1 # A genome-scale metabolic network model synergizes with
2 # statistical learning to predict amino acid concentrations in
3 # Chinese Hamster Ovary cell cultures
4

5 Song-Min Schinn[1], Carly Morrison[2], Wei Wei[2], Lin Zhang[2], Nathan E. Lewis[1,3,4]

6 1. Department of Pediatrics, University of California, San Diego

7 2. Pfizer, Biotherapeutics Pharmaceutical Sciences, Andover, MA

8 3. Department of Bioengineering, University of California, San Diego

9 4. Novo Nordisk Foundation Center for Biosustainability at UC San Diego

10

Abstract

12 The control of nutrient availability is critical to large-scale manufacturing of biotherapeutics.

13 However, the quantification of proteinogenic amino acids is time-consuming and thus is difficult

14 to implement for real-time *in situ* bioprocess control. Genome-scale metabolic models describe

15 the metabolic conversion from media nutrients to proliferation and recombinant protein

16 production, and therefore are a promising platform for *in silico* monitoring and prediction of

17 amino acid concentrations. This potential has not been realized due to unresolved challenges: (1)

18 the models assume an optimal and highly efficient metabolism, and therefore tend to

19 underestimate amino acid consumption, and (2) the models assume a steady state, and therefore

20 have a short forecast range. We address these challenges by integrating machine learning with

21 the metabolic models. Through this we demonstrate accurate and time-course dependent

22 prediction of individual amino acid concentration in culture medium throughout the production

23 process. Thus, these models can be deployed to control nutrient feeding to avoid premature

24 nutrient depletion or provide early predictions of failed bioreactor runs.

0

1   Short Communication

2   Chinese Hamster Ovary (CHO) cells are widely used to manufacture complex

3   biotherapeutic molecules at large scales. Industrial bioprocesses ensure high product yield and

4   quality by maintaining favorable growth conditions in cell culture environments, which requires

5   careful monitoring and control of nutrient availability. Chemically-defined serum-free media can

6   contain dozens or >100 components[1], but key nutrients include proteinogenic amino acids, which

7   are direct substrates and regulators[2,3] of proliferation and protein synthesis. Unfortunately,

8   conventional methods for amino acid quantification based on liquid chromatography and mass

9   spectrometry are time-consuming and difficult to use for decision making and control of cell

10   culture. Alternate spectroscopic approaches have been sensitive to a limited number of amino

11   acid species[4]. Here we present a computational method to forecast time-course amino acid

12   concentrations from routine bioprocess measurements, facilitating a timely and anticipatory

13   control of the bioprocess (Figure 1).

14   At the foundation of our method is a genome-scale metabolic network model, which

15   accounts for the complex conversion from media nutrients to biomass and recombinant protein

16   production. Such models have been increasingly utilized for CHO cells[5–7] and bioprocess

17   applications[8,9], such as predicting clonal performances[10], identifying metabolic bottlenecks[11], and

18   optimizing media formulation[12,13]. Metabolic network models can also estimate amino acid

19   uptake rates necessary to experimentally support observed proliferation and productivity[14].

20   However, challenges have limited their practical application.

21   First, metabolic network models are typically highly complex but under-constrained, and

22   therefore are easy to overfit. This is mitigated by training the model on a variety of bioprocess

1   conditions and metabolic phenotypes. Second, metabolic network models assume that cells

2   operate at some metabolic optimum, and thus tend to describe an idealized metabolism

3   specifically fit to the assumed objective (e.g., biomass production[15,16], minimization of redox[17]).

4   Third, for the present purpose, these models need to predict amino acid consumption fluxes,

5   typically on the order of 10^{-3} mmol·g_{DW}^{-1}·hr^{-1}(see Methods), from input data that are multiple

6   magnitudes larger, such as growth rate and glucose consumption (10^{-1} to 10^{-2} mmol·g_{DW}^{-1}·hr^{-1}).

