### 1. Reconstruction of the draft GEM

The reconstruction was based on the MetaCyc [(*1*)](https://paperpile.com/c/KOfw3T/d7D9) and a template model iML1515 of *E. coli* [*(2)*](https://paperpile.com/c/KOfw3T/Zuyr). For MetaCyc based reconstruction, the RAVEN 2.0 toolbox [(*3*)](https://paperpile.com/c/KOfw3T/vnm0) was applied with a bit-score cutoff of 100. The transport reactions were manually collected from MetaCyc database and added into the draft model. For iML1515 based reconstruction, the ortholog pairs between H. TD01 and E.coli were firstly identified with a bi-directional best hit (BBH) approach. The cutoff for evalue is 1e-3 and the length coverage is 45%. Then the reactions associated with genes in those pairs from iML1515 were extracted. The gene-reaction association rules (gr-rule) were then refined by replacing *E. coli* genes with orthologs from *H.* TD01 based on BBH. For those genes without an ortholog in H.TD01, a mapping based on homology was performed using the same evalue cutoff and length coverage. In the end, remove those reactions with one or more associated genes (involved in “and” association ) without any homolog in H. TD01. The two draft models based on either MetaCyc or iML1515 were merged based on the metacyc id, by giving the priority to iML1515 based draft model if there exists any conflicts. This yields the draft model Halo-GEM for H. TD01. The core-biomass function from iML1515 was used for Halo-GEM. Grouping the metabolites in the core-biomass for the synthesis of protein, DNA, RNA, lips and ions by creating the synthetic reaction for pseudo-metabolites protein\_c, dna\_c, rna\_c, lipid\_c and ions\_c. Manually add PHA synthesis reaction: “ PD-650\_c --> CO-A\_c + PHA\_c” and secretion “PHA\_c --> ”.

**Gap-filling**

Following medium composition was used

|  |  |  |
| --- | --- | --- |
| **Medium** | **Components** | **Concentration (g/L)** |
| 60MMG | Sodium chloride | 60 |
| Glucose | 30 |
| Urea | 0.6 |
| Yeast Extraction | 1 g/L |
| MgSO4 | 0.2 |
| Na2HPO4·12H2O | 9.65 |
| KH2PO4 | 1.5 |
| Trace element mixture I | 10mL/L |
| Trace element mixture II | 1mL/L |

|  |  |  |
| --- | --- | --- |
| **Medium** | **Components** | **Concentration (mg/L)** |
| Trace element I | Fe(III)-NH4-Citrate | 50 |
| CaCl2·2H2O | 20 |

Trace element II:

|  |  |  |
| --- | --- | --- |
| **Medium** | **Components** | **Concentration (**μg/L**)** |
| Trace element II | ZnSO4·7H2O | 100 |
| MnCl2·4H2O | 30 |
| **H3BO3** | 300 |
| **CoCl2·6H2O** | 200 |
| CuSO4·5H2O | 10 |
| NiCl2·6H2O | 20 |
| NaMoO4·2H2O | 30 |

For yeast extract, 20 amino acids as well as adenine, trehalose and lactate were considered as major components [(*4*)](https://paperpile.com/c/KOfw3T/AnM2).

Since the H. TD01 is able to grow in the medium without amino acids (REF), all uptake reactions for amino acids were blocked before gap-filling. Then a sequential gap-filling was performed to let the model be able to produce each metabolite in the biomass function, in the order of DNA, RNA, protein, lipids and other metabolites. The iML1515 model was used as the universal model for gap-filling. The “cobra.flux\_analysis.gapfilling” from cobrapy [(*5*)](https://paperpile.com/c/KOfw3T/MJFj) was used for this step.

Some statistics of the resulting GEM is as follows:

|  |  |
| --- | --- |
| **Reactions** |  |
| Metabolic rxn | 1863 |
| Transport rxn | 351 |
| **Gene associations** |  |
| Gene associated Metabolic rxn | 1847 |
| Gene associated Transport rxn | 301 |
| No gene associations metabolic reactions | 16 |
| No gene associations transport reactions | 50 |
| **Exchange reactions** | 56 |
| **Metabolites** |  |
| Unique metabolites | 1972 |
| Cytoplasmic | 1912 |
| Periplasmic | 350 |
| Extracellular | 77 |

### 2. Enzyme constrained GEM: Halo-ecGEM

The central concept of enzyme-constrained models is to consider the enzyme resource required to carry each flux catalyzed by enzymes.

