# A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese Hamster Ovary cell cultures

Song-Min Schinn[1], Carly Morrison[2], Wei Wei[2], Lin Zhang[2], and Nathan Lewis[3]

[1]University of California San Diego
[2]Pfizer Inc
[3]University of California, San Diego

September 11, 2020

**Abstract**

The control of nutrient availability is critical to large-scale manufacturing of biotherapeutics. However, the quantification of proteinogenic amino acids is time-consuming and thus is difficult to implement for real-time in situ bioprocess control. Genome-scale metabolic models describe the metabolic conversion from media nutrients to proliferation and recombinant protein production, and therefore are a promising platform for in silico monitoring and prediction of amino acid concentrations. This potential has not been realized due to unresolved challenges: (1) the models assume an optimal and highly efficient metabolism, and therefore tend to underestimate amino acid consumption, and (2) the models assume a steady state, and therefore have a short forecast range. We address these challenges by integrating machine learning with the metabolic models. Through this we demonstrate accurate and time-course dependent prediction of individual amino acid concentration in culture medium throughout the production process. Thus, these models can be deployed to control nutrient feeding to avoid premature nutrient depletion or provide early predictions of failed bioreactor runs.

## Short Communication

Chinese Hamster Ovary (CHO) cells are widely used to manufacture complex biotherapeutic molecules at large scales. Industrial bioprocesses ensure high product yield and quality by maintaining favorable growth conditions in cell culture environments, which requires careful monitoring and control of nutrient availability. Chemically-defined serum-free media can contain dozens or >100 components(Ritacco et al., 2018), but key nutrients include proteinogenic amino acids, which are direct substrates and regulators(Duarte et al., 2014; Fomina-Yadlin et al., 2014) of proliferation and protein synthesis. Unfortunately, conventional methods for amino acid quantification based on liquid chromatography and mass spectrometry are time-consuming and difficult to use for decision making and control of cell culture. Alternate spectroscopic approaches have been sensitive to a limited number of amino acid species(Bhatia et al., 2018). Here we present a computational method to forecast time-course amino acid concentrations from routine bioprocess measurements, facilitating a timely and anticipatory control of the bioprocess (Figure 1).

At the foundation of our method is a genome-scale metabolic network model, which accounts for the complex conversion from media nutrients to biomass and recombinant protein production. Such models have been increasingly utilized for CHO cells(Hefzi et al., 2016; Calmels et al., 2019; Huang & Yoon, 2020) and bioprocess applications(Sommeregger et al., 2017; Zhang & Hua, 2016), such as predicting clonal performances(Popp et

al., 2016), identifying metabolic bottlenecks(Zhuangrong & Seongkyu, 2020), and optimizing media formulation(Fouladiha et al., 2020; Traustason et al., 2019). Metabolic network models can also estimate amino acid uptake rates necessary to experimentally support observed proliferation and productivity(Chen et al., 2019). However, challenges have limited their practical application.

First, metabolic network models are typically highly complex but under-constrained, and therefore are easy to overfit. This is mitigated by training the model on a variety of bioprocess conditions and metabolic phenotypes. Second, metabolic network models assume that cells operate at some metabolic optimum, and thus tend to describe an idealized metabolism specifically fit to the assumed objective (e.g., biomass production(Feist & Palsson, 2010; Szeliova et al., 2020), minimization of redox(Savinell & Palsson, 1992)). Third, for the present purpose, these models need to predict amino acid consumption fluxes, typically on the order of $10^{-3}$mmol·$g_{DW}^{-1}$·$hr^{-1}$(see Methods), from input data that are multiple magnitudes larger, such as growth rate and glucose consumption ($10^{-1}$ to $10^{-2}$mmol·$g_{DW}^{-1}$·$hr^{-1}$). The preceding two challenges increase prediction error. Lastly, metabolic network models assume a steady state, which reduces the range of forecast. Typically, input data from one day are used to make predictions for the same day. However, such predictions cannot be extended to multiple days or subsequent culture phases, as cross-temporal shifts in metabolism would violate the steady state assumption. Thus, model predictions of amino acid concentrations can be overfit, ideal, and near-sighted – all of which dilutes their practicality for industrial bioprocess control. Here we demonstrate that these weaknesses can be addressed in a data-driven manner by coupling a metabolic network model with machine learning.

