating different flux data sets, using several subsets of elementary modes
210 (EMs), in order to assess if PEMA algorithm is capable of identifying them. The metabolic
211 model of *Escherichia coli*, presented in reference ¹⁹, is used for this purpose (see Figure 2).
212 The set of reactions can be found *online* in the Supplementary Materials section. The set of
213 255 EMs from the metabolic network of *E. coli* are obtained using EFMTOOL²⁰.

214 The simulated study is as follows: 100 different data sets are generated using from 1 to 10
215 EMs selected at random from the EM matrix. Ten different configurations of PEMA are
216 applied on the present data, varying the values of the relaxations and branches R - B : 1-1, 5-
217 1, 10-1, 20-1, 2-2, 5-2, 10-2, 3-3, 5-3, 4-4. The configurations are sorted in increasing
218 computation time.

219 The identifiability of each PEMA configuration can be assessed computing how many
220 times the complete set of EMs that generated the simulated flux data is identified. This
221 information is presented in Table 1. As expected, for a fixed value of B , the higher is R the
222 better tends to be the solution. Also, the more branch points are considered the more sets of
223 EMs tend to be completely identified.

224 Even though not all the EMs are identified when the number of generating ones increases,
225 all PEMA configurations are able to detect a subset of them. The precision and recall of the
226 EMs identifications are shown in Figure 3. The high precision implies that most of the EMs
227 identified are true ones, and also the high recall implies that the method identified most of
228 the original EMs. With the exception of the greedy approach, all PEMA configurations are
229 able to identify 80-100% of the original 3-4 EMs. The most complex configurations, *i.e.*
230 when $B=3$ or $B=4$, maintain this level of accuracy with 5-6 generating EMs.

231 It is also interesting to check the mean number of PEMs identified by the different
232 configurations and the percentage of explained variance. Since there exists a high degree of
233 redundancy in any EM matrix, different linear combinations of EMs can represent a given
234 flux distribution. This is clearly seen in Figure 4. Up to 5-6 generating EMs, the most
235 complex PEMA configurations identify the same number of PEMs, matching the original
236 ones (see Figure 4a). From 7 generating EMs onwards, the average number of PEMs grows

237 slower, identifying between 7 and 8 PEMs on average with 10 generating EMs. However,
238 the percentage of explained variance by these PEMs remains very high, more than 99%
239 having 7-10 generating EMs (see Figure 4b). The reduction in the number of EMs might
240 also be due to the fact that some of the randomly selected EMs, with a random weighting
241 on the model, have a small contribution to the variance in comparison to the EMs with
242 greater coefficients. A table with the minimum, mean and maximum values for Figure 4a
243 and the standard deviations for Figure 4b) can be found in Supplementary Materials.

244 3.2 *E. coli* real data

245 The flux data of *E. coli* presented in reference ¹⁹ is used in this section to check the
246 performance of PEMA with real data. Each observation in this dataset describes a flux
247 distribution in *E. coli*, after a specifically targeted gene knock-out. The metabolic network
248 and EMs set considered here are the same as in the simulated study (see Figure 2). The flux
249 data matrix, \mathbf{X} considered in this paper has 21 observations (rows) and 42 fluxes (columns).
250 In these 21 observations, a subset of the original 32 observations, the same set of reactions
251 is considered. The flux data set can be found *online*.

252 Based on the results of the simulated study, the tuning parameters R and B are both set to 4,
253 to obtain more accurate results. The computation time of PEMA in this case is 2 minutes
254 (2.9 GHz Intel Core I7, 8GB RAM 1600 MHz), while the computation time of the greedy
255 approach is less than a second. Figure 5a shows the cumulative scree plot of the PEMs.
256 This kind of plot is usually employed in PCA to assess the appropriate number of principal
257 components. Here, 8 PEMs are selected: EM₁₂₅, EM₁₆₇, EM₂₅₄, EM₂₇, EM₂₃₅, EM₁₆,
258 EM₁₄₃ and EM₁₄₅, explaining 97.8% of variance with the scaled data, and 99.4% of the real
259 variance. As opposed to PCA, in PEMA the PEMs are usually explaining common sources
260 of variability. This can be seen in Figure 5b, where the direct sum of all variances explained
261 by the PEMs is 150%. For instance, EM₁₂₅ explains more than 80% of variance in data, but
262 this variance is shared with other PEMs. Nevertheless, the PEMs explaining most variance
263 can be considered the most relevant in the model.

264 The degree of orthogonality of the PEMs can be obtained by dividing the variance
265 explained by the model (99.4%) by the sum of the explained variances of each PEM. Here,

266 the degree of orthogonality is 66.3%, which implies that the solution obtained by the
267 PEMA is strongly non-orthogonal and, therefore, quite different from the PCA one.

268 To assess if some observation is not well modelled the percentage of explained variance per
269 observation can be computed (see Figure 6a). Also the observed versus predicted plot can
270 be used to visualise the differences at a data point level (see Figure 6b). In the present case,
271 the percentage of explained variance is 97-99% for all observations, and the predicted
272 values lay close to the true ones.

273 The PEMA model can be easily interpreted using an adaptation of the classical PCA
274 loadings and scores plot. This way, Figures 7-8 shows the principal elementary modes plot
275 and the weightings plot, respectively. The PEMs plot shows which reactions are active for a
276 specific EM, while the weightings plot represents the contribution weight of each PEM on
277 each observation (*i.e.* knock-out). A first look at the selected PEMs shows that the whole
278 set captures the formation of all metabolic requirements for cell synthesis, that is, reactions
279 31-41 (see Figure 7). EM₁₂₅ is the PEM explaining most variance in data (see Figure 5b).
280 This pathway depicts the glucose flux into glycolysis and TCA, without any exchange
281 fluxes for cell synthesis metabolites. This leads to a high rate of NADH production, which
282 generally is used to synthesize ATP. For this, EM125 can be interpreted as the cell's
283 catabolic pathway, while the remainder PEMs capture the fluxes for cell synthesis
284 metabolites, thus representing anabolic pathways leading to synthesis of biomass.

285 Since EM₁₂₅ is related to the catabolism, it has a strong weight in each knock-out (see
286 Figure 8). Nevertheless, some observations seem to have a greater impact in this PEM than
287 others, in particular the knock-outs 2, 3, 10, 14, 15 and 16, representing the genes *glk*, *pgm*,
288 *gpmB*, *rpiB*, *tktB* and *talB*. The *pgm* gene codifies the phosphoglucomutase that converts
289 G6P into G1P and its deletion would likely direct the carbon flux to glycolysis or the
290 pentose phosphate pathway. The genes *rpiB*, *tktB* and *talB*, also scoring a high weight, are
291 related to pentose phosphate reactions.

292 The EMs related to anabolic metabolism represent all the remaining exchange fluxes that
293 produce the cell synthesis metabolites. EMs 16, 27, 143, 145 and 167 connect glucose
294 directly to the pentose phosphate pathway, which is fundamental in the metabolism, since it

295 generates NADPH, a reduced equivalent important in biosynthetic processes²¹. Moreover,
296 EM₁₆ and EM₁₆₇ are responsible for balancing the metabolic fluxes towards E4P and R5P,
297 being the sole PEMs that predict the fluxes of these metabolites to cell synthesis. With a
298 few exceptions, the knock-out experiments have similar weight values inside each anabolic
299 PEM. These exceptions are the observations 1, 5, 8, 12 and 14, representing the knockouts
300 *galM*, *pfkB*, *gapC*, *pykF* and *rpiB*. This group of genes has low weightings in EM₂₅₄ and
301 EM₂₃₅, meaning that these flux modes have a minor impact in the metabolism of these
302 mutants, that is, a lower flux in the synthesis of Pyr, 3-PG, 2-KG and OAA for biomass
303 synthesis. Conversely higher weightings from these mutants are observed for EM₂₇ and
304 EM₁₆, that is, in the production of E4P, PEP and G6P. Another curious aspect of EM₁₆ and
305 EM₂₇ is the activation of the glyoxylate bypass. This pathway is known to be active in low
306 glucose concentrations²², but repressed when glucose becomes available in higher
307 concentrations^{23,24}. The observations 18 to 21 reflect *E. coli* wild-type cultured at a dilution
308 rate of 0.2 h⁻¹, used as control experiments. In these observations, positive fluxes for the
309 glyoxylate pathway were registered, possibly due to a low glucose feed to the culture.

310 Finally, all the PEMs have a zero coefficient for fermentative pathways (reactions 28-30),
311 therefore these fluxes are not being represented by the model. However, looking at the
312 original data, all the observations have zero values for fluxes 28 and 29. Regarding flux 30,
313 few observations (4 out of 21) have a non-zero value for it. For the latter case, since PEMA,
314 as PCA, aims at explaining the covariance between the original variables using the PEMs,
315 if most of the values in a variable are 0 it is difficult for PEMA to identify the EM
316 generating these slight differences. The extraction of more PEMs may correct that,
317 however, the risk of overfitting is higher and the model would become less parsimonious.