[1]

[2]

In which A, B and E are substrate, product and the enzyme that catalyzes this reaction, respectively. is the turnover number. This enables the integration of absolute quantitative proteomics data by setting the upper bound for the reaction producing *E* (Eq. 2). For those enzymes lacking protein abundance data, the pool constraint was applied:

[3]

The conversion from GEM to enzyme-constrained GEM was done following the protocols described in [(*6*)](https://paperpile.com/c/KOfw3T/PEFs).

#### 2.1 Estimation of enzyme turnover numbers ()

As shown in Eq. 1, the values for all reactions in the model are required. For this, as there is no reported values from BRENDA [(*7*)](https://paperpile.com/c/KOfw3T/Ml3P) for any of enzymes from this species, some criteria was designed to estimate those values indirectly. Firstly, download “max\_KCAT.txt” file which contains kcat values collected from BRENDA [(*6*)](https://paperpile.com/c/KOfw3T/PEFs) from <https://github.com/SysBioChalmers/GECKO/blob/master/databases/max_KCAT.txt>. Only those kcat values from **Bacteria** will be used. For a given enzymatic reaction from Halo-GEM:

1. Try to match EC number and substrate. It will find kcat values of reactions from any bacteria that have the same EC number and substrate with the reaction from Halo-GEM.
2. If failed, try to match only EC number;
3. If no match is found, introduce one wildcard to the last digit of the EC number. Then repeat step (1) and (2). If still no match is found, introduce two wildcards to the last two digits of the EC number. Then repeat again. Repeat until all digits of the EC number are replaced with a wildcard.

In the end, a list of kcat values is collected for each reaction. Use the mean and standard deviation of log10-transformed values as the estimated kcat and associated uncertainties. For those reactions with only 1 matched value, use 1 as the uncertainty value for log10-transformed (Figure 1).

|  |  |  |
| --- | --- | --- |
| **Matched digits in EC number** | **Match substrate?** | **Number of rxns** |
| 4 (e.g. EC1.1.1.1) | Yes | 392 |
| 4 (e.g. EC1.1.1.1) | No | 679 |
| 3 (e.g. EC1.1.1.X) | Yes | 144 |
| 3 (e.g. EC1.1.1.X) | No | 587 |
| 2 (e.g. EC1.1.X.X) | Yes | 2 |
| 2 (e.g. EC1.1.X.X) | No | 9 |
| 1 (e.g. EC1.X.X.X) | No | 4 |
| 0 (e.g. ECX.X.X.X) | Yes | 2 |
| 0 (e.g. ECX.X.X.X) | No | 31 |

Note: EC1.1.1.1 is just for illustration. It refers to any EC number in the model.