We developed this hybrid approach on a diverse set of 10 CHO clones with different growth and productivity profiles from two different fed-batch production processes. These CHO clones were subject to different bioprocess conditions and recombinant antibody identities (see Methods), resulting in a variety of phenotypes and productivity performances (Fig. S1). For example, several high-performing clones were exceptionally proliferative or productive, suggesting an efficient conversion from nutrients to biomass or recombinant protein product. Other clones performed these conversions at lower rates, suggesting attenuated metabolic activity or inefficient resource utilization. The CHO cells adjusted their nutrient uptake according to these various metabolic phenotypes, leading to diverse amino acid consumption patterns (Fig, S2). For example, the consumption of glucose and serine differed by several fold across conditions and time. Furthermore, different clones varied in their consumption or secretion of key metabolites such as lactate, alanine, glycine, and glutamine.

We sought to predict these diverse consumption behaviors using a tailored model of CHO metabolism(Schinn et al., 2020). As input information, we utilized the following routinely measured industrial bioprocess data: (1) viable cell density and titer measurements, from which growth rate and specific productivity are calculated (Methods, equation 1), and (2) bioreactor concentrations of glucose, lactate, glutamate and glutamine, from which their respective consumption rates are calculated. These measurements were used as boundary conditions by constraining the fluxes of biomass production, recombinant protein synthesis and consumption of the four metabolites to observed values. Subsequently, we used Markov chain Monte Carlo sampling of metabolic fluxes(Schellenberger et al., 2011) to sample the range and magnitude of all reaction fluxes to calculate the likely uptake fluxes of the remaining 18 proteinogenic amino acids (see Methods). These predictions were applied to the CHO clones across 8 days of a 12-day production run (days 4 to 11), resulting in a total of 80 individual predictions.

We evaluated the resulting model predictions in two ways. First, we examined the differences in model predictions and experimental measurements of amino acid uptake and secretion (Figure 2A). For most amino acids, this difference was small compared to the scale of input data, suggesting that metabolic models can describe the conversion from nutrients to biomass and recombinent proteins. Second, we examined the fold changes between model predictions and experimental measurements. These fold change errors are summarized in Figure 2B by their mean and variance across the 80 observations. Overall, fold change error varied significantly across amino acids. For example, the model predicted some essential amino acids consistently well – e.g. phenylalanine, cysteine and tryptophan (fold change [?] 1), but predicted others poorly – e.g.

alanine, lysine, glycine, and methionine (fold change [?] 0). Overall, the sizeable fold change errors for many amino acids confirm the difficulty of using metabolic network models alone to predict amino acid consumption.

Notably, the model systematically underestimated consumption rates for almost all amino acids (fold change $< 1$). This is likely because the model doesn't consider certain metabolic inefficiencies – e.g. CHO cells consume more amino acids than needed for the observed production of biomass and recombinant protein, and catabolize them as byproducts(Mulukutla et al., 2017). Furthermore, the variance of fold change error was relatively low ([?]1) for most amino acids. This suggests that the difference between model ideality and biological reality remained consistent across many clones and conditions.

We hypothesized that this consistent gap could be bridged with data and statistical modeling. We constructed a series of linear regression models to 'correct' the predictions from metabolic modeling, using growth rate and the predictions from the metabolic model as explanatory variables (Methods, equation 2). The 80 observations were randomly divided into a training dataset and validation dataset, consisting of 48 and 32 observations, respectively. The regression coefficients were first estimated from the training dataset and then applied to the validation dataset. According to validation results, the regression models substantially improved predictions, as fold change error approached unity for most amino acids (Fig. 3B). As exceptions, predictions for alanine, glycine and histidine were not reliably improved (Fig. 3, red). These results were replicated in additional validation studies involving four distinct clones (Supplementary Document).

