Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states

Subham Choudhury[1], Bharath Narayanan[1], Michael Moret[1,2], Vassily Hatzimanikatis[1*], Ljubisa Miskovic[1*]

[1] Laboratory of Computational Systems Biology (LCSB), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[2] Present address: Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

[*] **Correspondence:**

Ljubisa Miskovic,
Laboratory of Computational Systems Biotechnology (LCSB),
École Polytechnique Fédérale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
Email: ljubisa.miskovic@epfl.ch
Phone: + 41 (0)21 693 98 92

Vassily Hatzimanikatis,
Laboratory of Computational Systems Biotechnology (LCSB),
École Polytechnique Fédérale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
Email: vassily.hatzimanikatis@epfl.ch
Phone: + 41 (0)21 693 98 70

Abstract

Large omics datasets are nowadays routinely generated to provide insights into cellular processes. Nevertheless, making sense of omics data and determining intracellular metabolic states remains challenging. Kinetic models of metabolism are crucial for integrating and consolidating omics data because they explicitly link metabolite concentrations, metabolic fluxes, and enzyme levels. However, the difficulties in determining kinetic parameters that govern cellular physiology prevent the broader adoption of these models by the research community. We present RENAISSANCE (REconstruction of dyNAmIc models through Stratified Sampling using Artificial Neural networks and Concepts of Evolution strategies), a generative machine learning framework for efficiently parameterizing large-scale kinetic models with dynamic properties matching experimental observations. We showcase RENAISSANCE's capabilities through three applications: generation of kinetic models of *E. coli* metabolism, characterization of intracellular metabolic states, and assimilation and reconciliation of experimental kinetic data. We provide the open-access code to facilitate experimentalists and modelers applying this framework to diverse metabolic systems and integrating a broad range of available data. We anticipate that the proposed framework will be invaluable for researchers who seek to analyze metabolic variations involving changes in metabolite and enzyme levels and enzyme activity in health and biotechnological studies.

Keywords: metabolism, integration of omics data, large-scale and genome-scale kinetic models, machine learning, evolution strategies, *E. coli*, kinetic parameters, nonlinear dynamics.

Abbreviations: ES, Evolution Strategies, **ORACLE,** Optimization and Risk Analysis of Complex Living Entities, **GEM,** GEnome-scale Model, **RENAISSANCE,** REconstruction of dyNAmIc models through Stratified Sampling using Artificial Neural networks and Concept of Evolution strategies, **REKINDLE,** REconstruction of KINetic models of metabolism using Deep LEarning, **TFA**, Thermodynamics-based Flux Balance Analysis, **ODE**, Ordinary Differential Equations

Advancement in biotechnology and health sciences hinges heavily on our capability to integrate different varieties of data produced by high-throughput techniques and obtain coherent insights into cellular processes[1–3]. Considerable effort has been invested in using genome-scale models, mathematical representations of metabolic information about living organisms, to reconcile and make sense of such constantly growing disparate datasets[4,5]. Genome-scale models integrate omics data by considering constraints imposed by genetics and physicochemical laws[6–10]. For instance, researchers use inequality constraints stemming from the second law of thermodynamics to relate metabolic fluxes (fluxome) to metabolite profiles (metabolome)[11–14]. However, data integration using such inequality constraints results in significant uncertainty about intracellular metabolic states[15]. Consequently, despite the availability of large omics datasets, determining the exact intracellular levels of metabolite profiles and metabolic reaction rates with these constraint-based models remains elusive.

Kinetic models of metabolism can address these issues by consolidating several types of omics data, such as metabolomics, fluxomics, transcriptomics, and proteomics, within a common and coherent mathematical framework[16]. Indeed, these models contain information about enzyme kinetics and metabolic regulation, allowing them to explicitly couple metabolite concentrations, metabolic reaction rates, and enzyme levels through mechanistic relations. Additionally, unlike constraint-based models, kinetic models capture time-dependent responses of cellular metabolism. Taken altogether, these models show great promise for addressing complex phenomena in biomedical sciences and biotechnology, such as metabolic reprogramming in the tumor microenvironment and disease[17–19], relationships between cancer, metabolism, and circadian rhythms[20], dynamics of drug absorption and drug metabolism[21], and engineering and modulating cell phenotypes[22–24].

Despite the capacity of kinetic models to reconcile data and identify metabolic features associated with phenotype, the application of these models is somewhat limited[16,25–30].The major challenge in developing kinetic models is the lack of knowledge about the characteristic kinetic parameter values that govern the cellular physiology of the studied organism *in vivo*. Overcoming this requires employing intricate computational procedures and the extensive expertise of researchers, and it is often impractical to build and use these models for studying multiple physiological conditions and large cohorts[31]. Therefore, there is a need for accelerated approaches for parameterizing kinetic models that would allow the broader research community access to these models.

Recent efforts employing new tailor-made parametrization[27] and machine learning[32–34] improved the efficiency of constructing near-genome-scale kinetic models. Nevertheless, challenges remain regarding extensive computational time[27] and the need for training data from traditional kinetic modeling approaches[32–34]. Here, we present RENAISSANCE (REconstruction of dyNAmIc models through Stratified Sampling using Artificial Neural networks and Concepts of Evolution strategies), a machine learning framework that efficiently parameterizes biologically relevant kinetic models of metabolism without requiring training data. RENAISSANCE uses Natural Evolutionary Strategies (NES)[35,36] to optimize a feed-forward neural network for parameterizing kinetic models with desired properties (Figure 1a). This way, it dramatically reduces the extensive computation time required by

traditional kinetic modeling methods, thus allowing its broad utilization for high-throughput dynamical studies of metabolism. We showcase this framework through three studies: (i) generating a population of large-scale dynamic models of *E. coli* metabolism, (ii) characterizing intracellular metabolic states in the *E. coli* metabolic network accurately, and (iii) integrating and reconciling available experimental data.