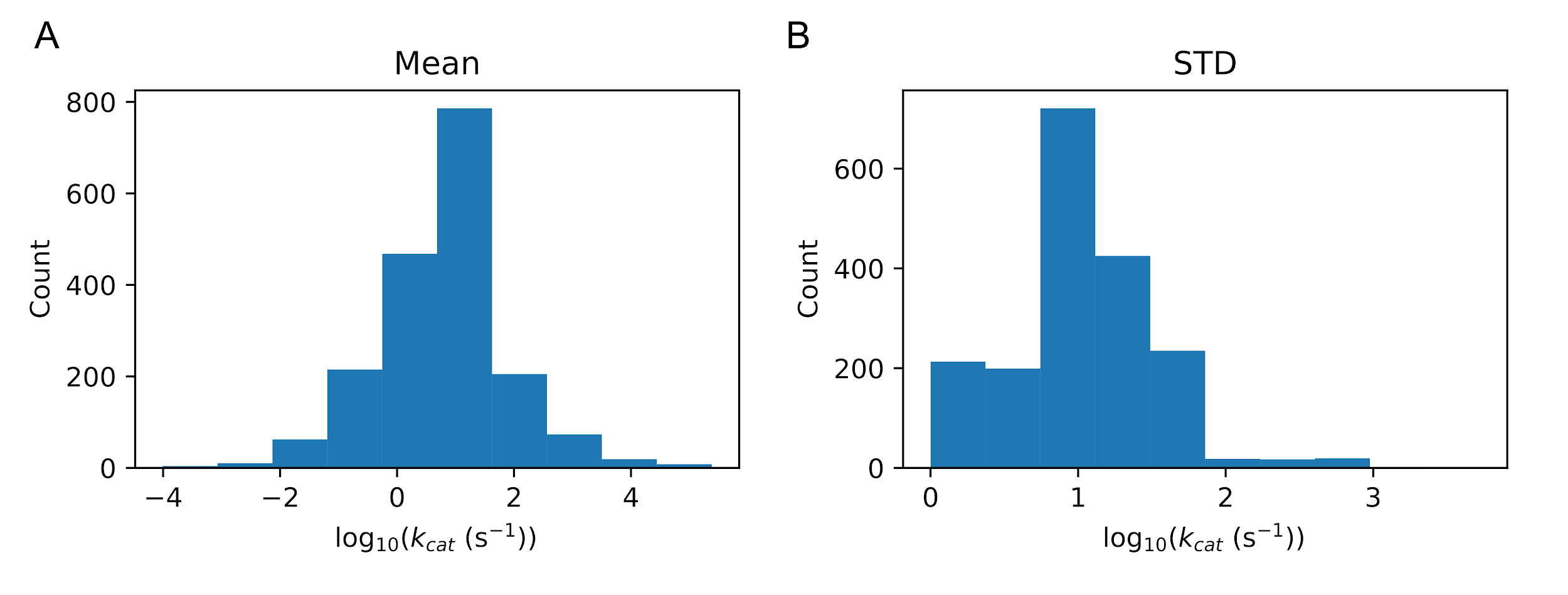


Figure 1. Estimated values. (A) Distribution of values; (B) Distribution of standard deviation of estimated values.

In the second step, single-reaction deletion analysis was performed to identify a list of essential reactions. The minimal medium was used. Those reactions with a predicted growth rate decreased to less than 10% of that of the wild-type were considered as ensiential. For those reactions with an estimated kcat value lower than 1 1/s were set as 1 1/s and a standard deviation of 2 in log10 space were used.

#### 2.2 Calibrating with proteomics data

We next tested if the Halo-ecGEM equipped with those estimated values can predict the cell growth. The experimental data used is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Condition | Total protein content  (g/gdw) | Specific growth rate (h-1) | P3HB (mmol/gcdw/h) |
| Batch | 0.17 | 0.16 | 0.57 |
| Batch-HN | 0.17 | 0.16 | 0 |
| Fed-batch 9h | 0.10 | 0.27 | 0.09 |
| Fed-batch 19h | 0.10 | 0.07 | 1.14 |
| Fed-batch 30h | 0.12 | 0.03 | 0.80 |

The Halo-ecGEM constrained by the total protein content and P3HB flux failed in maximizing the cell growth rate as inclusion of those constraints makes the optimization infeasible. This indicates that the model is over-constrained and estimated values of some essential enzymes are too small. This is not surprising considering the big large variance in estimated values (Figure 1B). We thereby performed Bayesian statistical learning [(*8*)](https://paperpile.com/c/KOfw3T/9unx) to calibrate those values.

The absolute quantitative proteomics data at those conditions were generated. Two different constraints strategies were tested:

1. Constraint the total protein abundance of individual enzymes with proteomics data
2. Group the enzymes into unique groups (each enzyme only belongs to one group) based on go-slims, and then constrain the total protein amount for each enzyme group instead of single enzymes