318 3.3 *Pichia pastoris* real data

319 A second real case study is analysed here: a fluxome for the growth of recombinant *P.*
320 *pastoris*. This data set was based on a statistical design of experiments to test the effects of
321 culture media factors in the flux data. The media composition was prepared according to
322 the Invitrogen's guidelines for *P. pastoris* fermentation, and consists mainly on mineral
323 salts. 26 shake flask experiments were performed with variations on 11 media factors

324 selected for statistical design. Glycerol was used as carbon source in every experiment.

325 The metabolic network for the central carbon metabolism of *P. pastoris* used here is largely
326 based on the network proposed in reference²⁵, with adaptations from other central carbon²⁶
327 and genome-scale networks²⁷. The network consists of 43 metabolic reactions, 34 internal
328 metabolites and 10 exchange reactions (see Figure 9). The main catabolic reactions are
329 represented in this network, namely glycolysis and gluconeogenesis pathways, the
330 tricarboxylic acid cycle (TCA), the pentose-phosphate pathway, anaplerotic, fermentative
331 and phosphorylative oxidation pathways. A biomass formation reaction is also included in
332 the model, from selected internal metabolites based on *P. pastoris* cells macromolecular
333 composition²⁸. There exist 158 EMs in the metabolic model. The flux data set and the
334 elementary modes matrix can be found *online* in the Supplementary Material section.

335 The results of PEMA with this data set are the same using either the greedy approach and
336 the most complex approach presented here ($R=B=4$), which takes 35 seconds. This
337 indicates that the results are stable against the different PEMA configurations. 99.5% of the
338 scaled data is explained using 3 PEMs, with a degree of orthogonality of 70% (*i.e.* the
339 variance explained by the 3 PEMs sums 141%). As in the previous real case study, this
340 implies that PCA cannot obtain these results using orthogonal components. The cumulative
341 scree plot and the variance explained by each PEM are shown in Figure 10.

342 All scenarios are being represented by the selected EMs, as can be seen in the explained
343 variance per observation plot (see Figure 11a); and the observed versus predicted plot (see
344 Figure 11b) shows an even better fitting than with *E. coli*, which could be due to different
345 levels of noise in the flux data sets.

346 Figure 12 shows the PEMs and weightings plots. The PEMs identified are EM₁₄₇, EM₁₀
347 and EM₁₄₉. The binary version of the PEM plot appears in the Supplementary Material
348 section. The binary version of the weightings plot is not included, since all observations use
349 all PEMs.

350 The first PEM consumes glycerol (reactions 35 and 29) and crosses half of the glycolytic
351 pathway (reactions 4-7) to activate the TCA cycle (reactions 15, 17-20), clearly

352 representing the cell's catabolism. EM10 uses also reactions 35, 29 and 4-7 to activate the
353 TCA cycle, but in this case reaction 16 is used instead of 17. It also activates the pentose
354 phosphate pathway (reactions 8-13), leading to the synthesis of redox equivalents
355 (NADPH), but also precursor metabolites for the synthesis of biomass. For this reason, this
356 PEM groups the reactions for the cell's anabolism. At the end, this is the PEM responsible
357 of the biomass production in all observations. The last PEM assimilates glycerol in the
358 same way as EM₁₄₇ and afterwards focuses on the production of ethanol (reactions 25 and
359 39). The occurrence of ethanol synthesis during aerobic respiration in yeast is a common
360 feature (Crabtree effect). Nonetheless, unlike most yeasts, *P. pastoris* does not typically
361 exhibit a significant ethanol production, favouring the aerobic metabolism. This fact is well
362 captured by the relative lower explained variance of EM₁₄₉ in comparison to EM₁₄₇ (see
363 Figures 10 and 12b).

364 Finally, as expected, no EM related to methanol assimilation (reactions 30-32 and 26) and
365 final products such as pyruvate or citrate (reactions 41 and 42, respectively) is selected,
366 since all fluxes are 0 for these reactions.

367

368 **4 Discussion**

369 In this paper a new method called principal elementary modes analysis (PEMA) is
370 presented with the aim of improving the interpretability of a traditional PCA modelling in
371 fluxomics. PEMA builds a PCA-like model using the complete set of elementary modes
372 (EMs) in order to identify which ones, the PEMs, are the driving forces generating the flux
373 distributions.

374 The simulated study on *E. coli* shows the high identifiability of PEMA. The most complex
375 PEMA configurations are able to detect completely 1-4 generating EMs and, a high
376 percentage of them, up to 6-7 EMs. Even though not all the EMs are identified by PEMA,
377 the method provides always a parsimonious solution explaining more than 99% of variance.
378 The analysis of actual flux data of the same organism confirms the tendency shown with
379 the simulated fluxes. 8 PEMs are identified explaining 99.4% of variance in the flux data.

380 This way, most of the PEMs identified are describing the glucose consumption, the
381 glycolytic pathway and the TCA cycle, but afterwards, each of them has a different
382 function in the cell synthesis. The results obtained with *P. pastoris* are coherent with *E.*
383 *coli*'s. In this case 3 PEMs are selected describing accurately the metabolic pathways being
384 activated when glycerol is used as main carbon source in aerobic conditions.

385 A significant number of graphical tools, all of them integrated in the PEMA toolbox, are
386 provided in this paper. The cumulative scree plot, the observed versus predicted plot, and
387 the variance explained per observation plot can be used to decide the number of PEMs to
388 extract. The plot showing the variance explained by each PEM and the PEMs and
389 weightings plots are useful to exploit the PEMA model in terms of relevance and biological
390 interpretation of the PEMs, and their activation among the observations.

391 Additionally, the theoretical estimation of the runs of PEMA algorithm when the tuning
392 parameters change permits to establish a relatively accurate upper bound of the
393 computation time, based on the greedy approach solution. This allows designing wisely a
394 set of trials to compare the results of the different configurations of PEMA.

395

396 **5 Conclusion**

397 In this work, PEMA is developed to explain the inherent variability on a fluxomics dataset,
398 while preserving biological meaning. This can be regarded as an exploratory technique that
399 allows researchers to interpret a data set by uncovering the most representative pathways
400 operating in a cell.

401 There is a potential use of this methodology in bioprocess engineering applications, such as
402 the development of structured metabolic models in cell culture fermentations. PEMA can be
403 useful in the identification of a specific set of EMs that explains variations in cellular
404 metabolic rates under certain operational conditions, such as temperature and pH. This
405 would allow the improvement of the process kinetics' modelling by the incorporation of
406 biological knowledge from the cellular system.

407

408 **Author's contributions**

409 AF-F, RM and IAI performed the analyses. AF-F and RM wrote the manuscript. AF-F
410 wrote the code of the PEMA toolbox. RO and AF conceived the study and reviewed the
411 manuscript. All authors read and approved the final manuscript.

412

413 **Acknowledgements**

414 Research in this study was partially supported by the Spanish Ministry of Economy and
415 Competitiveness and FEDER funds from the European Union through grants DPI2011-
416 28112-C04-02 and DPI2014-55276-C5-1R. We would also acknowledge Fundação para a
417 Ciência e Tecnologia for PhD fellowships with references SFRH/BD/67033/2009,
418 SFRH/BD/70768/2010 and PTDC/BBB-BSS/2800/2012.

419

420 **References**

- 421 1. B. Sariyar, S. Perk, U. Akman and A. Hortasu, *Journal of Theoretical Biology*, 2006,
422 **242**(2), 389–400.
- 423 2. C. Barrett, M. Herrgard, and B.O. Palsson, *BMC Systems Biology*, 2009, **3**, 30.
- 424 3. J.M. González-Martínez, A. Folch-Fortuny, F. Llaneras, M. Tortajada, J. Picó and A.
425 Ferrer, *Chemometrics and Intelligent Laboratory Systems*, 2014, **134**, 89–99.
- 426 4. J. Jaumot, R. A. de Juan, and R. Tauler, *Chemometrics and Intelligent Laboratory*
427 *Systems*, 2015, **140**, 1-12.
- 428 5. A. Folch-Fortuny, M. Tortajada, J.M. Prats-Montalbán, F. Llaneras, J. Picó and A.
429 Ferrer, *Chemometrics and Intelligent Laboratory Systems*, 2015, **146**, 293–303.
- 430 6. S.J. Wiback, R. Mahadevan and B.O. Palsson, *Journal of Theoretical Biology*, 2003, **3**,
431 313-324.