RENAISSANCE for parameterization of biologically relevant kinetic models

In its conception, RENAISSANCE can parameterize kinetic models to satisfy a broad range of biochemical properties or physiological conditions. For example, it can parameterize models reproducing experimentally observed fermentation curves or drug adsorption patterns. Herein, we use RENAISSANCE to parameterize kinetic models to be consistent with an experimentally observed steady-state. This approach to model construction was introduced within the ORACLE conceptual framework[15,28,34,37–40], which parameterizes kinetic models by unbiased sampling. In contrast, in RENAISSANCE, we leverage machine learning to perform stratified sampling biased toward kinetic models producing metabolic responses over time with timescales[41] matching experimental observations of studied organisms. Due to its capability to bias parameter sampling toward desired model properties, the proposed framework substantially improves model construction efficiency, enabling comprehensive studies of multiple physiological conditions.

In this context, before using RENAISSANCE, we compute a steady-state profile of metabolite concentrations and metabolic fluxes that will be used for parameterization (Methods). To this end, we integrate information about the structural properties of the metabolic network (stoichiometry, regulatory structure, rate laws) as well as available data (metabolomics, fluxomics, thermodynamics, proteomics, and transcriptomics) into the model (Figure 1b, c, Methods). A parameterized kinetic model exhibits highly nonlinear but deterministic responses that depend on the intracellular state determined by the network topology and the integrated data. To capture this nonlinear behavior and determine kinetic parameters, we require function approximators with similar complexity, such as neural networks[32]. In RENAISSANCE, we iteratively optimize the weights of feed-forward neural networks (generators) using NES (Figure 1a) to obtain kinetic parameters leading to biologically relevant kinetic models, meaning that the metabolic responses obtained from these models have experimentally observed dynamics (Methods).

An NES algorithm produces a population of candidate solutions to an optimization problem and assigns a fitness score to each candidate solution (Figure 1a). The algorithm uses the fitness scores of the current solutions to generate the next generation of candidate solutions, which are likely to have better fitness scores than the current generation. The iterative procedure stops as soon as the obtained solutions are satisfactory. Unlike traditional gradient-based deep learning methods that require data to train a neural network, NES requires only a scoring function.

The iterative process in RENAISSANCE consists of four steps (Figure 1d). We start by initializing a population of generators with random weights (step I). We select one generator at a time and, using

multivariate Gaussian noise as input, generate a batch of kinetic parameters consistent with the network structure and integrated data. We then parameterize the kinetic structure of the metabolic network (step II). Next, we evaluate the dynamics of each parameterized model by computing the eigenvalues of its Jacobian and the corresponding dominant time constants (Methods). These quantities allow us to assess if the generated kinetic models have dynamic responses corresponding to experimental observations (valid models) or not (invalid models). Based on this evaluation, we assign a reward to the generator (step III). NES repeats steps II and III for every generator in the