7   The preceding two challenges increase prediction error. Lastly, metabolic network models

8   assume a steady state, which reduces the range of forecast. Typically, input data from one day

9   are used to make predictions for the same day. However, such predictions cannot be extended to

10   multiple days or subsequent culture phases, as cross-temporal shifts in metabolism would violate

11   the steady state assumption. Thus, model predictions of amino acid concentrations can be overfit,

12   ideal, and near-sighted – all of which dilutes their practicality for industrial bioprocess control.

13   Here we demonstrate that these weaknesses can be addressed in a data-driven manner by

14   coupling a metabolic network model with machine learning.

15   We developed this hybrid approach on a diverse set of 10 CHO clones with different

16   growth and productivity profiles from two different fed-batch production processes. These CHO

17   clones were subject to different bioprocess conditions and recombinant antibody identities (see

18   Methods), resulting in a variety of phenotypes and productivity performances (Fig. S1). For

19   example, several high-performing clones were exceptionally proliferative or productive,

20   suggesting an efficient conversion from nutrients to biomass or recombinant protein product.

21   Other clones performed these conversions at lower rates, suggesting attenuated metabolic

22   activity or inefficient resource utilization. The CHO cells adjusted their nutrient uptake

23   according to these various metabolic phenotypes, leading to diverse amino acid consumption

2

1   patterns (Fig, S2). For example, the consumption of glucose and serine differed by several fold

2   across conditions and time. Furthermore, different clones varied in their consumption or

3   secretion of key metabolites such as lactate, alanine, glycine, and glutamine.

4   We sought to predict these diverse consumption behaviors using a tailored model of CHO

5   metabolism[18]. As input information, we utilized the following routinely measured industrial

6   bioprocess data: (1) viable cell density and titer measurements, from which growth rate and

7   specific productivity are calculated (Methods, equation 1), and (2) bioreactor concentrations of

8   glucose, lactate, glutamate and glutamine, from which their respective consumption rates are

9   calculated. These measurements were used as boundary conditions by constraining the fluxes of

10   biomass production, recombinant protein synthesis and consumption of the four metabolites to

11   observed values. Subsequently, we used Markov chain Monte Carlo sampling of metabolic

12   fluxes[19] to sample the range and magnitude of all reaction fluxes to calculate the likely uptake

13   fluxes of the remaining 18 proteinogenic amino acids (see Methods). These predictions were

14   applied to the CHO clones across 8 days of a 12-day production run (days 4 to 11), resulting in a

15   total of 80 individual predictions.

16   We evaluated the resulting model predictions in two ways. First, we examined the

17   differences in model predictions and experimental measurements of amino acid uptake and

18   secretion (Figure 2A). For most amino acids, this difference was small compared to the scale of

19   input data, suggesting that metabolic models can describe the conversion from nutrients to

20   biomass and recombinent proteins. Second, we examined the fold changes between model

21   predictions and experimental measurements. These fold change errors are summarized in Figure

22   2B by their mean and variance across the 80 observations. Overall, fold change error varied

23   significantly across amino acids. For example, the model predicted some essential amino acids

3

1 consistently well – e.g. phenylalanine, cysteine and tryptophan (fold change ≈ 1), but predicted

2 others poorly – e.g. alanine, lysine, glycine, and methionine (fold change ≈ 0). Overall, the

3 sizeable fold change errors for many amino acids confirm the difficulty of using metabolic

4 network models alone to predict amino acid consumption.

5   Notably, the model systematically underestimated consumption rates for almost all amino

6 acids (fold change < 1). This is likely because the model doesn't consider certain metabolic

7 inefficiencies – e.g. CHO cells consume more amino acids than needed for the observed

8 production of biomass and recombinant protein, and catabolize them as byproducts[20].

9 Furthermore, the variance of fold change error was relatively low (≤1) for most amino acids.

10 This suggests that the difference between model ideality and biological reality remained

11 consistent across many clones and conditions.