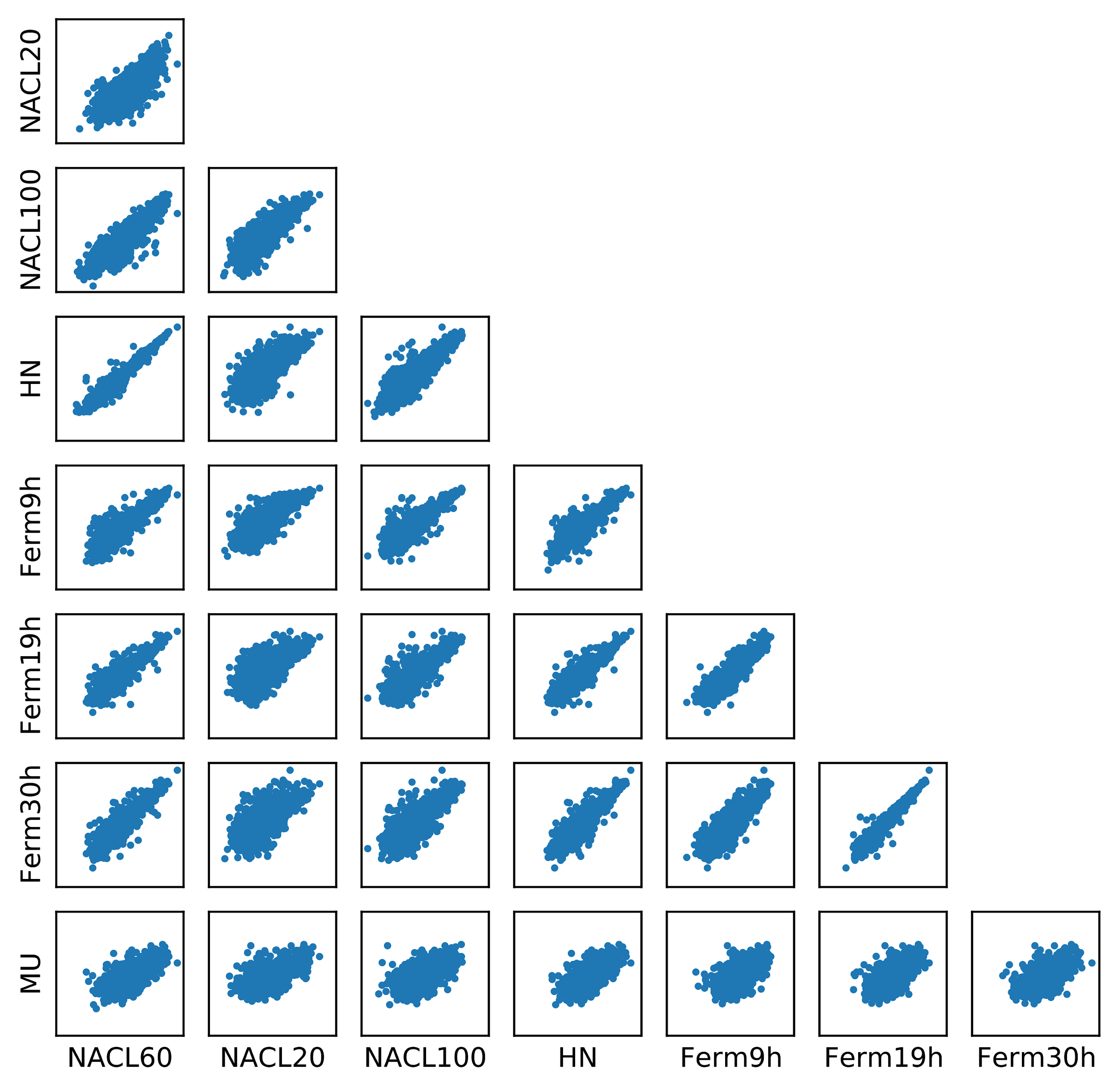


Figure 3. Protein abundance comparison among individual proteins