- 432 7. J.-M. Schwartz and M. Kanehisa, *Bioinformatics*, 2005, **21**, 204-205.
- 433 8. H.-S. Song and D. Ramkrishna, *Biotechnology and Bioengineering*, 2009, **2**, 554-568.
- 434 9. J.E. Jackson, *A User's Guide to Principal Components*. Wiley Series in Probability and
435 Statistics, Wiley Online Library, 2004.
- 436 10. R. Bro and A.K. Smilde, *Analytical Methods*, 2014, **6**(9), 2812–2831.
- 437 11. S. Schuster, D.A. Fell and T. Dandekar, *Nature Biotechnology*, 2000, **18**(3), 326–332.
- 438 12. S. Schuster, C. Hilgetag, J.H. Woods and D.A. Fell, *Journal of Mathematical Biology*,
439 2002, **45**, 153-181.
- 440 13. S. Klamt, and J. Stelling, *Molecular Biology Reports*, 2002, **29**(1-2), 233–236.
- 441 14. L.-E. Quek and L.K. Nielsen, *BMC Systems Biology*, 2014, **8**(1), 94.
- 442 15. M.B. Badsha, R. Tsuboi and H. Kurata, *Biochemical Engineering Journal*, 2014, **90**,
443 121–130.
- 444 16. F. Llaneras and J. Picó, *Journal of Bioscience and Bioengineering*, 2008, **105**(1), 1–11.
- 445 17. J. Camacho and A. Ferrer, *Journal of Chemometrics*, 2012, **26**(7), 361–373.
- 446 18. J. Camacho and A. Ferrer, *Chemometrics and Intelligent Laboratory Systems*, 2014,
447 **131**, 37–50.
- 448 19. N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M.
449 Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T.
450 Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N.
451 Iwata, Y. Toya, T. Nakayama, T. Nishioka, K. Shimizu, H. Mori and M. Tomita, *Science*,
452 2007, **316**(5824), 593–597.
- 453 20. M. Terzer and J. Stelling, *Bioinformatics*, 2009, **19**, 2229-2235
- 454 21. M. Madigan, J. Martinko and J. Parker, *Brock Biology of Microorganisms*, Pearson
455 Education, Inc., New Jersey, 2003.
- 456 22. A. Nanchen, A. Schicker, O. Revelles and U. Sauer, *Journal of Bacteriology*, 2008,
457 **190**(7), 2323–2330.

- 458 23. A. Nanchen, A. Schicker and U. Sauer, *Applied and Environmental Microbiology*,
459 2006, **72**(2), 1164–1172.
- 460 24. R. Carlson and F. Srienc, *Biotechnology and Bioengineering*, 2004, **85**(1), 1–19.
- 461 25. M. Tortajada, F. Llaneras and J. Pico, *BMC Systems Biology*, 2010, **4**, 115.
- 462 26. K. Baumann, M. Carnicer, M. Dragosits, A.B. Graf, J. Stadlmann, P. Jouhten, H.
463 Maaheimo, B. Gasser, J. Albiol, D. Mattanovich and P. Ferrer, *BMC Systems Biology*,
464 2010, **4**(1), 141.
- 465 27. B.K. Chung, S. Selvarasu, A. Camattari, J. Ryu, H. Lee, J. Ahn, H. Lee and D.-Y. Lee,
466 *Microbial Cell Factories*, 2010, **9**(1), 50.
- 467 28. M. Dragosits, J. Stadlmann, J. Albiol, K. Baumann, M. Maurer, B. Gasser, M. Sauer, F.
468 Altmann, P. Ferrer and D. Mattanovich, *Journal of Proteome Research*, 2009, **8**(3), 1380–
469 1392.
- 470

471 **Tables**

472

473 **Table 1.** Complete identifications of the generating elementary modes.

Configuration R-B	Number of generating elementary modes						
	1	2	3	4	5	6	7-10
1-1	10/10	7/10	2/10	2/10	0/10	0/10	0/10
5-1	10/10	10/10	5/10	3/10	1/10	1/10	0/10
10-1	10/10	10/10	5/10	4/10	1/10	0/10	0/10
20-1	10/10	10/10	5/10	5/10	1/10	0/10	0/10
2-2	10/10	9/10	5/10	4/10	1/10	0/10	0/10
5-2	10/10	10/10	5/10	2/10	1/10	0/10	0/10
10-2	10/10	10/10	7/10	7/10	2/10	1/10	0/10
3-3	10/10	9/10	7/10	6/10	4/10	1/10	0/10
5-3	10/10	10/10	7/10	8/10	5/10	1/10	0/10
4-4	10/10	10/10	7/10	8/10	6/10	3/10	0/10

474

475 **List of Figure captions**

476

477 **Figure 1.** Relaxation (R) and branch point (B) parameters. When $B=R=1$ the EM explaining
478 most variance is chosen and fixed at each step. If these parameters change, different subsets
479 are considered for each PEM identification.

480 **Figure 2.** Metabolic network of *E. coli* considered for the present study.

481 **Figure 3.** Precision and recall of the different configurations. Precision is calculated by
482 dividing the sum of the true identified EMs by the sum of the true identified plus the false
483 identified ones. The recall is calculated by dividing the true identified EMs divided by the
484 true ones plus the true non-identified ones.

485 **Figure 4.** a) Mean number of identified EMs. b) Mean percentage of explained variance.

486 **Figure 5.** a) PEMA Cumulative scree plot and b) Percentage of variance explained by each
487 PEM in *E. coli* study: 8 PEMs are selected explaining 97.4% of variance in the scaled data.

488 **Figure 6.** a) Explained variance per observation and b) Observed versus predicted plot in *E.*
489 *coli* study.

490 **Figure 7.** Principal elementary modes plot in *E. coli* study. The PEMs are represented by
491 columns and the corresponding reactions by rows. Blue squares represent positive values,
492 and dashed red squares the negative ones. The darker the colour, the more highly
493 positive/negative is the value.

494 **Figure 8.** Weightings plot in *E. coli* study. The weightings of the PEMs are represented by
495 columns and the observations by rows. The darker the colour, the more important is the
496 PEM for the corresponding observation.

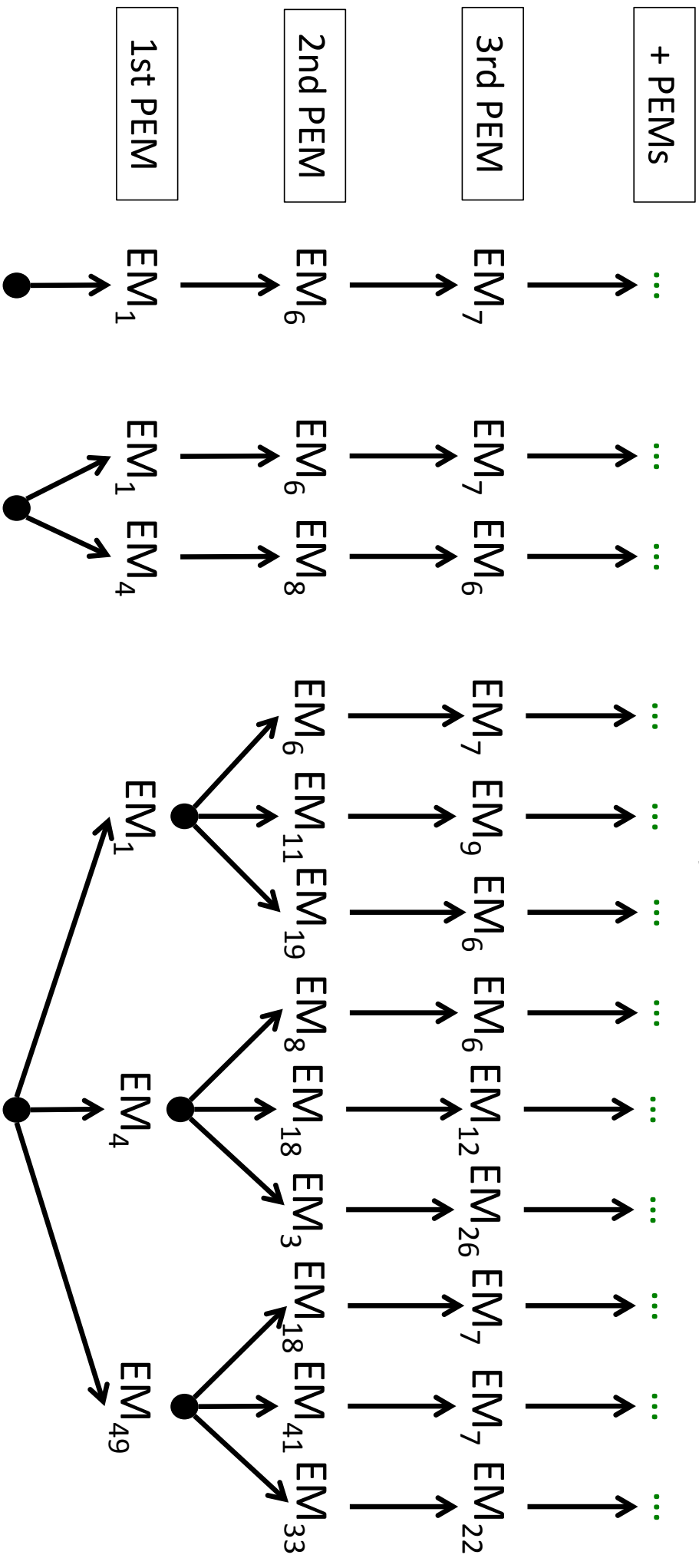
497 **Figure 9.** Metabolic network of *P. pastoris* considered for the real case study.

498 **Figure 10.** a) PEMA Cumulative scree plot and b) Percentage of variance explained by
499 each PEM in *P. pastoris* study: 3 PEMs are selected explaining 99.5% of variance in the

500 scaled data.