12   We hypothesized that this consistent gap could be bridged with data and statistical

13 modeling. We constructed a series of linear regression models to 'correct' the predictions from

14 metabolic modeling, using growth rate and the predictions from the metabolic model as

15 explanatory variables (Methods, equation 2). The 80 observations were randomly divided into a

16 training dataset and validation dataset, consisting of 48 and 32 observations, respectively. The

17 regression coefficients were first estimated from the training dataset and then applied to the

18 validation dataset. According to validation results, the regression models substantially improved

19 predictions, as fold change error approached unity for most amino acids (Fig. 3B). As exceptions,

20 predictions for alanine, glycine and histidine were not reliably improved (Fig. 3, red). These

21 results were replicated in additional validation studies involving four distinct clones

22 (Supplementary Document).

1   These results show that our hybrid modeling approach estimates amino acid

2   consumptions well for a small timescale of 1 day, when the steady state assumption holds true.

3   This assumption is not valid at larger timescales of multiple days, where nutrient consumption

4   declines asymptotically as cellular metabolism shifts from exponential growth phase to

5   stationary phase. However, we found this limitation could be remedied by modeling the multi-

6   phase decline in amino acid consumption with a simple sigmoid function (Methods, equation 3;

7   line in Fig. 4), which can be fitted from only a few datapoints. Specifically, we further adapted

8   our hybrid modeling approach by first predicting amino acid consumption rates of several early

9   culture days as heretofore described. Then, these datapoints were used to fit a sigmoid function

10   that described the entire consumption profile, including later culture days (Fig. 4A). Using this

11   approach, we accurately predicted the time-course consumption rates of 13 out of 18 amino acids

12   (Spearman $\rho > 0.65$; Fig. 4B), with only few amino acids remaining difficult to predict (alanine,

13   glycine, and histidine). Notably, our approach accurately predicted the consumption profiles of

14   amino acids that are highly abundant in recombinant antibodies (e.g. serine, valine, and

15   leucine)[21], or that complicate media formulation due to low solubility (e.g. tyrosine). These

16   results highlight the method's value in monitoring and forecasting the bioreactor environment.

17   In summary, the presented modeling workflow forecasted the entire amino acid

18   consumption profile from early bioprocess measurements, facilitating anticipatory and *in situ*

19   control of bioreactor nutrient availability. This was realized by a novel combination of metabolic

20   and statistical models. A metabolic network model estimated amino acid uptake rates necessary

21   for observed proliferation and productivity, assuming an ideally efficient metabolism and steady

22   state conditions. Two subsequent regression models refined these predictions by offsetting

23   prediction errors empirically and by describing the time-course relationship of individual

1   predictions. Our efforts are part of a growing trend of synergizing metabolic network models

2   with machine learning methods[22], and demonstrates the power of hybrid modeling for on-line

3   control of bioprocesses.

4

Methods

Cell culture experiments

7   Two production fed batch processes were used, Fed batch 1 and Fed batch 2. Both fed

8   batch processes used chemically defined media and feeds over the 12-day cell culture. Fed batch

9   1 used a glucose restricted fed batch process called HiPDOG[23]. Glucose concentration is kept

10   low during the initial phase of the process, Day 2-7, through intermittent addition of feed

11   medium containing glucose at the high end of pH dead-band and then glucose was maintained

12   above 1.5 g/L thereafter. These conditions help restrict lactate production in fed batch cultures

13   without compromising the proliferative capability of cells. In Fed batch 2 a conventional cell

14   culture process was used where glucose was maintained above 1.5 g/L throughout the process.

15   For both process conditions, bioreactor vessels were inoculated at 2 x 10^6 viable cells/mL.

16   The following bioprocess characteristics were quantified daily using a NOVA Flex BioProfile

17   Analyzer (Nova Biomedical, Waltham, MA): viable cell density, average live cell diameter and

18   concentrations of glucose, lactate, glutamate, and glutamine. Viable cell density data were

19   converted to growth rates by following equation to be compared to model-predicted growth rates.