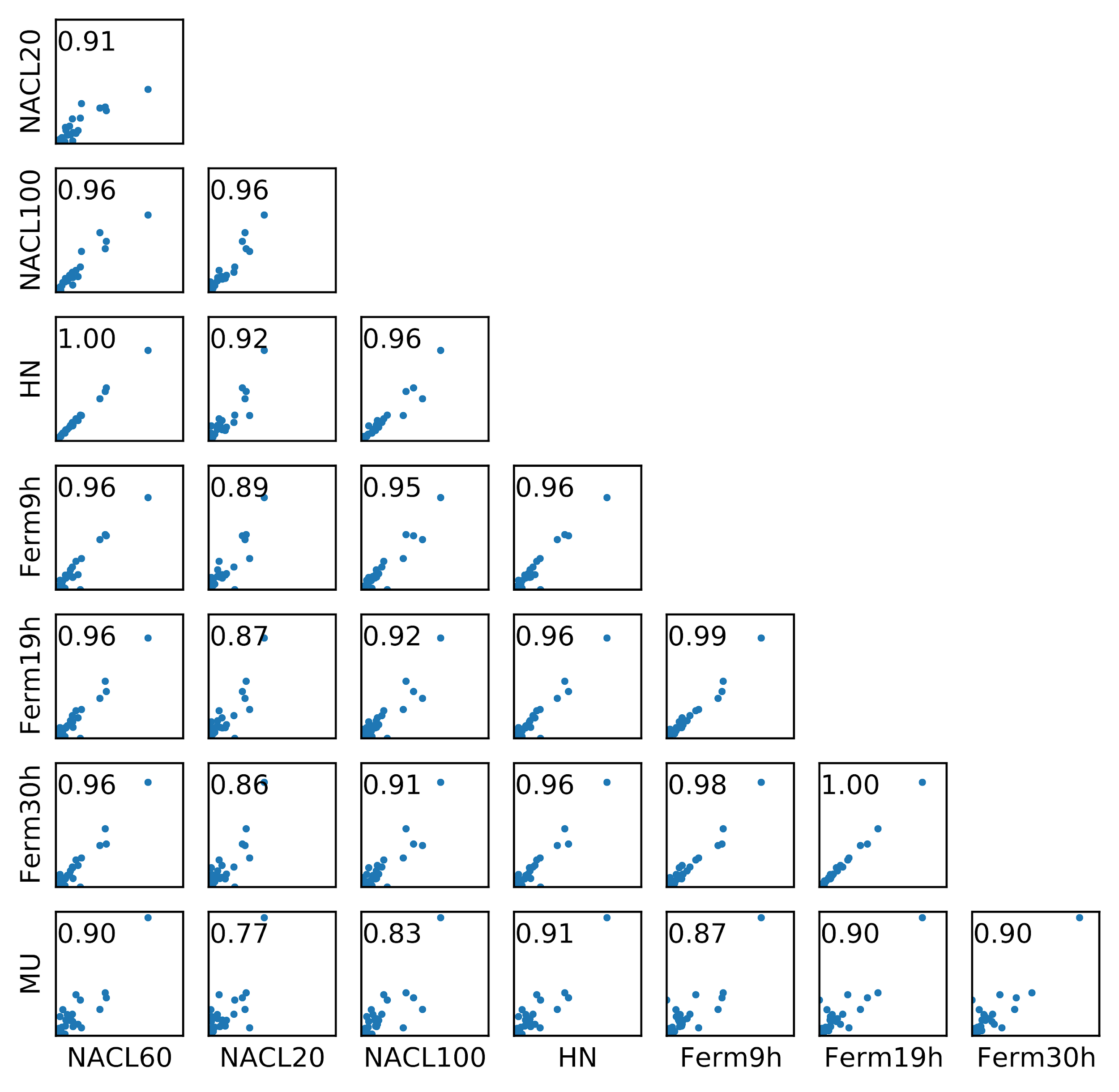


Figure 4. Mass fraction comparison when group enzymes into groups based on go-slims.