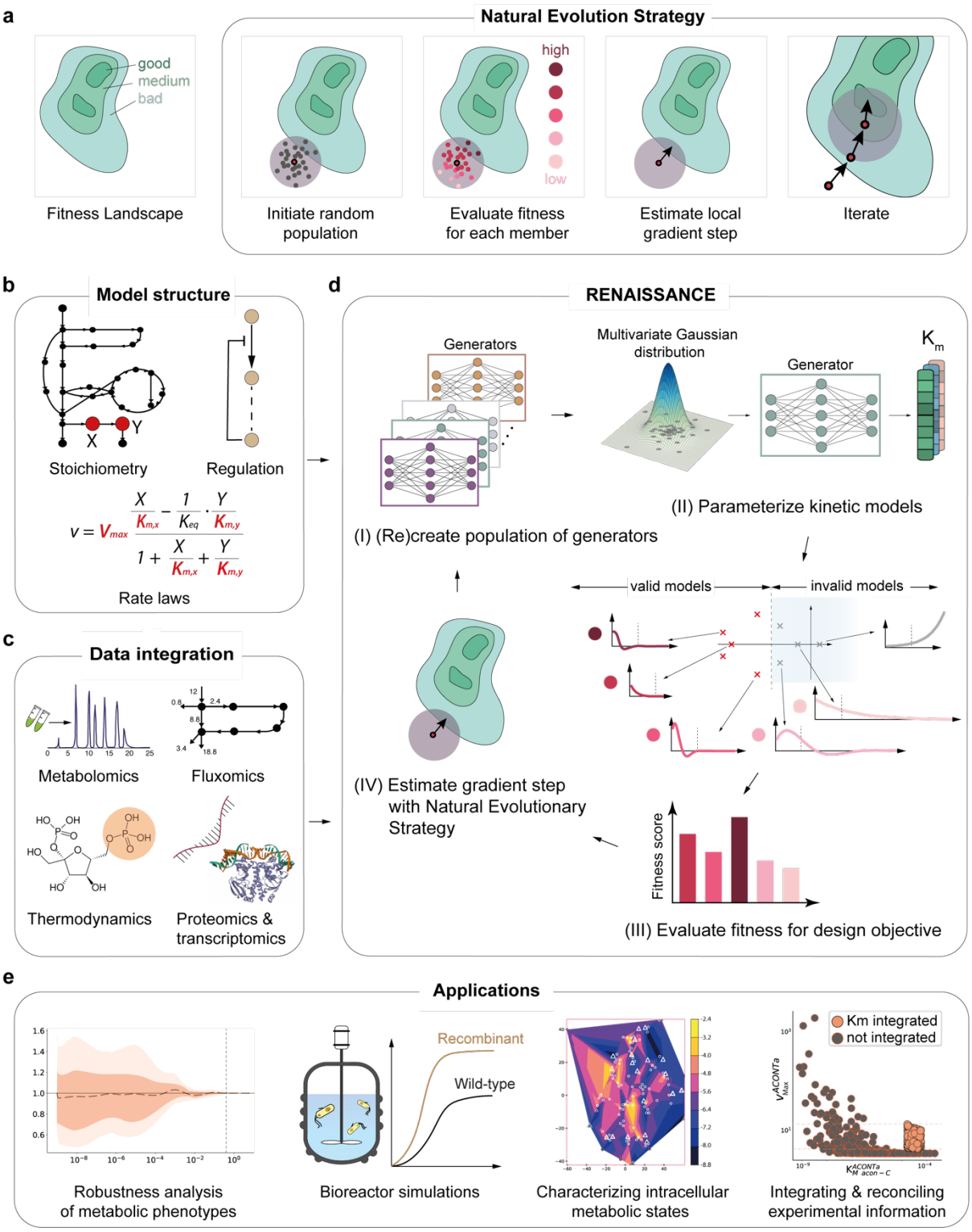

Figure 1 | Overview and applications of the RENAISSANCE framework a): Conceptual representation of the steps of Natural Evolutionary Strategy (NES), an optimization algorithm used in RENAISSANCE. **b)** Context-specific structural properties of the metabolic networks are established and incorporated into the model. **c)** Once the model structure is fixed, available omics data are integrated into the model. **d)** Generators for parameterizing biologically relevant (valid) kinetic models are optimized iteratively in four steps to meet the design objective: a population of generators is randomly initialized (step I); generators produce parameters needed to parameterize kinetic models (step II); the fitness of the kinetic models (circles and bars in shades of red) is assessed based on the largest eigenvalues of the Jacobian (red and grey crosses) corresponding to the dominant time constants of the model responses (Methods); the generator is assigned a score based on this performance (step III); the rewards for each generator are fed back to NES to find the best-performing generator (step IV); the best-performing generator is then perturbed to obtain the next generation of generators (step I). **e)** A few applications of RENAISSANCE-generated models presented in this paper.

population and uses the rewards of the entire population to estimate the local gradient landscape and find the weights of the generator that improve the design objective (step IV). Then, we mutate the obtained generator by injecting random noise in its weights and recreate the new population of generators (step I). We iterate steps I-IV until we obtain a generator that meets the design objective, i.e., it can generate biologically relevant kinetic models (Methods).

The generated kinetic models are applicable to a broad range of metabolism studies. Here, we present a few of these (Figure 1e).

Results

Generating large-scale kinetic models of *E. coli* metabolism

To test and validate RENAISSANCE, we generated biologically relevant kinetic parameter sets for central carbon pathways of *E. coli* metabolism (Methods, Supplementary Note 2). The objective was to find kinetic parameters resulting in dynamic models consistent with an experimentally observed doubling time of 134 minutes for the studied *E. coli* strain[42]. A valid kinetic model satisfying this requirement should produce metabolic responses with the dominant time constant of 24 mins, which corresponds to having the largest eigenvalue $\lambda_{max} < -2.5$ (Methods). The model structure consisted of 113 nonlinear ordinary differential equations (ODEs) parameterized by 502 kinetic parameters, including 384 Michaelis constants, $K_M s$ (Methods, Supplementary Figure 4). To integrate the experimental data[42] and compute a steady-state profile of metabolite concentrations and fluxes, we used Thermodynamically-based flux balance analysis[13] (Methods).

We ran RENAISSANCE for 50 evolution generations. We repeated the optimization process 10 times with a randomly initialized generator population to obtain statistical replicates. At every generation, we generated 100 kinetic parameter sets for every generator in the population and computed the maximum eigenvalue, λ_{max}, for each parameter set. To evaluate the generators, we used the incidence of valid models, defined as the proportion of the generated models that are valid (with $\lambda_{max} < -2.5$, Methods). We observed that the incidence of valid models steadily increases with the number of generations, with the mean incidence converging around 92% after 50 generations (Figure

2a, thick black line). For some repeats, we could achieve incidence up to 100% (Figure 2a, green-shaded region).

For further analysis of the generated models, we selected a statistical repeat with fast convergence (Figure 2a, dashed line) and chose 10 generators from that repeat with monotonically increasing incidence over generations (Figure 2a, black diamonds). For each of the 10 chosen generators, we generated 500 kinetic parameter sets and examined the distribution of the resulting maximum