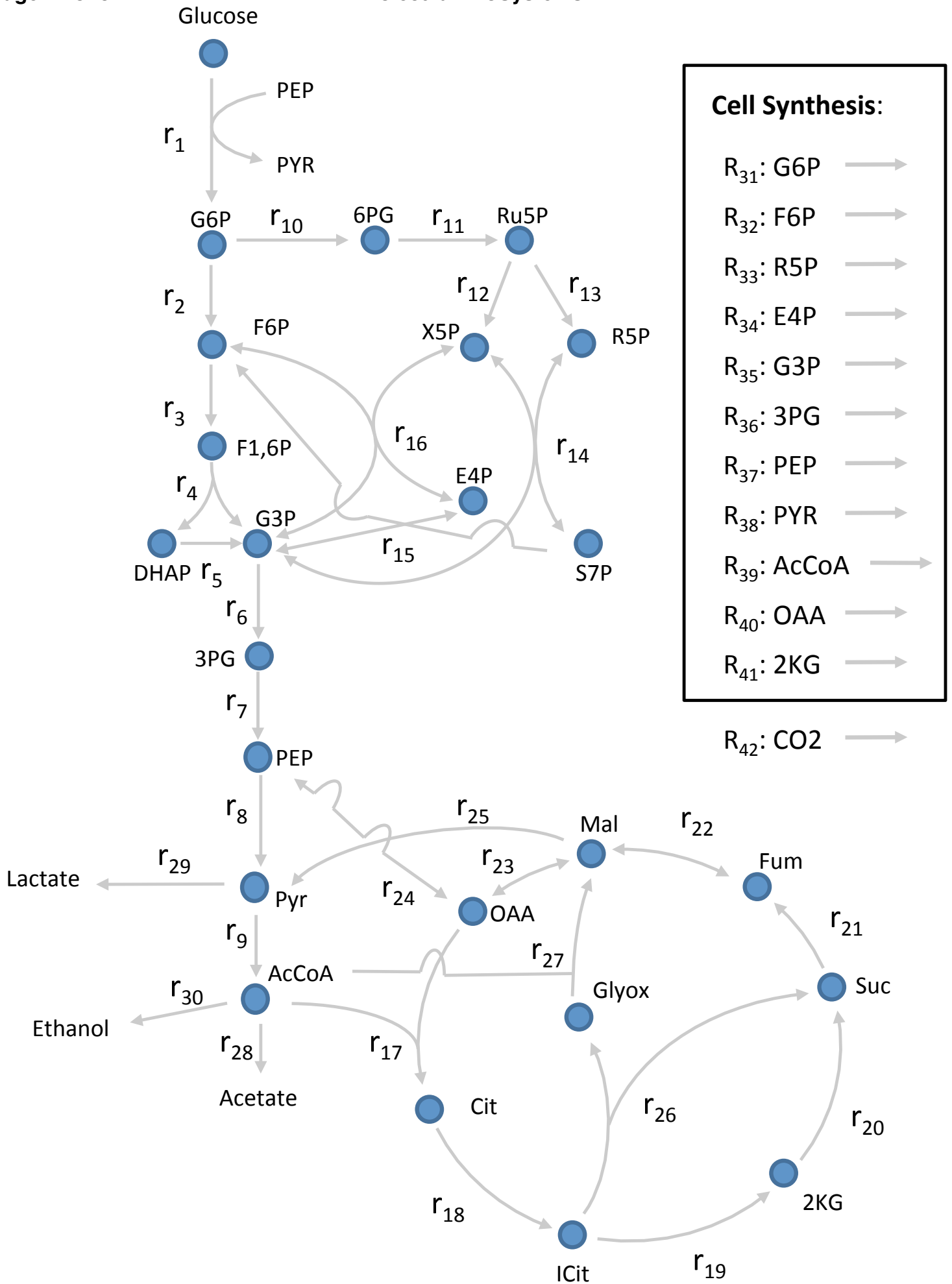
501 **Figure 11.** a) Explained variance per observation and b) Observed versus predicted plot in
502 *P. pastoris* study.

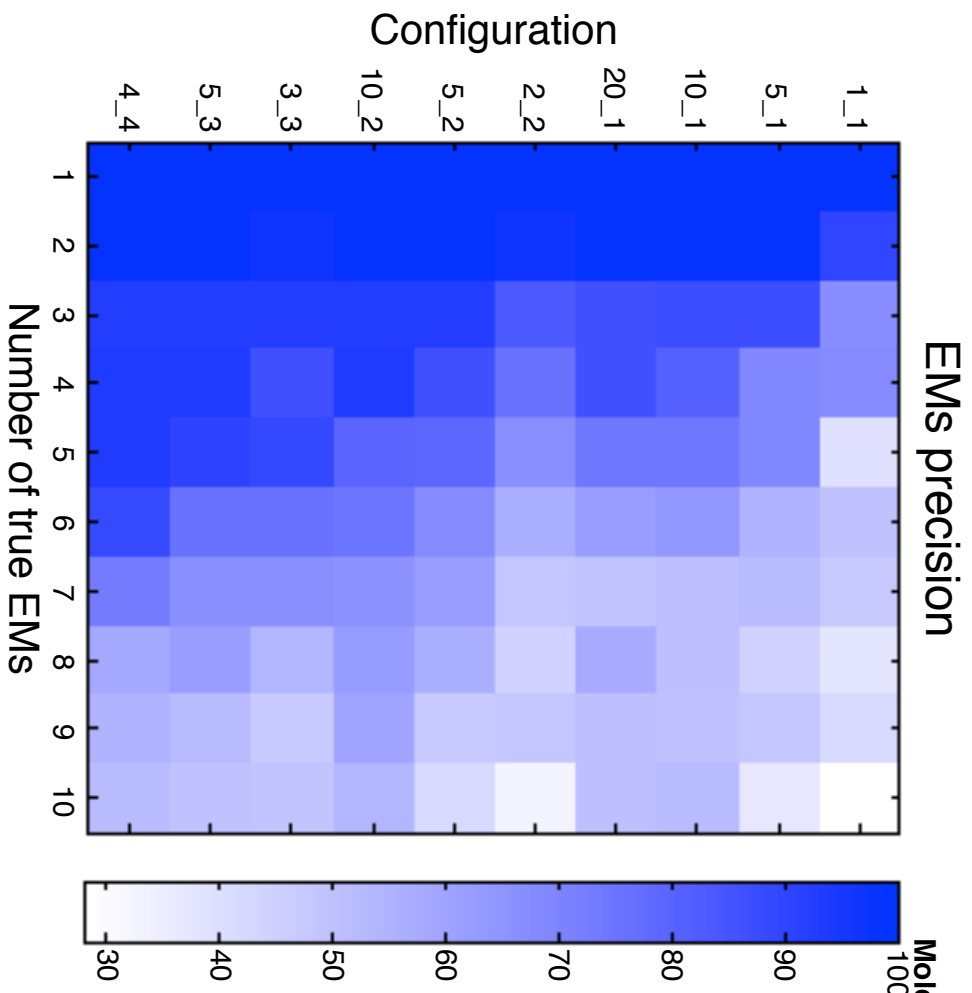
503 **Figure 12.** Principal elementary modes and weightings plots in *P. pastoris* study. The
504 PEMs are represented by columns in both plots; reactions and observations appear row-
505 wise in each plot, respectively. Blue squares represent positive values, and dashed red ones
506 the negatives. The darker the colour, the more highly positive/negative is the value.