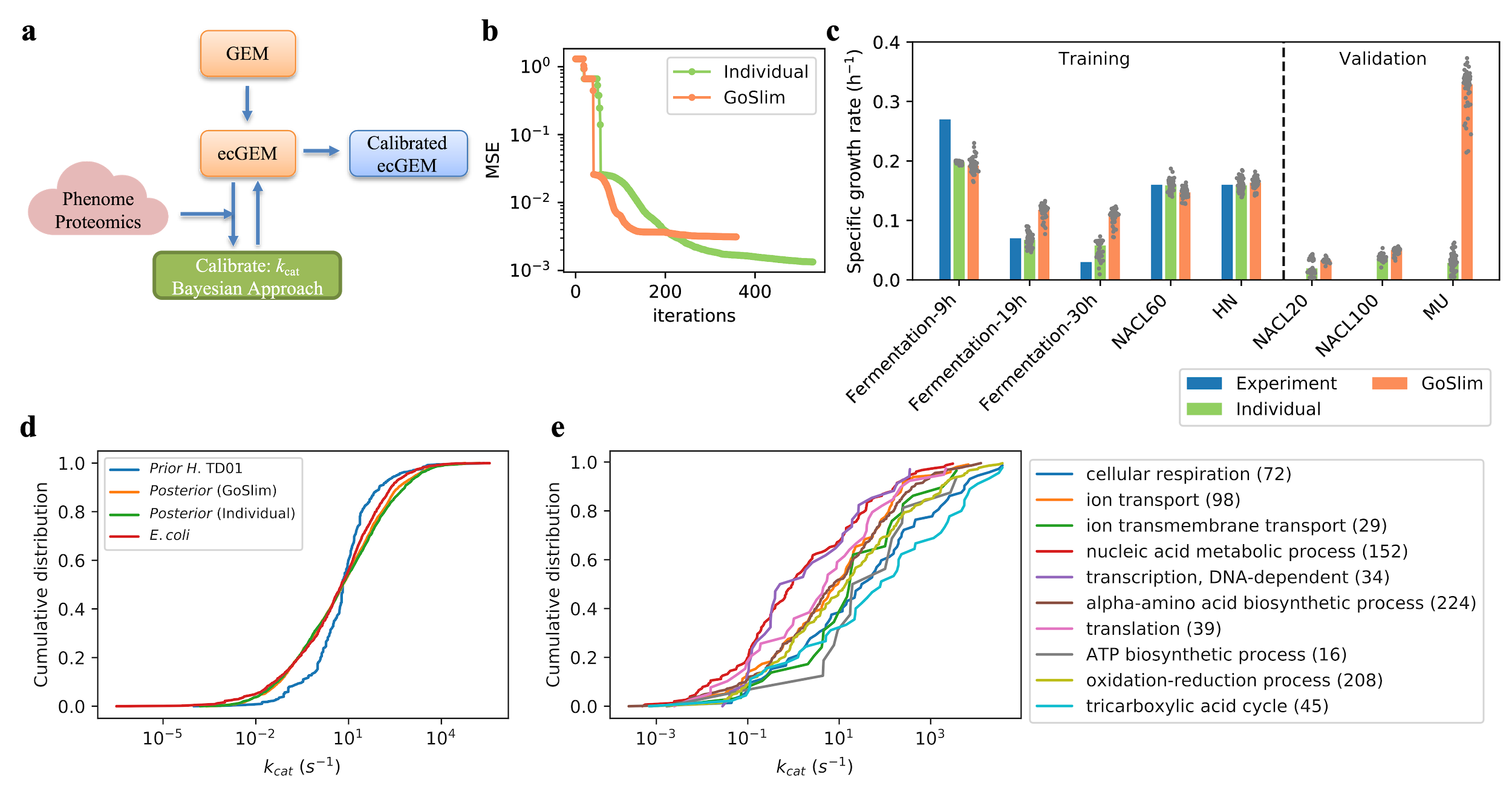
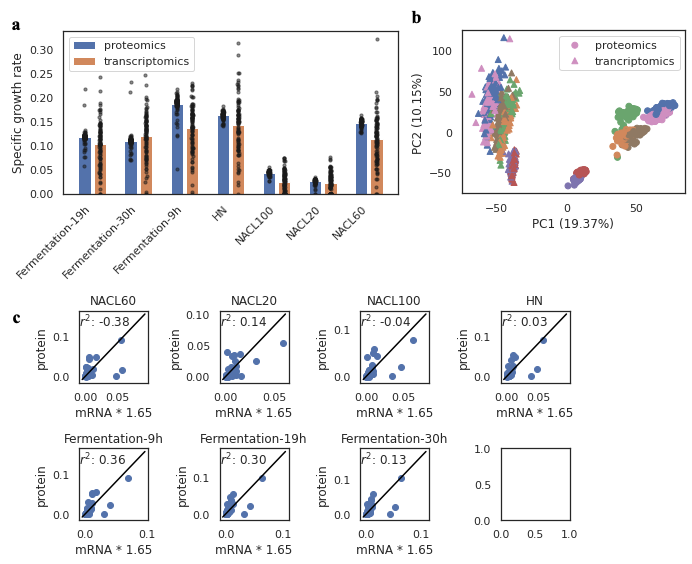