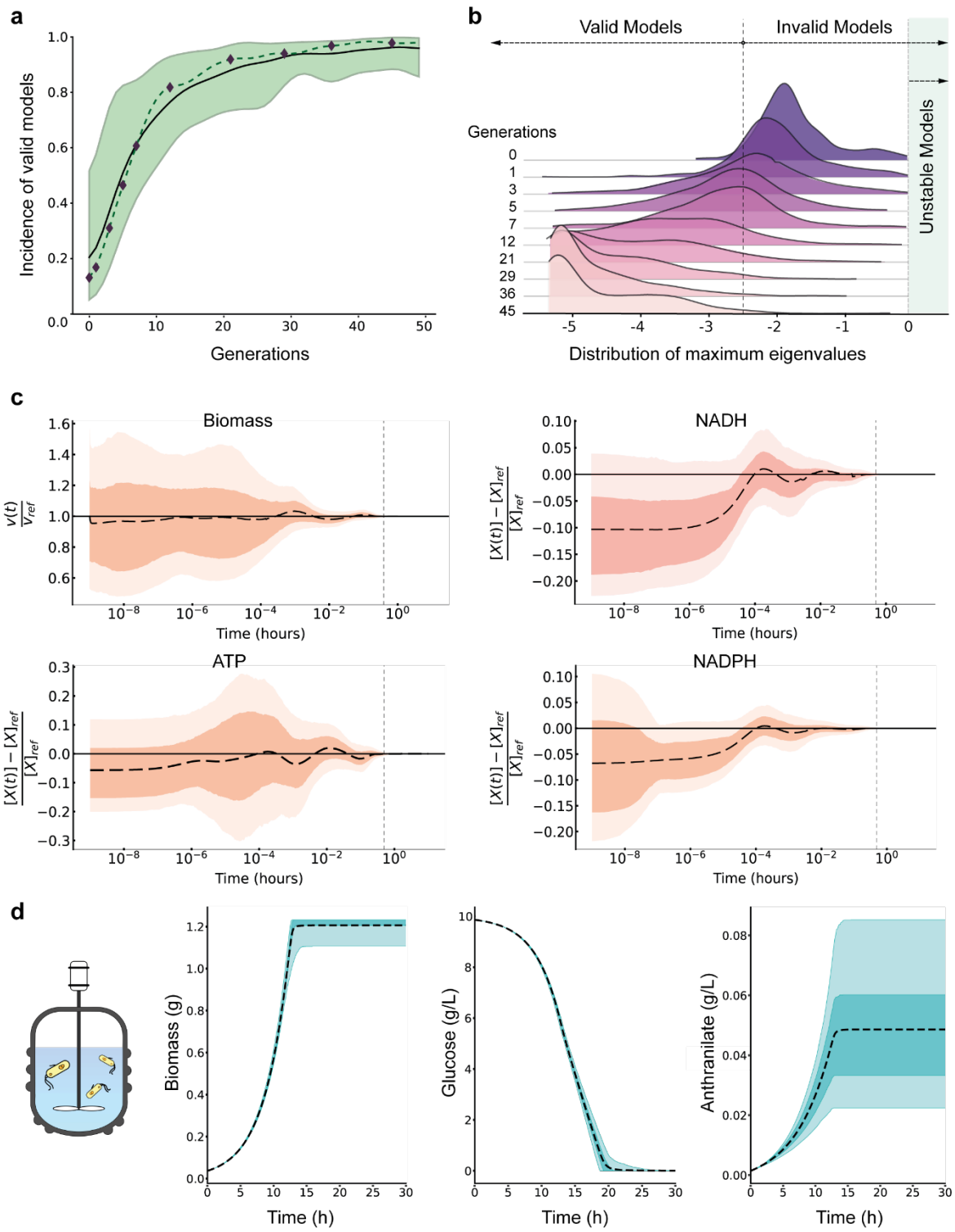

Fig. 2 | Generation, validation, and application of RENAISSANCE-parameterized kinetic models. a) The incidence of models with desired dynamic properties increases with the number of generations, as indicated by the mean incidence (black line) and the maximum and minimum incidence (green-shaded region) observed over 10 statistical repeats for every generation. The dashed line indicates the incidence for a repeat with a fast convergence. The black diamonds indicate the generators selected for subsequent analysis from that repeat. **b)** The distribution of the maximum eigenvalues (λ_{max}) for the generated models over generations. The vertical dashed lines indicate $\lambda_{max} = -2.5$ (left) and $\lambda_{max} = 0$ (right). **Robustness analysis: c)** The time evolution of the normalized perturbed biomass, $v(t)/v_{ref}$, (upper left) and concentrations, $(X(t) - X_{ref})/X_{ref}$, Nicotinamide adenine dinucleotide reduced, NADH (upper right), Adenosine triphosphate, ATP (lower left), and Nicotinamide adenine dinucleotide phosphate reduced, NADPH (lower right), respectively: the mean response (dashed black line), the 25th-75th percentile (dark orange region), and the 5th-95th percentile (light orange region) of the ensemble of responses. The vertical dashed line corresponds to $t = 24\ mins$. **Bioreactor simulations: d)** The time evolution of the biomass (left), glucose concentration (middle), and anthranilate concentration (right) in the bioreactor runs: the mean response (dashed black line), the 25th-75th percentile (dark cyan region), and the 5th-95th percentile (light cyan region) of the ensemble of responses.

eigenvalues (Figure 2b). Remarkably, the generated models gradually shifted over the optimization process from having slow dynamics ($\lambda_{max} > -2.5$) to having fast dynamics, with the metabolic processes settling before the subsequent cell division, indicating that RENAISSANCE-generated models could capture the experimentally observed dynamics.

Since cellular organisms maintain phenotypic stability when faced with perturbations[43], the generated models that describe cellular metabolism should possess the same property. To test the robustness of the models, we perturbed the steady-state metabolite concentrations up to ±50% and verified if the perturbed system returned to the steady state. To this end, we generated 1000 relevant kinetic models using the last of 10 selected generators (Figure 2a, generation 45). Inspection of the time evolution of the normalized biomass showed that the biomass came back to the reference steady state ($v(t)/v_{ref} = 1$) within 24 minutes for 100% of the perturbed models (Figure 2c). Similarly, the perturbed time responses of a few critical metabolites, namely, NADH, ATP, and NADPH, returned to their steady-state values within 24 minutes for 99.9%, 99.9%, and 100% of the 1000 generated kinetic models, respectively (Figure 2c). Examining every cytosolic metabolite collectively revealed that 75.4% of the models returned to the steady state within 24 minutes and 93.1% returned within 34 mins, demonstrating that the generated kinetic models are robust and obey imposed context-specific observable biophysical timescale constraints.

Next, we tested the generated models in nonlinear dynamic bioreactor simulations closely mimicking real-world experimental conditions[42,44]. The temporal evolution of biomass production showed similar trends as typical experimental observations with clear exponential and stationary phases of *E. coli* growth (Figure 2d, Supplementary Figure 5). Similarly, glucose uptake and anthranilate production also reproduce trends observed in experiments with glucose completely consumed and anthranilate production saturated around 20 hours[29,30]. This study indicates that the RENAISSANCE models can accurately reproduce the physiologically observable and emergent properties of cellular metabolism, even without implicit training to reproduce fermentation experiments.

Characterizing the intracellular states of *E. coli* metabolism

Accurately determining the intracellular levels of metabolite profiles and metabolic reaction rates is crucial for associating metabolic signatures with phenotype. Yet, our capabilities to establish the intracellular metabolic state are limited. Notwithstanding the ever-increasing availability of physiological and omics data, a significant amount of uncertainty in the intracellular states remains. To reduce this uncertainty, we propose using kinetic models because of their explicit coupling of enzyme levels, metabolite concentrations and metabolic fluxes. Moreover, kinetic models allow us to consider dynamic constraints in addition to steady-state data, thus allowing us further uncertainty reduction.

After integrating available physiology and omics data[42,45–47] using the constraint-based thermodynamics-based flux balance analysis[13], significant uncertainty was present in the intracellular metabolic state as indicated by the wide ranges of metabolite concentrations and metabolic fluxes. We sampled 5000 steady-state profiles of metabolite concentrations and metabolic fluxes from this uncertain space and deployed RENAISSANCE to find the fastest possible dynamics (maximum negative eigenvalues, λ_{max}) for each steady state (Methods, Supplementary figure 6). We visualized the steady-state profiles by performing dimension reduction with Principal Component Analysis (PCA)[48] and t-Distributed Stochastic Neighbor Embedding (t-SNE)[49] (Methods) and colored each steady-state profile according to the obtained λ_{max} (Figure 3a). We observed a high variation in the dynamics (λ_{max}) of the studied steady-state profiles (Figure 3c, blue distribution). Out of 5000 steady-state profiles, 918 (18.4%) had λ_{max} larger than -2.5, meaning that these intracellular metabolic states could not correspond to the experimental observations. Indeed, the dynamic responses corresponding to these states are with a time constant superior to 24 mins, i.e., slower than the experimental observations.