20         $$(1)\ Growth\ rate\ =\ \frac{1}{vcd} \cdot \frac{\Delta vcd}{\Delta time}$$

1   Flash-frozen cell pellets (10^6 cells) and supernatant (1 mL) were collected from

2 bioreactor runs for each sampling day. Collected samples were sent to Metabolon (Metabolon

3 Inc, Morrisville, NC) for metabolomics analyses. Proteins were removed by methanol

4 precipitation and the metabolites were recovered by vigorous shaking and centrifugation. The

5 extracted samples were run for reverse-phase Ultrahigh Performance Liquid Chromatography-

6 Tandem Mass Spectrometry with negative ion mode ESI. Raw data was extracted, peak-

7 identified and processed for quality control using Metabolon's hardware and software. The raw

8 ion count data was normalized against the extracted proteins quantified using a Bradford assay.

9 These measurements were used as input data to the model by converting their units to model

10 units of mmol per gram of dry weight of cell per hour.

11

12 **Metabolic network modeling**

13   We used a previously described metabolic network model that is tailored to the

14 investigated CHO clones[18]. Experimental measurements for clone and culture day were used to

15 constrain model reactions for biomass production, monoclonal antibody secretion and

16 consumption of glucose, lactate, glutamate, and glutamine. Then, we computed distributions of

17 likely amino acid consumption rates by stochastically sampling 5000 points within the model's

18 solution space via a Markov chain Monte Carlo sampling algorithm, as described previously[24,25],

19 using *optGpSampler*[25] and COBRApy[26]. A set of non-uniform 'points' or flux values was

20 generated within the feasible flux space. Each point was subsequently moved randomly, while

21 remaining within the feasible flux space. To do this, a random direction was first chosen. Second,

22 the limit for how far the point can travel in the randomly-chosen direction was calculated. Lastly,

23 a new random point on this line was selected. This process was iterated until the set of points

1   approached a uniform sample of the solution space. Upon completion, the sampled distributions'

2   statistical features were noted – that is, their mean, median, standard deviation, 25 percentile, and

3   75 percentile values.

4

5   **Statistical methods**

6   For each amino acid, the mean of the sample distribution was interpreted as the likely

7   consumption rates. These predicted consumption rates deviated from experimental observations

8   by a consistent fold amount. Fold change error was also correlated with culture day, as the model

9   predicted the exponential growth phase better than the subsequent stationary phase. Therefore,

10   the model predictions were refined by a regression model as follows, with growth rate and the

11   predictions themselves as explanatory variables.

12         $(2)\ Corrected\ prediction = \ \beta_0 + \beta_1 \cdot prediction + \beta_2 \cdot growth\ rate$

13   The time-course amino acid consumption profiles were described mathematically by the

14   Monod equation, as follows:

15         $(3)\ Consumption\ rate = \ \beta_0 \cdot \dfrac{time}{\beta_1 + time}$

16   Here, β_0 represents the minimum consumption rate which the cells asymptotically

17   approach during later stationary phase. The variable β_1 is the half-velocity constant, or the time

18   point at which the consumption rate reaches half of β_0. These analyses were carried out and

19   visualized using COBRA Toolbox 2.0[19] in MATLAB R2018b (MathWorks; Natick,

20   Massachusetts, USA)

References

1. Ritacco FV, Wu Y, Khetan A. Cell culture media for recombinant protein expression in Chinese hamster ovary (CHO) cells: History, key components, and optimization strategies. *Biotechnology Progress*. 2018;34(6):1407-1426. doi:10.1002/btpr.2706

2. Duarte TM, Carinhas N, Barreiro LC, Carrondo MJT, Alves PM, Teixeira AP. Metabolic responses of CHO cells to limitation of key amino acids. *Biotechnology and Bioengineering*. 2014;111(10):2095-2106. doi:10.1002/bit.25266

3. Fomina☐Yadlin D, Gosink JJ, McCoy R, et al. Cellular responses to individual amino-acid depletion in antibody-expressing and parental CHO cell lines. *Biotechnology and Bioengineering*. 2014;111(5):965-979. doi:10.1002/bit.25155