These results show that our hybrid modeling approach estimates amino acid consumptions well for a small timescale of 1 day, when the steady state assumption holds true. This assumption is not valid at larger timescales of multiple days, where nutrient consumption declines asymptotically as cellular metabolism shifts from exponential growth phase to stationary phase. However, we found this limitation could be remedied by modeling the multi-phase decline in amino acid consumption with a simple sigmoid function (Methods, equation 3; line in Fig. 4), which can be fitted from only a few datapoints. Specifically, we further adapted our hybrid modeling approach by first predicting amino acid consumption rates of several early culture days as heretofore described. Then, these datapoints were used to fit a sigmoid function that described the entire consumption profile, including later culture days (Fig. 4A). Using this approach, we accurately predicted the time-course consumption rates of 13 out of 18 amino acids (Spearman $\rho > 0.65$; Fig. 4B), with only few amino acids remaining difficult to predict (alanine, glycine, and histidine). Notably, our approach accurately predicted the consumption profiles of amino acids that are highly abundant in recombinant antibodies (e.g. serine, valine, and leucine)(Fan et al., 2015), or that complicate media formulation due to low solubility (e.g. tyrosine). These results highlight the method's value in monitoring and forecasting the bioreactor environment.

In summary, the presented modeling workflow forecasted the entire amino acid consumption profile from early bioprocess measurements, facilitating anticipatory and *in situ* control of bioreactor nutrient availability. This was realized by a novel combination of metabolic and statistical models. A metabolic network model estimated amino acid uptake rates necessary for observed proliferation and productivity, assuming an ideally efficient metabolism and steady state conditions. Two subsequent regression models refined these predictions by offsetting prediction errors empirically and by describing the time-course relationship of individual predictions. Our efforts are part of a growing trend of synergizing metabolic network models with machine learning methods(Zampieri et al., 2019), and demonstrates the power of hybrid modeling for on-line control of bioprocesses.

# Methods

### Cell culture experiments

Two production fed batch processes were used, Fed batch 1 and Fed batch 2. Both fed batch processes used chemically defined media and feeds over the 12-day cell culture. Fed batch 1 used a glucose restricted fed

batch process called HiPDOG(Gagnon et al., 2011). Glucose concentration is kept low during the initial phase of the process, Day 2-7, through intermittent addition of feed medium containing glucose at the high end of pH dead-band and then glucose was maintained above 1.5 g/L thereafter, restricting lactate production without compromising the proliferative capability of cells. In Fed batch 2 a conventional cell culture process was used where glucose was maintained above 1.5 g/L throughout the process.

For both process conditions, bioreactor vessels were inoculated at 2 x $10^6$ viable cells/mL. The following bioprocess characteristics were quantified daily using a NOVA Flex BioProfile Analyzer (Nova Biomedical, Waltham, MA): viable cell density, average live cell diameter and concentrations of glucose, lactate, glutamate, and glutamine. Viable cell density data were converted to growth rates by following equation to be compared to model-predicted growth rates.

$$\text{Growth rate} = \frac{1}{\text{vcd}} \frac{\Delta \text{vcd}}{\Delta \text{time}}$$

Flash-frozen cell pellets ($10^6$ cells) and supernatant (1 mL) were collected from bioreactor runs for each sampling day. Collected samples were sent to Metabolon (Metabolon Inc, Morrisville, NC) for metabolomics analyses. Metabolomics measurements were used as input data to the model by converting their units to model units of mmol per gram of dry weight of cell per hour.

### Metabolic network modeling

We used a previously described metabolic network model that is tailored to the investigated CHO clones(Schinn et al., 2020). Experimental measurements for clone and culture day were used to constrain model reactions for biomass production, monoclonal antibody secretion and consumption of glucose, lactate, glutamate, and glutamine. Then, we computed distributions of likely amino acid consumption rates by stochastically sampling 5000 points within the model's solution space via a Markov chain Monte Carlo sampling algorithm, as described previously(Megchelenbrink et al., 2014; Nam et al., 2012), using *optGpSampler* (Megchelenbrink et al., 2014) and COBRApy(Ebrahim et al., 2013). Upon completion, the sampled distributions' statistical features were noted – that is, their mean, median, standard deviation, 25 percentile, and 75 percentile values.