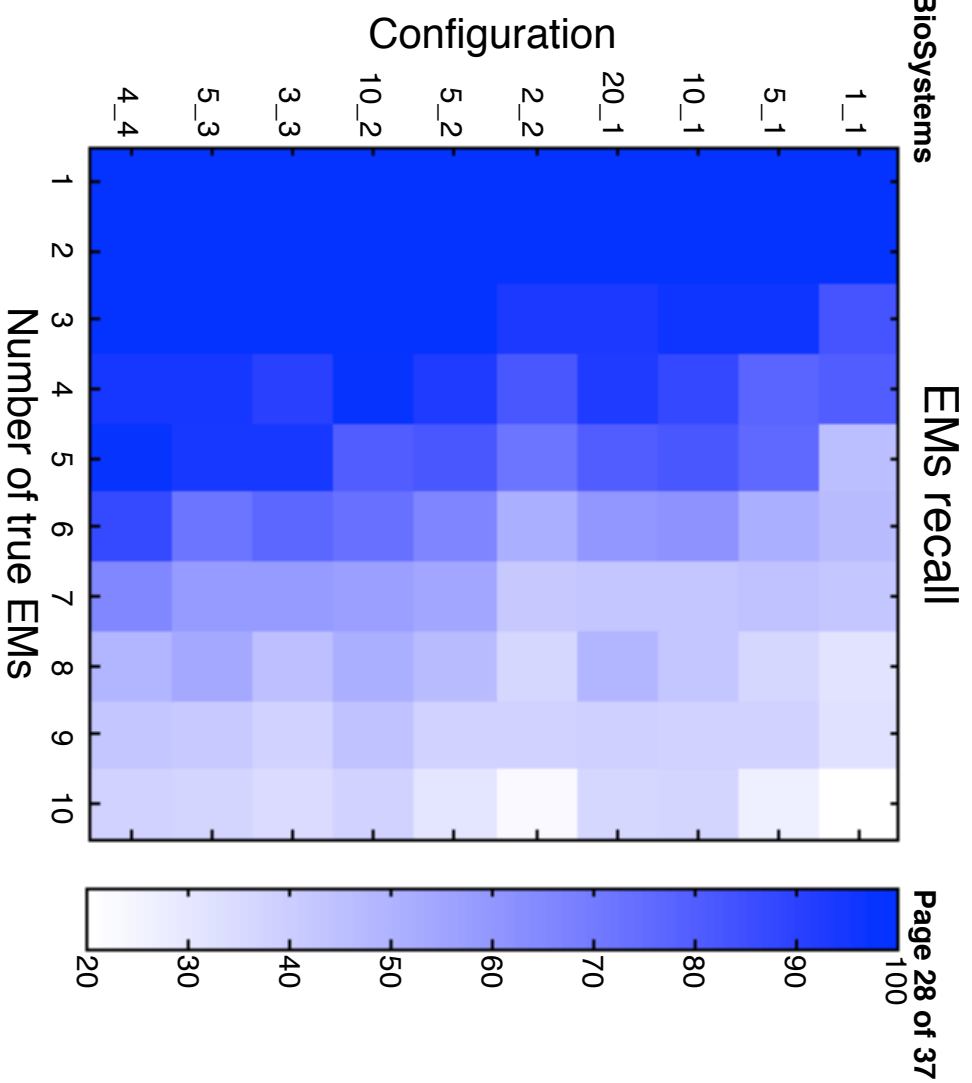
$R = B = 1$ $R = 2, B = 1$

$R = 3, B = 2$

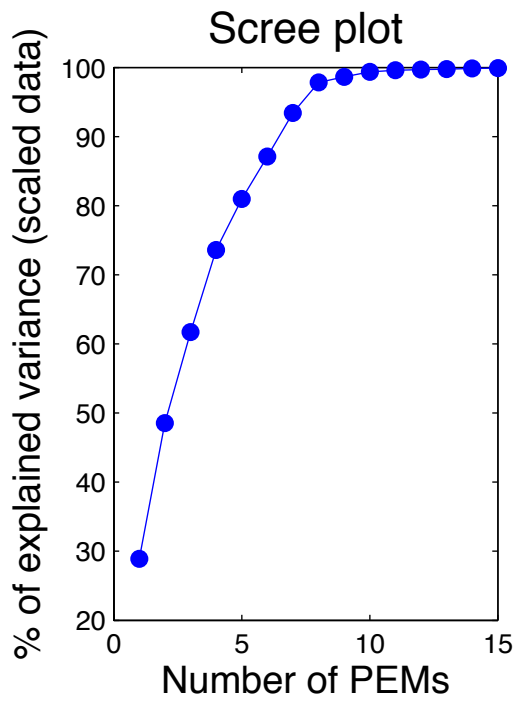




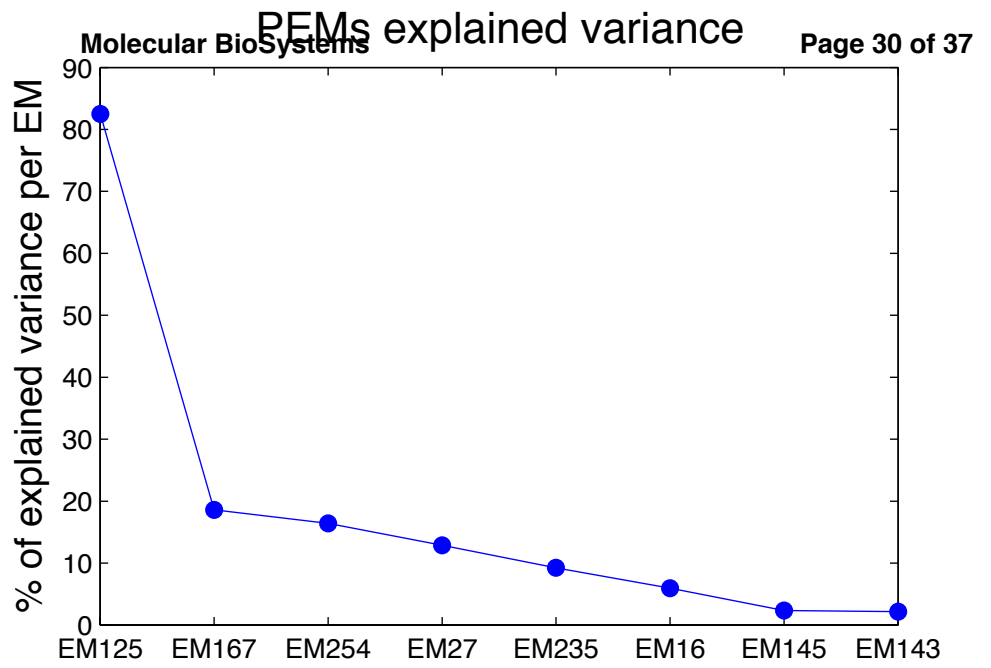
a)



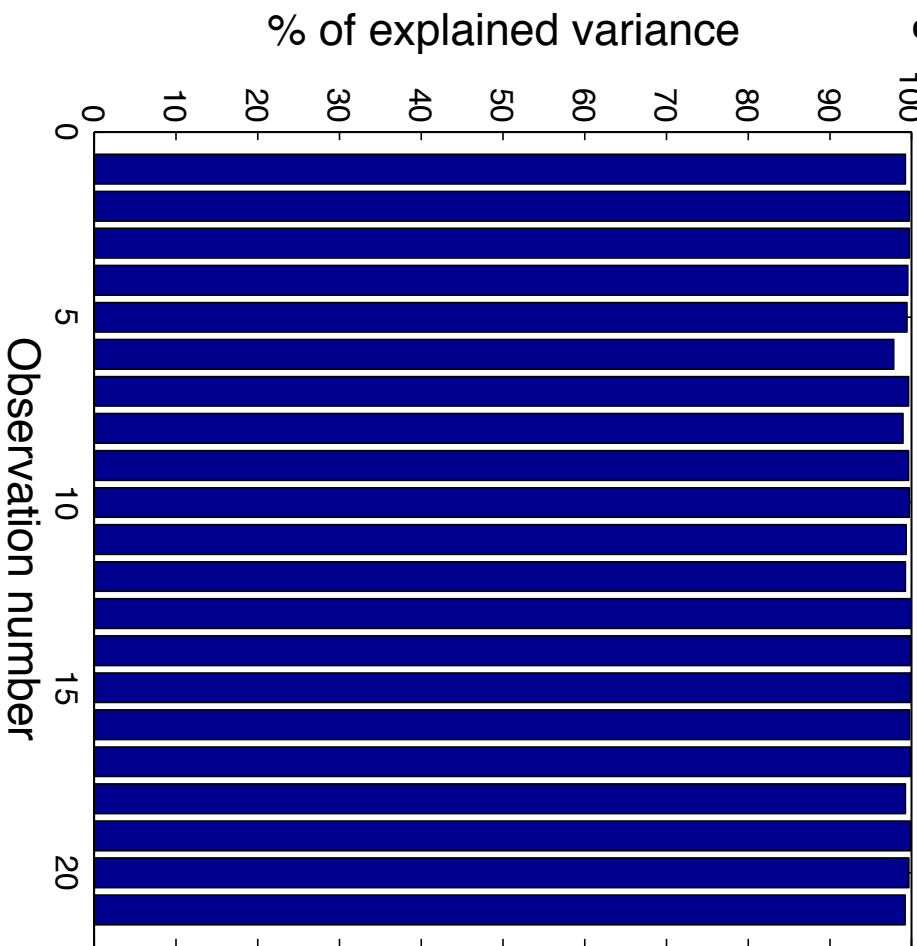
b)