4. Bhatia H, Mehdizadeh H, Drapeau D, Yoon S. In-line monitoring of amino acids in mammalian cell cultures using raman spectroscopy and multivariate chemometrics models. *Engineering in Life Sciences*. 2018;18(1):55-61. doi:10.1002/elsc.201700084

5. Hefzi H, Ang KS, Hanscho M, et al. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Systems*. 2016;3(5):434-443.e8. doi:10.1016/j.cels.2016.10.020

6. Calmels C, McCann A, Malphettes L, Andersen MR. Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. *Metab Eng*. 2019;51:9-19. doi:10.1016/j.ymben.2018.09.009

7. Huang Z, Yoon S. Integration of Time-Series Transcriptomic Data with Genome-Scale CHO Metabolic Models for mAb Engineering. *Processes*. 2020;8(3):331. doi:10.3390/pr8030331

8. Zhang C, Hua Q. Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Front Physiol*. 2016;6. doi:10.3389/fphys.2015.00413

9. Sommeregger W, Sissolak B, Kandra K, von Stosch M, Mayer M, Striedner G. Quality by control: Towards model predictive control of mammalian cell culture bioprocesses. *Biotechnol J*. 2017;12(7). doi:10.1002/biot.201600546

10. Popp O, Müller D, Didzus K, et al. A hybrid approach identifies metabolic signatures of high-producers for chinese hamster ovary clone selection and process optimization. *Biotechnology and Bioengineering*. 2016;113(9):2005-2019. doi:10.1002/bit.25958

11. Zhuangrong H, Seongkyu Y. Identifying metabolic features and engineering targets for productivity improvement in CHO cells by integrated transcriptomics and genome-scale metabolic model. *Biochemical Engineering Journal*. Published online May 5, 2020:107624. doi:10.1016/j.bej.2020.107624

9

12. Fouladiha H, Marashi S-A, Torkashvand F, Mahboudi F, Lewis NE, Vaziri B. A metabolic network-based approach for developing feeding strategies for CHO cells to increase monoclonal antibody production. *bioRxiv*. Published online 2020:751347. doi:10.1101/751347

13. Traustason B, Cheeks M, Dikicioglu D. Computer-Aided Strategies for Determining the Amino Acid Composition of Medium for Chinese Hamster Ovary Cell-Based Biomanufacturing Platforms. *International Journal of Molecular Sciences*. 2019;20(21):5464. doi:10.3390/ijms20215464

14. Chen Y, McConnell BO, Gayatri Dhara V, et al. An unconventional uptake rate objective function approach enhances applicability of genome-scale models for mammalian cells. *NPJ Syst Biol Appl*. 2019;5. doi:10.1038/s41540-019-0103-6

15. Feist AM, Palsson BO. The biomass objective function. *Curr Opin Microbiol*. 2010;13(3):344-349. doi:10.1016/j.mib.2010.03.003

16. Szeliova D, Ruckerbauer D, Galleguillos S, et al. What CHO is made of: Variations in the biomass composition of Chinese hamster ovary cell lines. *Metabolic Engineering*. Published online 2020.

17. Savinell JM, Palsson BO. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J Theor Biol*. 1992;154(4):421-454. doi:10.1016/s0022-5193(05)80161-4

18. Schinn S-M, Morrison C, Wei W, Zhang L, Lewis N. Systematic evaluation of parameterization for genome-scale metabolic models of cultured mammalian cells. Published online 2020.

19. Schellenberger J, Que R, Fleming RMT, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc*. 2011;6(9):1290-1307. doi:10.1038/nprot.2011.308

20. Mulukutla BC, Kale J, Kalomeris T, Jacobs M, Hiller GW. Identification and control of novel growth inhibitors in fed-batch cultures of Chinese hamster ovary cells. *Biotechnology and Bioengineering*. 2017;114(8):1779-1790. doi:10.1002/bit.26313

21. Fan Y, Val IJD, Müller C, et al. Amino acid and glucose metabolism in fed-batch CHO cell culture affects antibody production and glycosylation. *Biotechnology and Bioengineering*. 2015;112(3):521-535. doi:10.1002/bit.25450

22. Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput Biol*. 2019;15(7). doi:10.1371/journal.pcbi.1007084

23. Gagnon M, Hiller G, Luan Y-T, Kittredge A, DeFelice J, Drapeau D. High-End pH-controlled delivery of glucose effectively suppresses lactate accumulation in CHO Fed-

1   batch cultures. *Biotechnology and Bioengineering*. 2011;108(6):1328-1337.
2   doi:10.1002/bit.23072

3   24. Nam H, Lewis NE, Lerman JA, et al. Network Context and Selection in the Evolution to
4   Enzyme Specificity. *Science*. 2012;337(6098):1101-1104. doi:10.1126/science.1216861

5   25. Megchelenbrink W, Huynen M, Marchiori E. optGpSampler: An Improved Tool for
6   Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks. *PLOS*
7   *ONE*. 2014;9(2):e86587. doi:10.1371/journal.pone.0086587

8   26. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based
9   Reconstruction and Analysis for Python. *BMC Systems Biology*. 2013;7(1):74.
10   doi:10.1186/1752-0509-7-74

11

12

1 # Figures
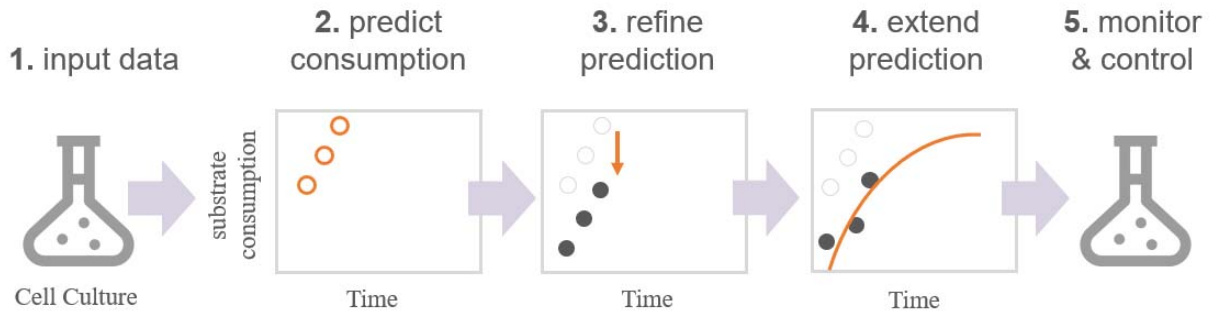
2

3 *Figure 1:* **Overview of method**. A novel combination of a metabolic and statistical models forecast the time-course amino acid
4 consumption profiles in CHO cell cultures, as follows: (1) Routine bioprocess measurements are used as input data. (2) A
5 metabolic model initially estimates early amino acid consumption rates.(3) A regression model refines these predictions. (4)
6 These refined predictions are fit to a curve describing the time-course profile. (5) This would allow for anticipatory control of
7 amino acid availability in bioreactors.