### Statistical methods

For each amino acid, the mean of the sample distribution was interpreted as the likely consumption rates. These predicted consumption rates deviated from experimental observations by a consistent fold amount. Fold change error was also correlated with culture day, as the model predicted the exponential growth phase better than the subsequent stationary phase. Therefore, the model predictions were refined by a regression model as follows, with growth rate and the predictions themselves as explanatory variables.

$$\text{Corrected prediction} = \beta_0 + \beta_1 \bullet \textit{prediction} + \beta_2 \bullet \textit{growth rate}$$

The time-course amino acid consumption profiles were described mathematically by the Monod equation, as follows:

$$\text{Consumption rate} = \beta_0 \bullet \frac{\text{time}}{\beta_1 + \textit{time}}$$

Here, $\beta_0$ represents the minimum consumption rate which the cells asymptotically approach during later stationary phase. The variable $\beta_1$ is the half-velocity constant, or the time point at which the consumption rate reaches half of $\beta_0$. These analyses were carried out and visualized using COBRA Toolbox 2.0(Schellenberger et al., 2011) in MATLAB R2018b (MathWorks; Natick, Massachusetts, USA)

# References

Bhatia, H., Mehdizadeh, H., Drapeau, D., & Yoon, S. (2018). In-line monitoring of amino acids in mammalian cell cultures using raman spectroscopy and multivariate chemometrics models. *Engineering in Life Sciences*

, *18* (1), 55–61. https://doi.org/10.1002/elsc.201700084

Calmels, C., McCann, A., Malphettes, L., & Andersen, M. R. (2019). Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. *Metabolic Engineering* , *51* , 9–19. https://doi.org/10.1016/j.ymben.2018.09.009

Chen, Y., McConnell, B. O., Gayatri Dhara, V., Mukesh Naik, H., Li, C.-T., Antoniewicz, M. R., & Betenbaugh, M. J. (2019). An unconventional uptake rate objective function approach enhances applicability of genome-scale models for mammalian cells. *NPJ Systems Biology and Applications* , *5* . https://doi.org/10.1038/s41540-019-0103-6

Duarte, T. M., Carinhas, N., Barreiro, L. C., Carrondo, M. J. T., Alves, P. M., & Teixeira, A. P. (2014). Metabolic responses of CHO cells to limitation of key amino acids. *Biotechnology and Bioengineering* ,*111* (10), 2095–2106. https://doi.org/10.1002/bit.25266

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python.*BMC Systems Biology* , *7* (1), 74. https://doi.org/10.1186/1752-0509-7-74

Fan, Y., Val, I. J. D., Müller, C., Sen, J. W., Rasmussen, S. K., Kontoravdi, C., Weilguny, D., & Andersen, M. R. (2015). Amino acid and glucose metabolism in fed-batch CHO cell culture affects antibody production and glycosylation. *Biotechnology and Bioengineering* ,*112* (3), 521–535. https://doi.org/10.1002/bit.25450

Feist, A. M., & Palsson, B. O. (2010). The biomass objective function.*Current Opinion in Microbiology* , *13* (3), 344–349. https://doi.org/10.1016/j.mib.2010.03.003

Fomina-Yadlin, D., Gosink, J. J., McCoy, R., Follstad, B., Morris, A., Russell, C. B., & McGrew, J. T. (2014). Cellular responses to individual amino-acid depletion in antibody-expressing and parental CHO cell lines. *Biotechnology and Bioengineering* , *111* (5), 965–979. https://doi.org/10.1002/bit.25155

Fouladiha, H., Marashi, S.-A., Torkashvand, F., Mahboudi, F., Lewis, N. E., & Vaziri, B. (2020). A metabolic network-based approach for developing feeding strategies for CHO cells to increase monoclonal antibody production. *BioRxiv* , 751347. https://doi.org/10.1101/751347

Gagnon, M., Hiller, G., Luan, Y.-T., Kittredge, A., DeFelice, J., & Drapeau, D. (2011). High-End pH-controlled delivery of glucose effectively suppresses lactate accumulation in CHO Fed-batch cultures.*Biotechnology and Bioengineering* , *108* (6), 1328–1337. https://doi.org/10.1002/bit.23072

Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jimenez, N. E., Zielinski, D. C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., ... Lewis, N. E. (2016). A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Systems* , *3* (5), 434-443.e8. https://doi.org/10.1016/j.cels.2016.10.020