Figure 5. (a) Illustration of Bayesian modelling approach. (b) The mean squared error between experimental and predicted values during training iterations. (c) the simulated specific growth rate with *Posterior* models from two approaches. (d) Cumulative distributions of kcat values in *Prior*, *Posterior* of both strategies, as well as experimental kcats of E. coli. (e) Cumulative distributions of kcat values of enzymes belonging to different GoSlim terms. The number in the parentheses indicates the number of kcat values in that GoSlim term.

Test if we can use mRNA level to constrain the model:

Experiment setup: (1) calculate the mass fraction of summation of unique proteins in each GoSlim term based on transcriptomics data. (2) constrain the model with this new fraction times a coefficient of 1.65. (3) simulate the maximal specific growth rate and perform flux variability analysis (FVA). (4) do the same analysis for models constrained with proteomics.

Results: the simulated specific growth rates based on RNAseq seems to be consistent with ones based on proteomics, as shown below (Figure 6a). However, with much bigger variations. Furthermore, if we use principal component analysis to cluster the simulation results based from FVA (Figure 6b), those simulations were primarily clustered based on the omics data secondly on conditions within, which against our expectation that it should only be clustered based on conditions. This indicates that transcriptomics data can not be used in this way to constrain the ecGEM. This can be explained poor explained variance by 1.65\*mRNA fraction for proteomics based mass fraction (Figure 6c).

****

**Figure 6. b is colored based on conditions. In c, r2 indicate the coefficient of determination**

### Reference

1. [R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, P. D. Karp, The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **36**, D623–31 (2008).](http://paperpile.com/b/KOfw3T/d7D9)

2. [J. M. Monk, C. J. Lloyd, E. Brunk, N. Mih, A. Sastry, Z. King, R. Takeuchi, W. Nomura, Z. Zhang, H. Mori, A. M. Feist, B. O. Palsson, iML1515, a knowledgebase that computes Escherichia coli traits. *Nature Biotechnology*. **35** (2017), pp. 904–908.](http://paperpile.com/b/KOfw3T/Zuyr)

3. [H. Wang, S. Marcišauskas, B. J. Sánchez, I. Domenzain, D. Hermansson, R. Agren, J. Nielsen, E. J. Kerkhoven, RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor, , doi:](http://paperpile.com/b/KOfw3T/vnm0)[10.1101/321067](http://dx.doi.org/10.1101/321067)[.](http://paperpile.com/b/KOfw3T/vnm0)

4. [J. Zhang, J. Reddy, B. Buckland, R. Greasham, Toward consistent and productive complex media for industrial fermentations: studies on yeast extract for a recombinant yeast fermentation process. *Biotechnol. Bioeng.* **82**, 640–652 (2003).](http://paperpile.com/b/KOfw3T/AnM2)

5. [A. Ebrahim, J. A. Lerman, B. O. Palsson, D. R. Hyduke, COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7**, 74 (2013).](http://paperpile.com/b/KOfw3T/MJFj)

6. [B. J. Sánchez, C. Zhang, A. Nilsson, P.-J. Lahtvee, E. J. Kerkhoven, J. Nielsen, Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).](http://paperpile.com/b/KOfw3T/PEFs)

7. [L. Jeske, S. Placzek, I. Schomburg, A. Chang, D. Schomburg, BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).](http://paperpile.com/b/KOfw3T/Ml3P)

8. [G. Li, Y. Hu, H. Wang, A. Zelezniak, B. Ji, J. Zrimec, J. Nielsen, Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. *bioRxiv* (2020), p. 2020.04.01.019620.](http://paperpile.com/b/KOfw3T/9unx)