8
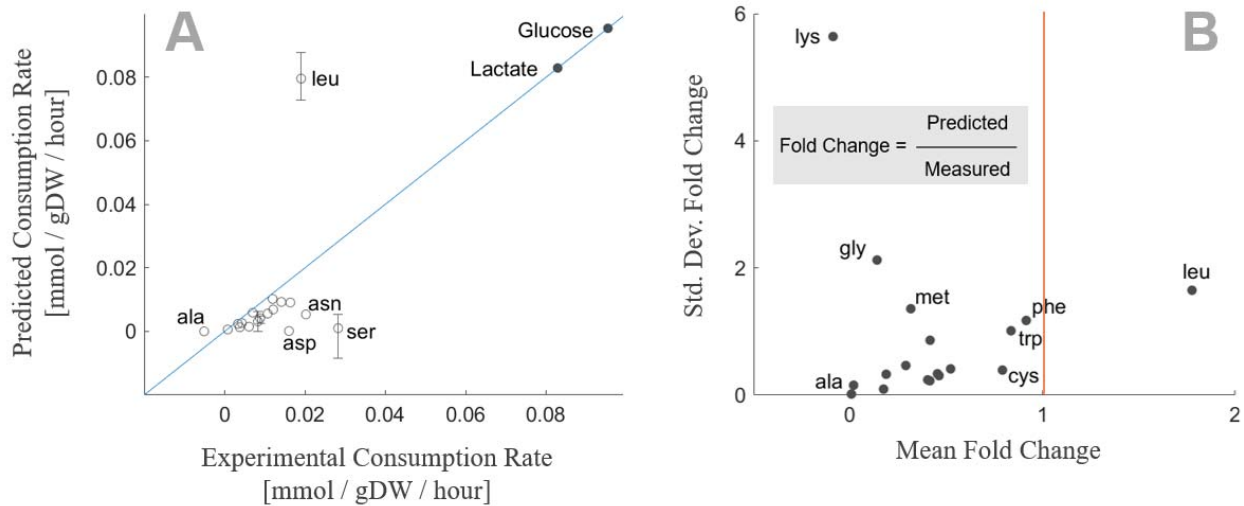
9

10 *Figure 2:* **Metabolic network model shows moderate accuracy in estimating amino acid consumption**. (a) Model predictions
11 compared well to experimental observations, given the scale of input data such as the consumption rates of glucose and lactate
12 (upper right, filled circles). (b) The fold change of model predictions and experimental measurements was also explored. The
13 mean and variance of fold change across all 10 clones and 8 timepoints are shown. Prediction accuracy was particularly good for
14 phenylalanine, tryptophan and cysteine, whose fold change approached unity (red line, *x*-axis). However, for many amino acids,
15 model predictions were prone to significant fold change errors. Notably, the relatively low variance in fold change error (*y*-axis)
16 suggests that predictions could be improved empirically.