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

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

Mulukutla, B. C., Kale, J., Kalomeris, T., Jacobs, M., & Hiller, G. W. (2017). Identification and control of novel growth inhibitors in fed-batch cultures of Chinese hamster ovary cells. *Biotechnology and Bioengineering* , *114* (8), 1779–1790. https://doi.org/10.1002/bit.26313

Nam, H., Lewis, N. E., Lerman, J. A., Lee, D.-H., Chang, R. L., Kim, D., & Palsson, B. O. (2012). Network Context and Selection in the Evolution to Enzyme Specificity. *Science* , *337* (6098), 1101–1104. https://doi.org/10.1126/science.1216861

Popp, O., Muller, D., Didzus, K., Paul, W., Lipsmeier, F., Kirchner, F., Niklas, J., Mauch, K., & Beaucamp, N. (2016). A hybrid approach identifies metabolic signatures of high-producers for chinese hamster ovary clone selection and process optimization. *Biotechnology and Bioengineering* , *113* (9), 2005–2019. https://doi.org/10.1002/bit.25958

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

Savinell, J. M., & Palsson, B. O. (1992). Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of Theoretical Biology* ,*154* (4), 421–454. https://doi.org/10.1016/s0022-5193(05)80161-4

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., & Palsson, B. O. (2011). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nature Protocols* , *6* (9), 1290–1307. https://doi.org/10.1038/nprot.2011.308

Schinn, S.-M., Morrison, C., Wei, W., Zhang, L., & Lewis, N. (2020).*Systematic evaluation of parameterization for genome-scale metabolic models of cultured mammalian cells* .

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

Szeliova, D., Ruckerbauer, D., Galleguillos, S., Petersen, Hanscho, M., Troyer, Causon, Schoeny, Christensen, Lee, D. Y., Lewis, N. E., Koellensperger, Hann, Nielsen, L. K., Borth, N., & Zanghellini, J. (2020). What CHO is made of: Variations in the biomass composition of Chinese hamster ovary cell lines. *Metabolic Engineering* .

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

Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling.*PLoS Computational Biology* , *15* (7). https://doi.org/10.1371/journal.pcbi.1007084

Zhang, C., & Hua, Q. (2016). Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Frontiers in Physiology* , *6* . https://doi.org/10.3389/fphys.2015.00413

Zhuangrong, H., & Seongkyu, Y. (2020). Identifying metabolic features and engineering targets for productivity improvement in CHO cells by integrated transcriptomics and genome-scale metabolic model.*Biochemical Engineering Journal* , 107624. https://doi.org/10.1016/j.bej.2020.107624

# Figures

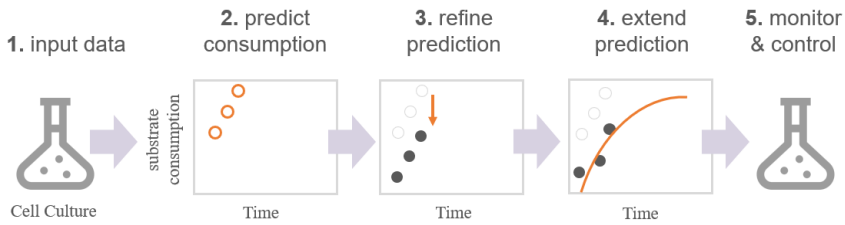

**Figure 1: Overview of method** . A novel combination of a metabolic and statistical models forecast the time-course amino acid consumption profiles in CHO cell cultures, as follows: (1) Routine bioprocess measurements are used as input data. (2) A metabolic model initially estimates early amino acid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