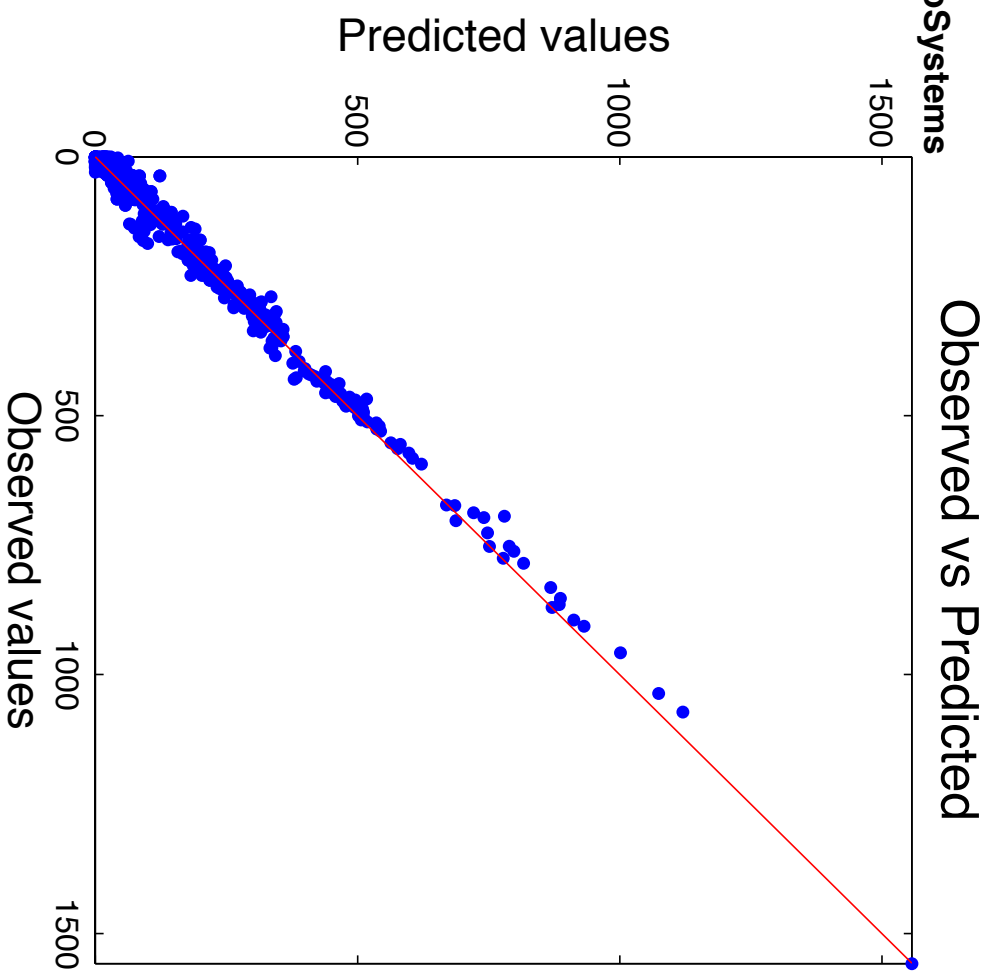
a)



b)

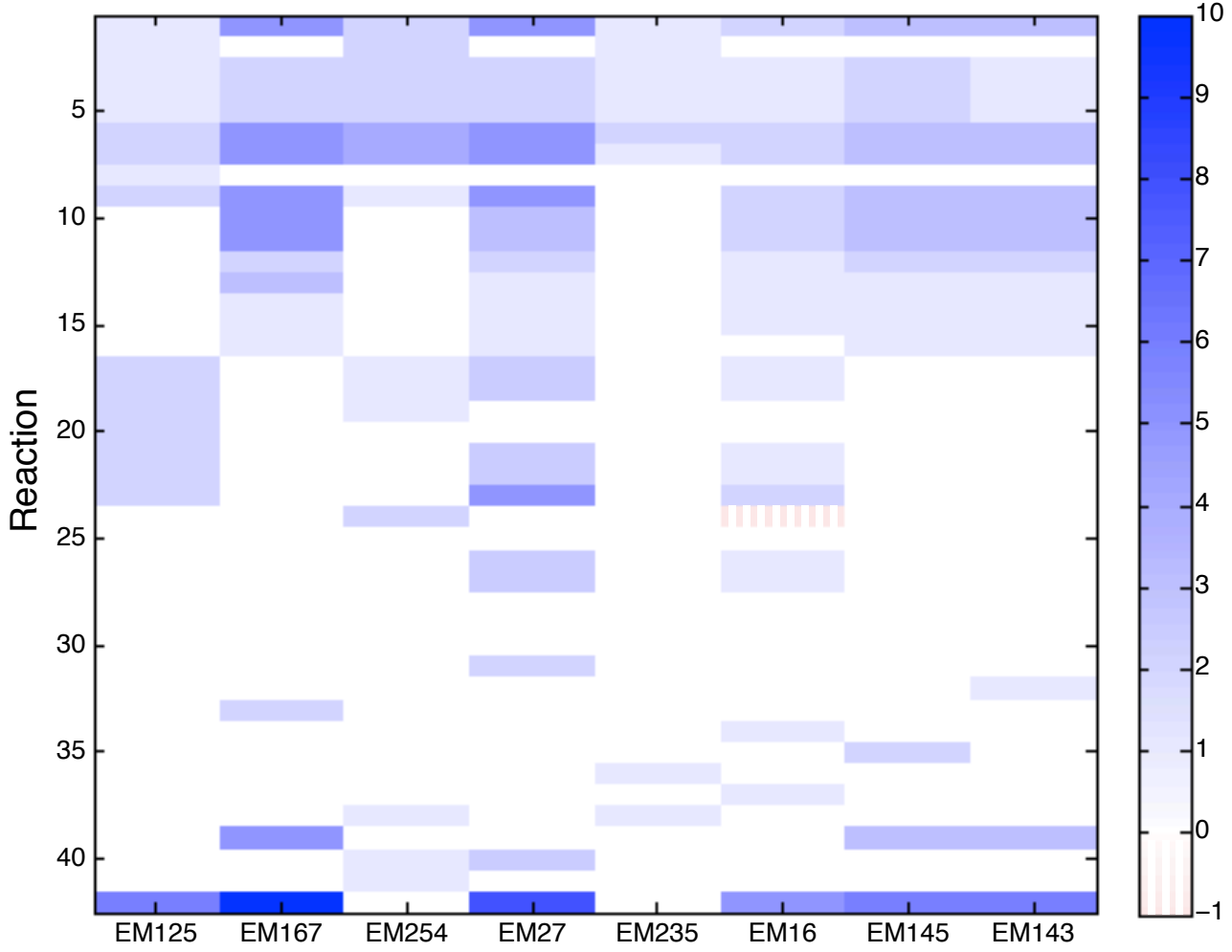


a)