Inspection of the intracellular steady state space suggested that the steady-state profiles corresponding to slow (Figure 3a, yellow dots) and fast (Figure 3a, blue dots) are locally clustered. From this observation, we hypothesized that distinct subregions corresponding to the experimental observations exist and that steady-state profiles sampled in the vicinity of the chosen local cluster would likely satisfy dynamic requirements.

To test this hypothesis, we selected one of these local clusters (Figure 3b), which contained 22 steady states with fast dynamics with $-3.8 \leq \lambda_{max} \leq -8.5$ (Figure 3c, green distribution), and analyzed its neighborhood (Figure 3d, left). We sampled 90 additional steady states within this neighborhood from the Gaussian distribution with a mean and standard deviation estimated on the initial 22 steady states. The sampled steady states allowed us to improve the resolution of the initial dynamic landscape (Figure 3d, right, circles). Crucially, the sampled steady states had linearized dynamics in the same range as the initial 22 states (Figure 3d, e), confirming our hypothesis. Indeed, RENAISSANCE allows us to select subsets of intracellular states consistent with experimentally

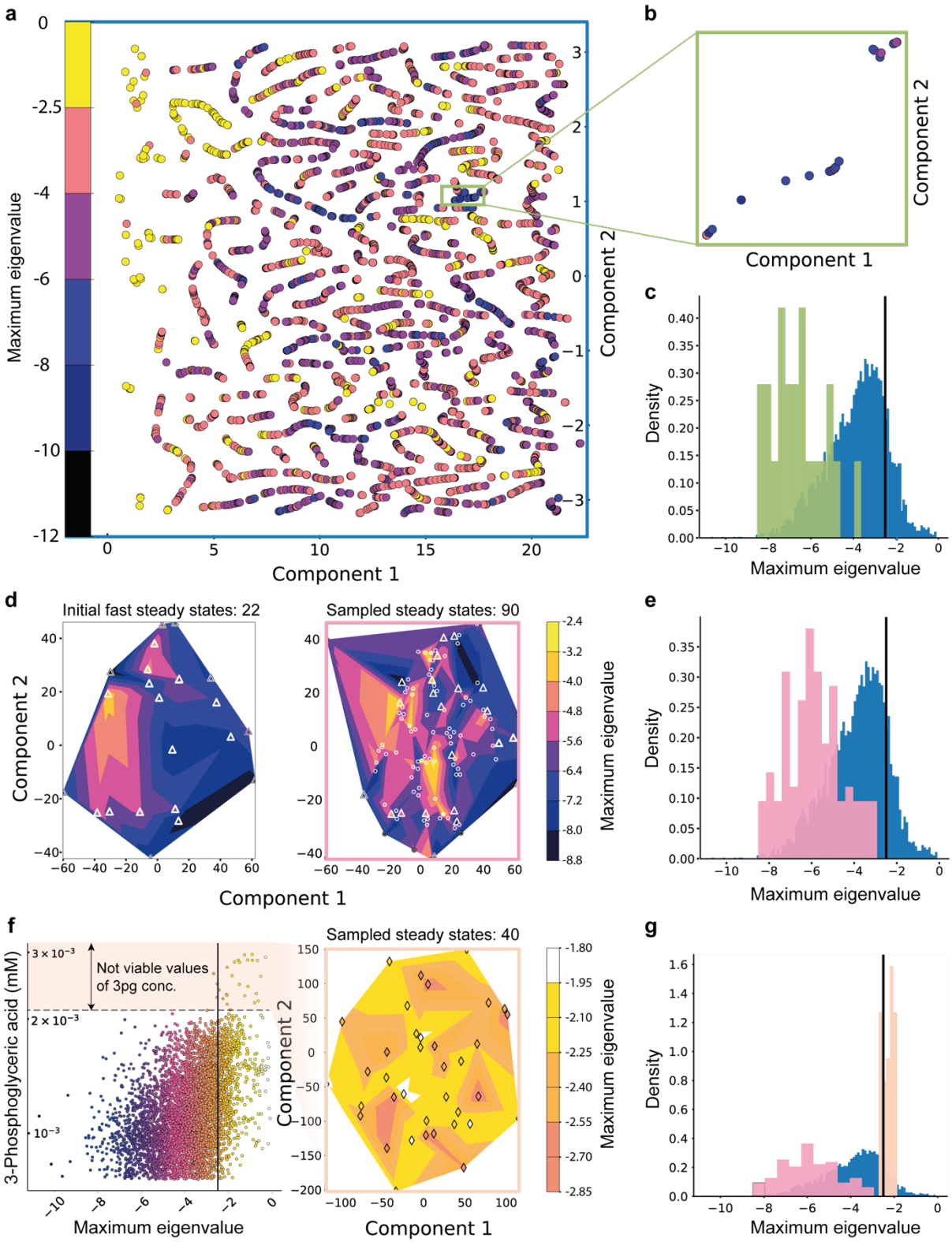

Fig. 3 | Dynamic characterization reduces uncertainty in intracellular metabolic states. a) The two-dimensional representation of the fastest linearized dynamic modes (corresponding to the maximum eigenvalue λ_{max}) of 5000 intracellular steady states (reaction fluxes and metabolite concentrations) obtained with Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE)[49] (Methods). Each point represents a steady state colored according to λ_{max} computed by RENAISSANCE for that steady state. **b)** Magnified view of 22 neighboring steady states with fast dynamics ($-3.8 \leq \lambda_{max} \leq -8.5$). Color scheme is the

same as 3a. **c)** Distributions of the fastest linearized dynamic (λ_{max}) for all 5000 steady states (blue) and for the 22 steady states shown in 3b (green). **d)** Left: the linearized dynamics landscape of the 22 fast steady states in the reduced space (Methods). The triangles represent the location of the steady states in the landscape. Right: The landscape on the left is enhanced by sampling 90 additional steady states in the neighborhood of the initial 22 steady states. The circles represent the location of the newly sampled steady states in the same landscape as on the left. **e)** Distributions of the fastest linearized dynamic (λ_{max}) for all steady states (blue) and for 90 steady states sampled in 3d (pink). **f)** Left: concentration of 3-Phosphoglyceric acid (mM) vs. the fastest linearized dynamic in every steady state. Color scheme is the same as that in 3a. The horizontal black line indicates the cutoff for valid models ($\lambda_{max} = -2.5$). The peach shaded region indicates the range of 3-Phosphoglyceric acid concentration that does not allow fast dynamics. Right: the dynamic landscape of 40 steady states sampled by constraining the metabolite concentrations of 30 metabolites to ranges that do not support fast dynamics. The diamonds represent the location of the steady states. **g)** Distributions of the fastest linearized dynamic (λ_{max}) for the 40 steady states sampled in 3f (peach), compared to all steady states (blue) and those sampled in 3d (pink).

observed dynamics and generate additional ones with the same characteristics. Moreover, it allows us to discard subregions with experimentally inconsistent states, thus reducing uncertainty.