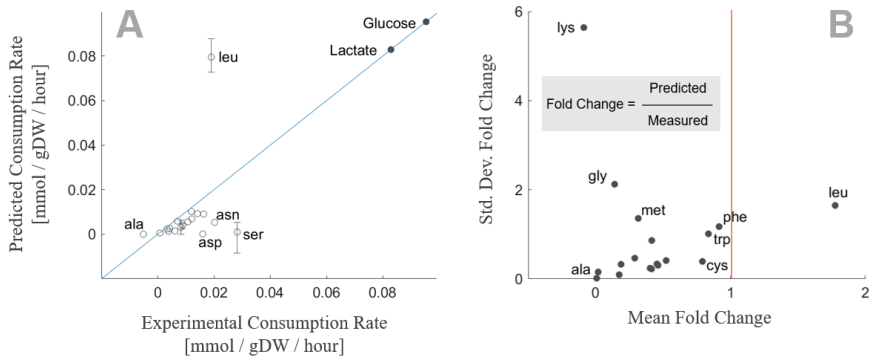


**Figure 2: Metabolic network model shows moderate accuracy in estimating amino acid consumption** . (a) Model predictions compared well to experimental observations, given the scale of input data such as the consumption rates of glucose and lactate (upper right, filled circles). (b) The fold change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

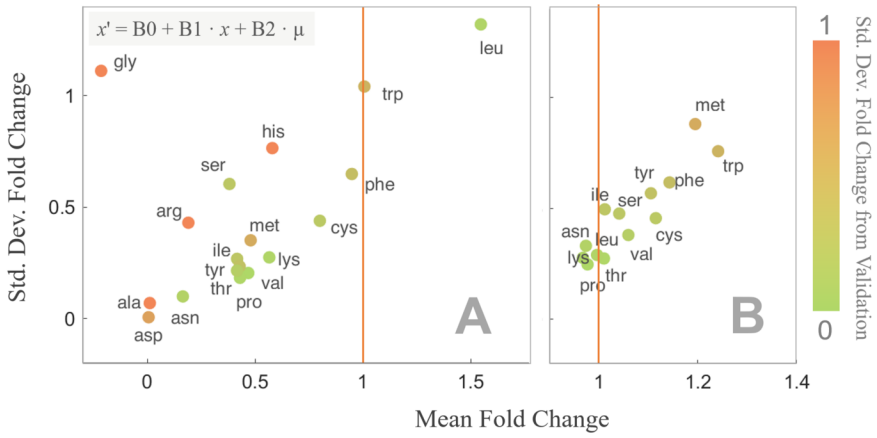


**Figure 3: Statistical learning refines predictions of the metabolic model.** A regression model was devised for each amino acid. The variables μ, x, and x' represent growth rate, predicted consumption rate and revised consumption rate, respectively. The regression coefficients were fitted from the training dataset. (a) Here we show the prediction qualities by fold change for a validation dataset before refinement by regression. The mean and variance of fold changes are comparable to the entire dataset (Figure 2b). (b) The prediction values were transformed by the regression model. Prediction of nearly all amino acids approach unity (red line). The datapoints are colored by the variance in fold change after transformation.

Amino acids with high variance in fold change (y [?] 1, red) – alanine, glycine, arginine and histidine – failed to be reliably corrected by the regression model.

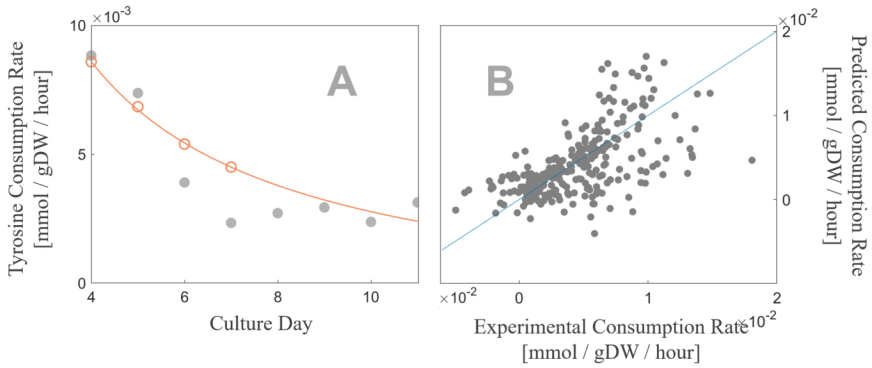


**Figure 4: Curve fitting forecasts amino acid profiles** . (A) Bioprocess data from days 4-7 are used to estimate tyrosine consumption rates (empty circles). These estimations are used to parameterize a sigmoid curve (line; see Methods, equation 3) that describes the consumption profile for the entire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