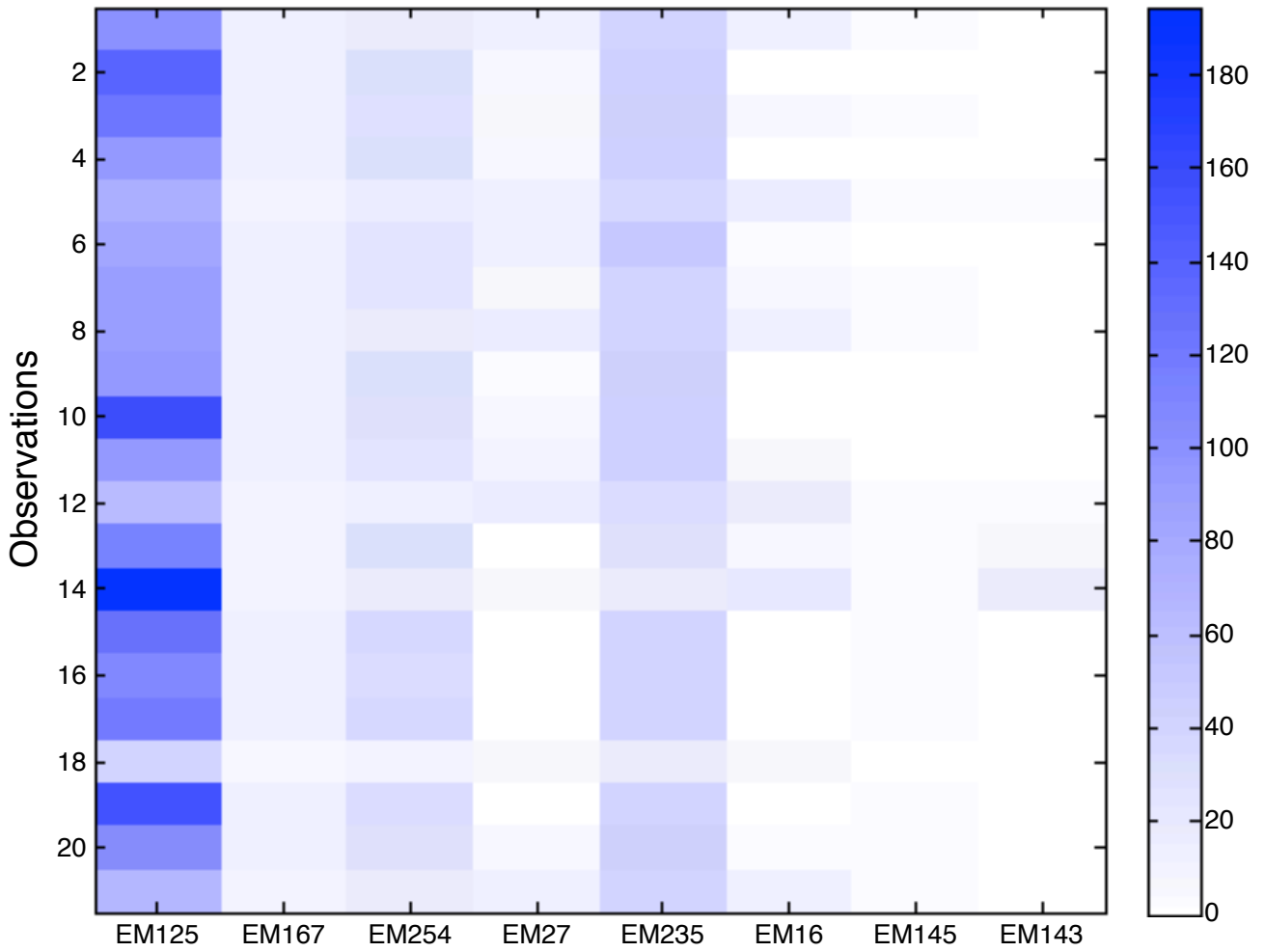


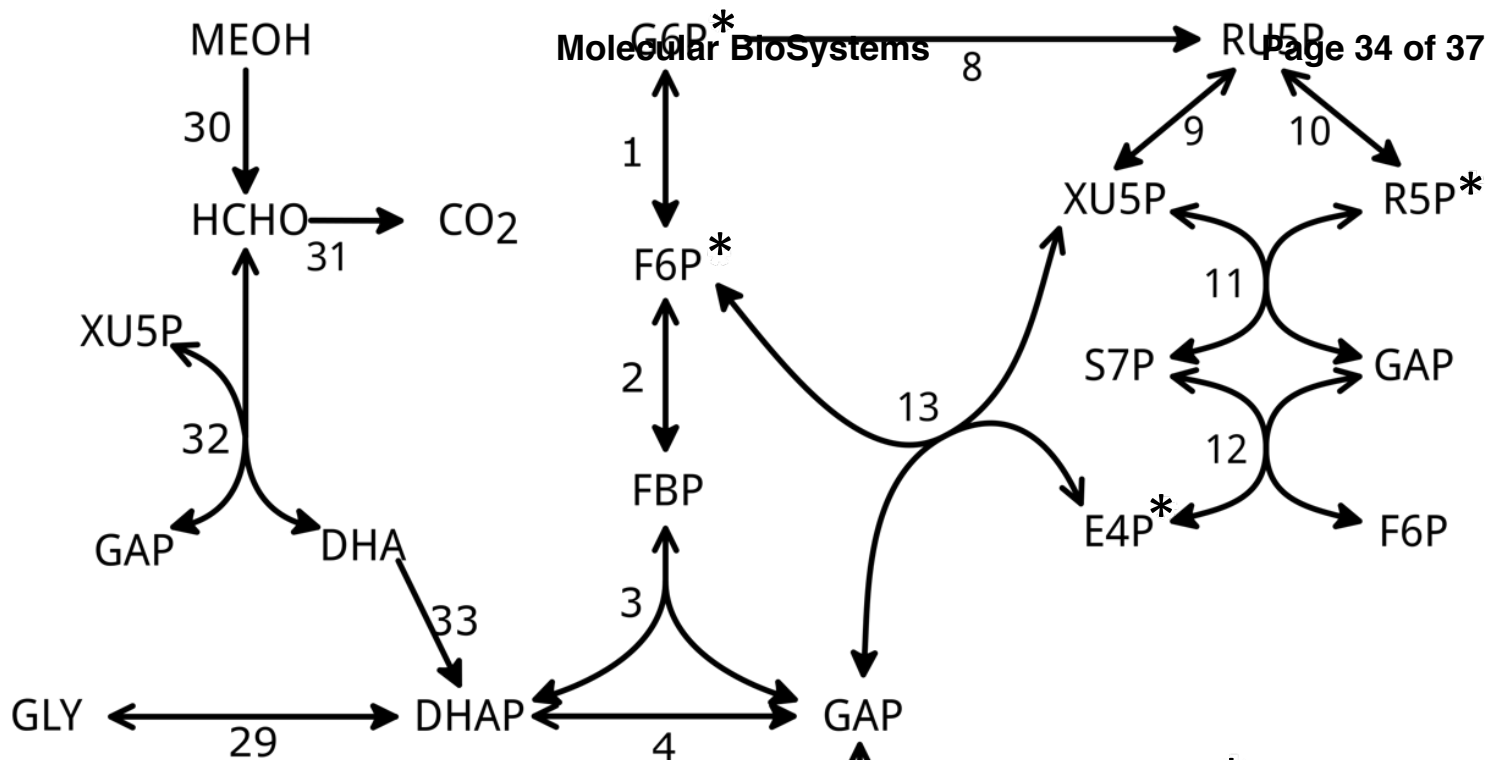
b)

Principal elementary modes plot

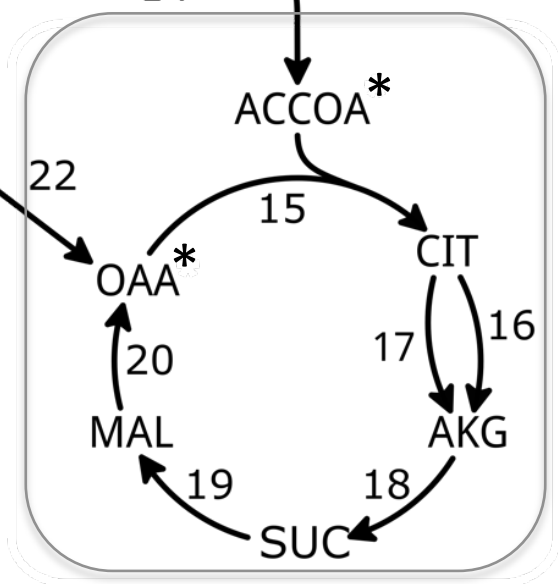


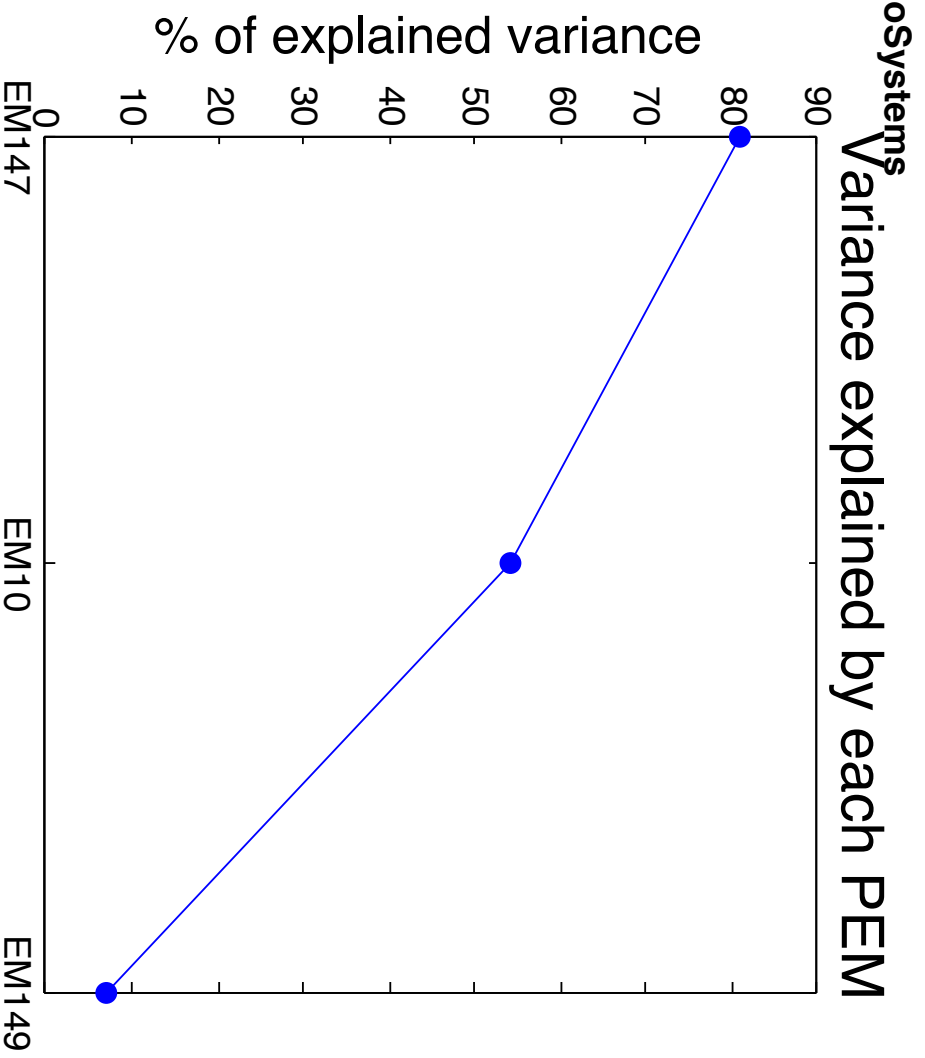
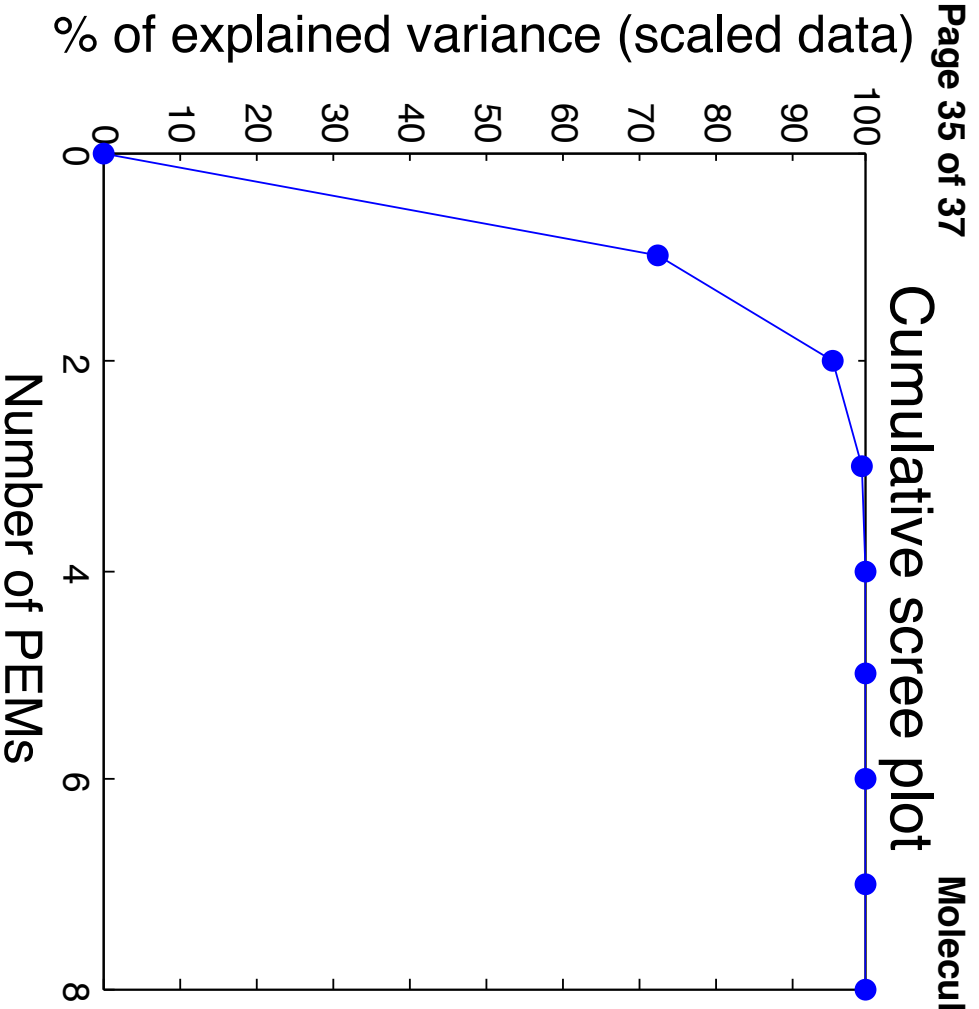
Weightings plot