We next examined individual metabolite concentrations of the 5000 steady-state profiles to identify patterns corresponding to the experimentally observed phenotype. We observed a clear bias in the dynamics depending on the concentrations for some of the metabolites (Figure 3f, Supplementary Figure 7). For example, in the case of 3-Phosphoglyceric acid (3pg), we obtain models with relevant dynamics only when the concentration of this metabolite is less than $\sim 0.002\ mM$. In contrast, steady-state profiles with 3pg concentrations between $0.002 - 0.003\ mM$ do not have relevant dynamics (Figure 3f). To investigate this further, we identified 30 cytosolic metabolites that showed such concentration biases by visual inspection (Supplementary Figure 7) and sampled 40 new steady states from the same Gaussian distribution as before (Figure 3d, left) but constrained the selected 30 metabolites to concentration ranges that do not support relevant dynamics (e.g., peach shaded region in Figure 3f). As expected, almost all of these new intracellular states did not yield models with relevant dynamics (Figure 3f, right and 3g). This result demonstrates that information stemming from the dynamic responses can be used to constrain values of intracellular metabolites to specific ranges.

Overall, dynamic characterization of a broad range of intracellular states allows us to reduce uncertainty at the level of steady-state profiles and individual metabolite concentrations and metabolic fluxes.

Integration and reconciliation of experimental information

Experimentally measured Michaelis constants, K_Ms, are curated in comprehensive databases like BRENDA[50]. Integrating experimental results from *in vivo* and *in vitro* studies, despite the disparities in their parameter values, can help further constrain uncertainty and lead to a more accurate description of intracellular metabolic states. To this end, we retrieved from BRENDA experimentally measured values for 108 out of 384 K_Ms in our model (Methods).