1
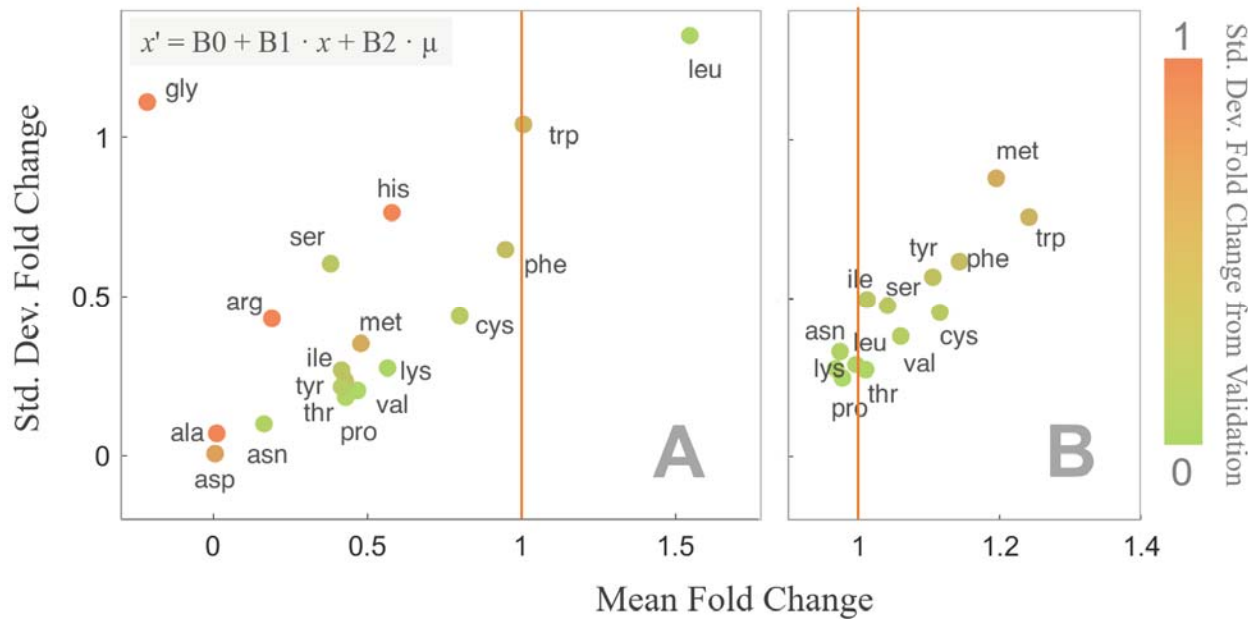
2

3   *Figure 3:* **Statistical learning refines predictions of the metabolic model.** A regression model was devised for each amino acid.
4   The variables μ, x, and x' represent growth rate, predicted consumption rate and revised consumption rate, respectively. The
5   regression coefficients were fitted from the training dataset. (a) Here we show the prediction qualities by fold change for a
6   validation dataset *before* refinement by regression. The mean and variance of fold changes are comparable to the entire dataset
7   (Figure 2b). (b) The prediction values were transformed by the regression model. Prediction of nearly all amino acids approach
8   unity (red line). The datapoints are colored by the variance in fold change after transformation. Amino acids with high variance in
9   fold change (y ≥ 1, red) – alanine, glycine, arginine and histidine – failed to be reliably corrected by the regression model.
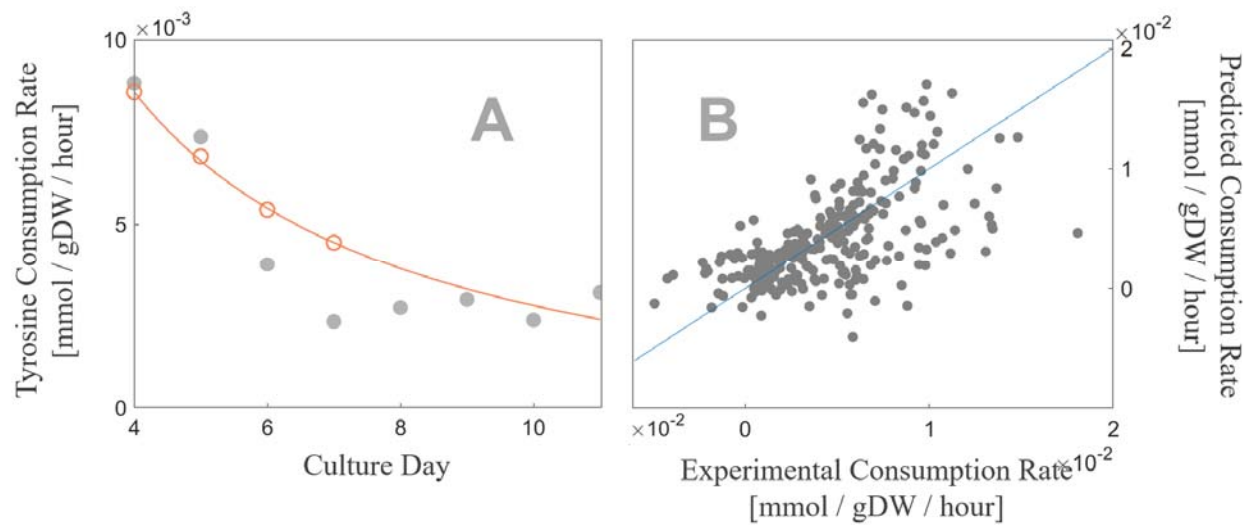
10

11

12

13

1

2

3 *Figure 4:* **Curve fitting forecasts amino acid profiles**. (A) Bioprocess data from days 4-7 are used to estimate tyrosine
4 consumption rates (empty circles). These estimations are used to parameterize a sigmoid curve (line; see Methods, equation 3)
5 that describes the consumption profile for the entire culture duration. This predicted time-course consumption profile agreed well
6 with experimental measurements (filled circle). Time-course profiles of other amino acids are provided in Supplementary Figures.
7 (B) Overall, model predictions and experimental measurements agreed fairly well for all amino acids (Spearman $\rho = 0.54$).

8