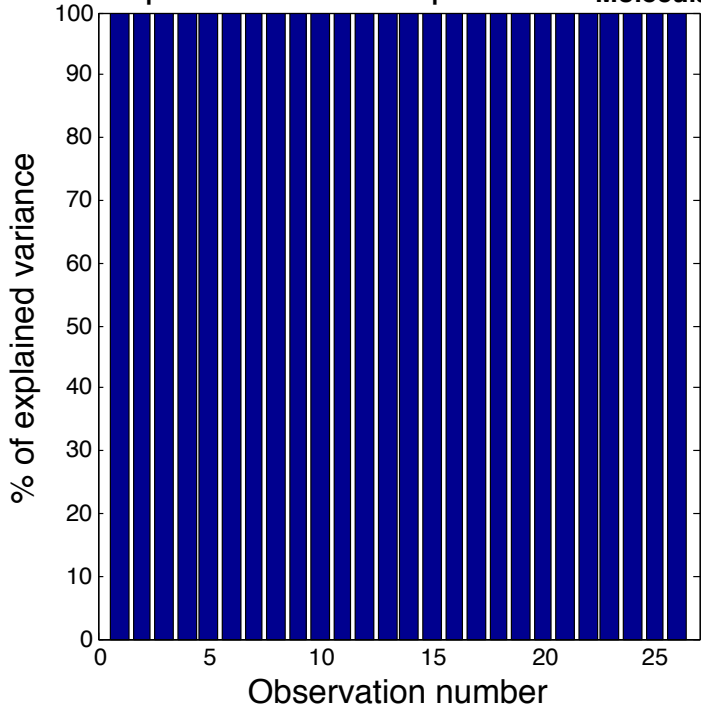


- 23 AKG[m] → AKG[c]*
- 28 O₂ → H₂O
- 43 * → Biom
- 34 GLC[e] → G6P[c]
- 35 GLY[e] ↔ GLY[c]
- 36 MEOH[e] → MEOH[c]
- 37 O₂[e] → O₂[i]*
- 38 CO₂[e] ← CO₂[i]
- 39 ETOH[e] → ETOH[c]
- 40 AC[e] ← AC[c]
- 41 PYR[e] ← PYR[c]
- 42 CIT[e] ← CIT[c]

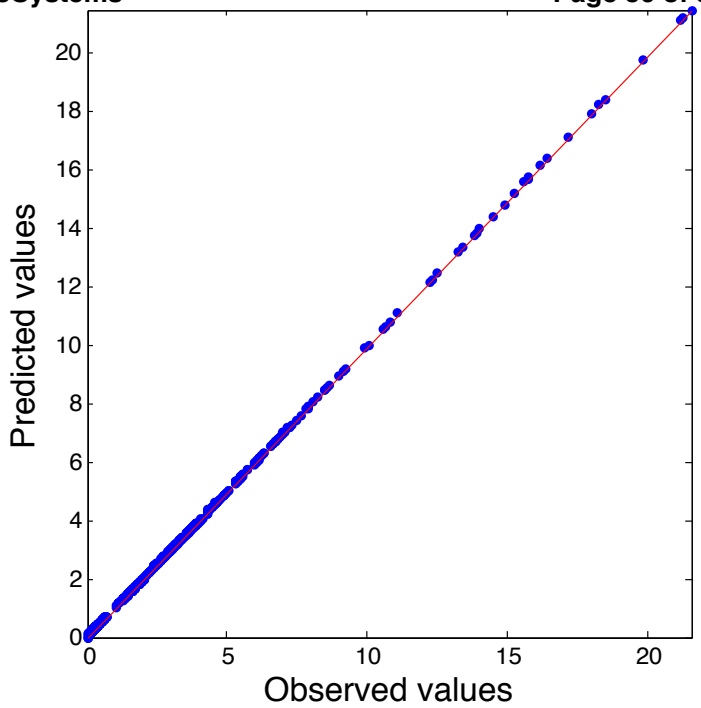




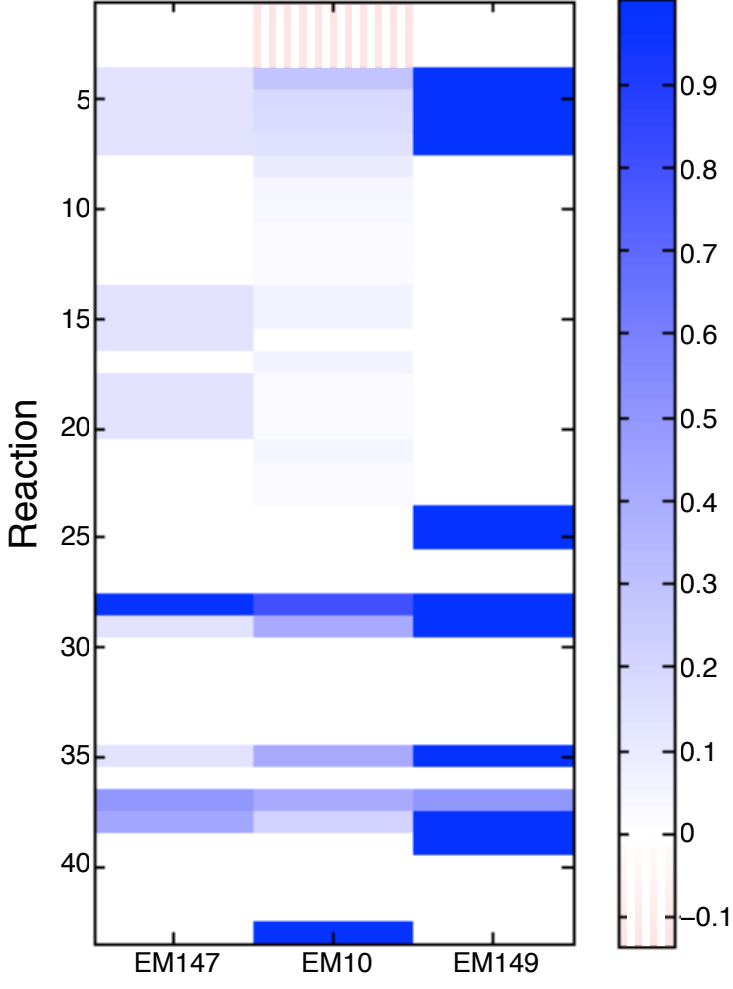
Explained variance per observation



Observed vs Predicted



Principal elementary modes plot



Weightings plot