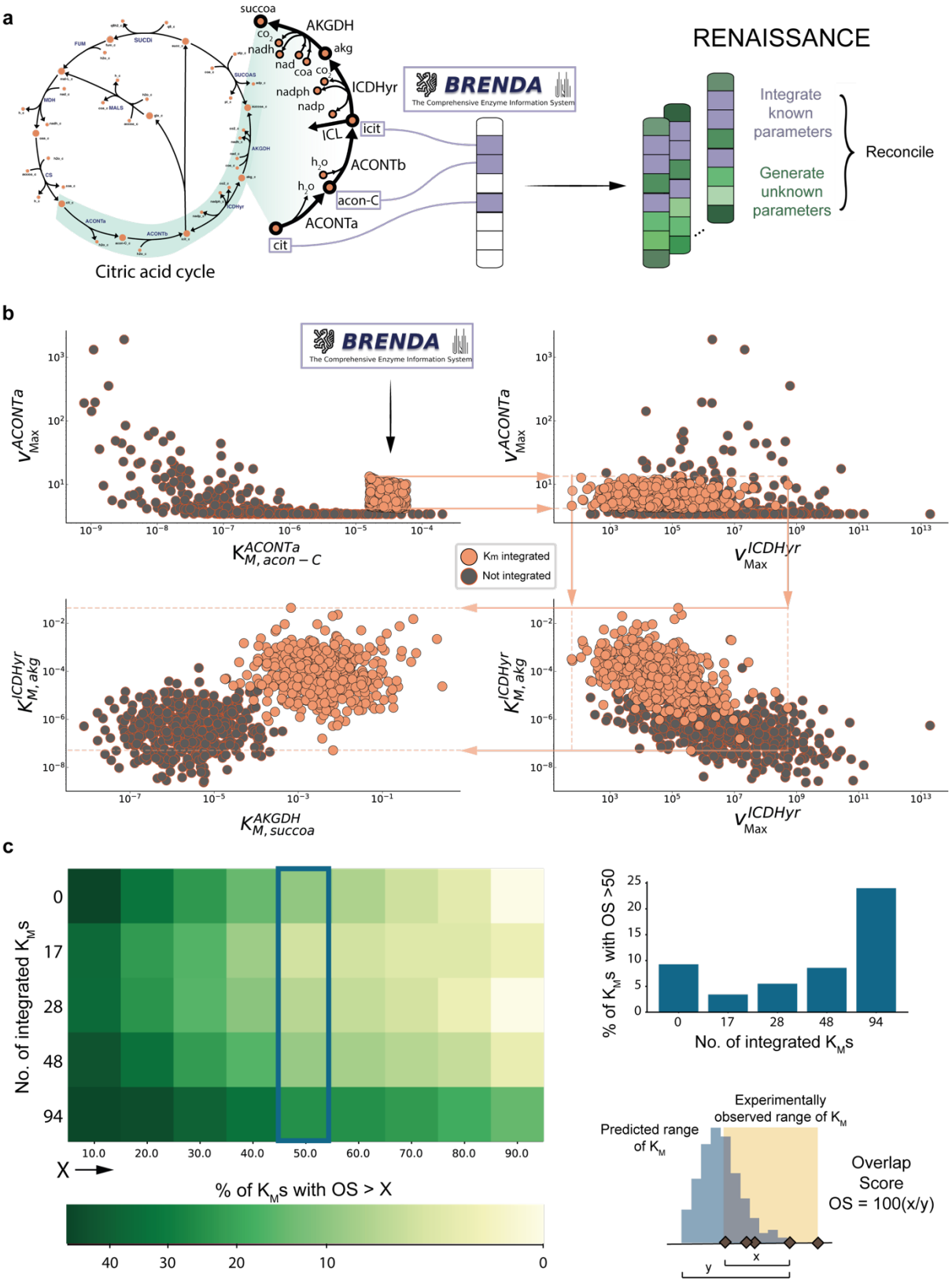

Fig. 4 | Integrated experimentally measured data improve estimates of kinetic parameters. a) RENAISSANCE allows direct integration of K_M values from literature and reconciling them with unknown parameters that collectively lead to valid kinetic models. **b)** Propagation of the integrated K_M experimental data around ACONTa and ACONTb through the metabolic network. Comparison of RENAISSANCE generated values for: $K_{M,acon-C}^{ACONTa}$ vs v_{max}^{ACONTa} (upper left), v_{max}^{ICDHyr} vs v_{max}^{ACONTa} (upper right), v_{max}^{ICDHyr} vs $K_{M,akg}^{ICDHyr}$ (lower right), and $K_{M,succoa}^{AKGDH}$ vs $K_{M,akg}^{ICDHyr}$ (lower lef) when (i) no kinetic parameters (grey circles) (ii) 4 parameters are integrated

(orange circles, Methods). Abbreviations: *ACONTa*: Aconitate A, *ICDHyr*: Isocitrate dehydrogenase, *AKGDH*: 2-Oxoglutarate dehydrogenase, *acon-C*: Cis-Aconitate, *akg*: 2-Oxoglutarate, *succoa*: Succinyl Coenzyme-A. **c)** The integrated K_M experimental data from the BRENDA database improves the RENAISSANCE predictions of not integrated parameters. Overlap Score (OS) for a kinetic parameter is calculated as the percentage overlap of the RENAISSANCE predicted range (x) with the experimentally observed range (yellow) by the total predicted range (blue, y) (right bottom). Percentage of known K_Ms with an overlap score of >50% for different amounts of data (0, 17, 28, 48 and 94 K_ms) integrated (right top). Percentage of known K_Ms with an overlap score of >10% to >90% for different amounts of data integrated (left). Dark blue box represents the exhibited right top case.

To investigate how the integrated kinetic data constrain unknown kinetic parameters, we started by integrating 4 K_Ms of Aconitate hydratase from the citric acid cycle (Figure 4a, ACONTa, b, Methods), trained a generator with a high incidence of valid models (>99%), and generated 500 valid kinetic models (Supplementary Figure 8). To quantify the effect of integrating one experimental K_M value on the generated values of the other kinetic parameters, we compared the estimates of the other K_Ms and maximum velocities, v_{max}, with ones obtained when no kinetic parameters were integrated.

Integration of K_M values of ACONTa at a reaction level restricted the maximum velocity estimates of that reaction, v_{max}^{ACONTa} (Figure 4b, upper left). Due to the correlation in the v_{max} values throughout the network, restricting v_{max}^{ACONTa} estimates through K_M integration constrained the estimated ranges of other maximal velocities, such as v_{max}^{ICDHyr} (Figure 4b, upper right). This restriction further affected downstream K_M values in the network, such as $K_{m,akg}^{ICDHyr}$ and $K_{M,succoa}^{AKGDH}$ (Figure 4b, bottom left and right). These results suggest that integrating only a small amount of experimental data, localized to one enzyme (ACONTa, b), propagates throughout the whole metabolic network and significantly alters the rest of the kinetic parameters.

We next enquired if RENAISSANCE improves its K_M estimates as the number of integrated experimental K_M values increases. To this end, we integrated 17 out of 108 experimentally measured K_M values and generated 500 valid kinetic models (Methods). Then, for the remaining 91 experimental K_M values, we calculated the overlap of the distributions of the generated kinetic parameters with the experimentally known range for each parameter (Fig. 4c, bottom right). We repeated this procedure by integrating 28, 48, and 94 experimentally measured K_M values and calculated the overlaps for the remaining 80, 60, and 14 K_M values, respectively. The overlap of the estimated distributions and experimentally measured ranges increased as we integrated more experimental information. Indeed, the number of parameters with an overlap score (OS) of more than 50% increased with the number of integrated K_M values (Fig. 3c, top right). The statistics for a range of overlap scores, OS > 10% to OS > 90%, showed the same trend (Figure 4c, left). Overlap scores were low when only a few parameters were integrated (e.g., light green, 0%-2% of parameters with OS>90%) but increased with more integrated parameters (e.g., dark green, 15% of parameters with OS>90%). These results suggest that RENAISSANCE estimates improve by incorporating experimental kinetic information from literature and that this framework can make informed predictions about the unknown values of parameters. As more experimental measurements become available, we foresee that it will further reduce uncertainties in not yet measured parameter values.

Discussion

Metabolism plays a defining role in shaping the overall health of living organisms. A reprogrammed or altered metabolism is not only associated with the most common causes of death in humans – cancer, stroke, diabetes, heart disease, and others – but is also related to many congenital diseases[51]. Thus, a better understanding of metabolic processes is crucial to accelerate the development of new drugs, personalized therapies, and nutrition. Biotechnological advances like the bioproduction of industrially important compounds and environmental bioremediation also hinge on our ability to describe cellular metabolism accurately.

Kinetic models provide the most thorough mathematical representation of metabolism. The efficient construction of these models will open new possibilities for various biomedical and biotechnological applications. However, acquiring the parameters of these models with traditional kinetic modeling approaches is computationally expensive and arduous[15,32]. To improve the efficiency of generating kinetic models, we have recently proposed REKINDLE[32]. This unsupervised deep-learning method uses generative adversarial networks (GANs)[52] for this task. While REKINDLE offered several orders of magnitude improvement in model generation efficiency, it required existing kinetic modelling approaches to create the data required for the GAN training. Herein proposed RENAISSANCE retains the model generation efficiency of REKINDLE without requiring training data.

RENAISSANCE can achieve more than 90% incidence of valid models within 10 to 20 minutes of computational time on a standard workstation. Once trained, the generators generate valid models with a rate of ~1 million valid models in 18 seconds, making it 150-200 times faster than traditional sampling based kinetic frameworks. RENAISSANCE also does not require specialized hardware to execute. The proof-of-concept applications shown here demonstrate RENAISSANCE's applicability to a broad range of studies. In this work, we deployed RENAISSANCE to parameterize valid models of metabolism consistent with an experimentally observed steady-state, with validity being characterized by the biological relevance of their timescales. However, conceptually, any other requirement can be imposed or data used, such as consistency with knockout studies or time series from drug absorption trials.

As RENAISSANCE is agnostic to the nature, range, and number of the parameters it needs to generate, it is straightforward to adjust them to the requirements the models need to satisfy. These parameters are not restricted to kinetic parameters only and can include other unknown quantities in the studied system such as metabolite concentrations.

Crucially, given proteomic data, RENAISSANCE can predict unknown enzyme turnover number, k_{cat}, values and consolidate them with the experimentally measured k_{cat} values from databases such as BRENDA and SABIO-RK[53]. As such, it represents a valuable complement to current machine learning methods that estimate k_{cat} values directly[54–56].

In summary, we provide a fast and efficient framework that leverages machine learning to generate biologically relevant kinetic models. The open-access code of RENAISSANCE will facilitate experimentalists and modelers to apply this framework to their metabolic system of choice and integrate a broad range of available data.

Methods

E. coli model structure and data integration

The studied metabolic network included central carbon pathways of *E. coli* such as glycolysis, pentose phosphate pathway (PPP), tricarboxylic cycle (TCA), anaplerotic reactions, the shikimate pathway, glutamine synthesis, and had a lumped reaction for growth generated using lumpGEM[57]. The resulting model structure had 113 mass balances, including one for biomass accumulation, involving 123 reactions. The kinetic mechanism for each reaction was assigned based on the reaction stoichiometry. The overall model was characterized by 507 kinetic parameters consisting of 384 $K_M s$ and 123 $v_{max} s$ (Supplementary Figure 4).

We integrated different data types into the model to create a context-specific model. We integrated exo-fluxomics and exometabolomics data, such as the growth rate, uptake rates, and extracellular concentrations of different medium components from an earlier experimental study[42]. We used data from other experimental works to impose constraints on the ranges of intracellular concentrations for different metabolites in *E. coli*[47]. Additionally, we imposed constraints on thermodynamic variables calculated using the Group Contribution Method[45,46] to ensure that any sampled flux directionalities and metabolite concentrations were consistent with the second law of thermodynamics.

After defining the model structure, and integrating the available data, we then sampled 5000 sets of steady-state profiles consistent with the integrated data using thermodynamics-based flux balance analysis implemented in the pyTFA tool[58]. Each steady-state profile consists of metabolite concentrations, metabolic fluxes, and thermodynamic variables. Once these profiles are available, we can generate kinetic models around these steady-states[15,28,34,37–40] using the RENAISSANCE framework.

Determining validity of kinetic models

Herein, we consider a kinetic model valid (biologically relevant) if all time constants of the aperiodic model response are consistent with the experimental observations of the studied organism. The time constant defines the time required for the system response to decay to $\frac{1}{e} \approx 36.8\%$ of its initial value. To test the model's time constants, we compute the Jacobian of the dynamic system formed by the model[37]. The dominant time constant of the linearized system is defined as the inverse of the real part of the largest eigenvalue of the Jacobian. The dominant time constants allow us to characterize the model dynamics - fast metabolic processes such as electron transport chain and glycolysis are characterized by small time constants. In contrast, the slower timescale emerges from biosynthetic processes. Additionally, the sign of the Jacobian eigenvalues provides us information on the local stability of the generated models, where a model is locally stable if the real parts of all eigenvalues are negative.
We consider that the dominant time constants of aperiodic model response should be five times faster than the cell's doubling time. This way, a perturbation of the metabolic processes settles within 1% of the steady state before the cell division. The biochemical response should also have a characteristic time slower than the timescale of proton diffusion within the cell[32]. With these properties, models can reliably describe the experimentally measured metabolic responses.
The doubling time of the *E. coli* strain used in this study is $t_{doubling} = 134\ mins$, which corresponds to a growth rate of $ln2 \cdot \frac{60}{t_{doubling}} = 0.31\frac{1}{h}$. Therefore, the dominant time constant of the model's responses should be smaller than one-fifth of the doubling time (26.8 $mins$). Here, we imposed a stricter dominant time constant of 24 minutes, corresponding to an upper limit of $Re(\lambda_i) < -2.5$ (or $-60/24$), on the real parts of the eigenvalues, λ_i, of the Jacobian. All kinetic parameter sets that result in the model obeying this constraint are labelled valid and the rest are labelled invalid.

Assigning rewards to determine fitness in RENAISSANCE.

RENAISSANCE uses Natural Evolution Strategy, NES, (Supplementary Note 1) to optimize the weights of the generator network. However, in order to calculate the local gradient estimate NES also requires an objective

function, F, to evaluate the fitness of each generator network, G. In our study, we use the incidence of the generator, $I(G)$, as the objective function, which is defined as the fraction of the generated models that are relevant $(0 \leq I(G) \leq 1)$. Thus, generator networks that have a higher incidence of relevant models are 'fitter' than those with low incidence and have higher weight in determining the parameters of the seed generator network for the next generation. In many cases, we observed that initially the generator neural networks do not generate any relevant models $(I(G) = 0)$ and thus the optimisation does not proceed as the fitness is always 0. To mitigate this, we added an additional sigmoidal term defined as follows,

$$r = \frac{0.01}{1 + e^{(\lambda_{fastest} - \lambda_{partition})}}$$

where, $\lambda_{fastest}$ corresponds to smallest maximal eigenvalue of the generated models and $\lambda_{partition}$ is the maximal eigenvalue partition that determines relevancy of the kinetic model. In this study, $\lambda_{partition} = -2.5$ (see previous section). This term rewards generators that generate models with dynamics closer to the relevant range more than those which generate models with slower, irrelevant or unstable dynamics. This effectively pushes the optimisation process towards finding generators that generate relevant models. So, the overall reward, R, for a generator, G, can be summarized as,

$$R(G) = \begin{cases} r, & I(G) = 0 \\ I(G), & I(G) > 0 \end{cases}$$

For the large-scale analysis of intracellular states (Fig. 3), the fitness for NES was no longer the incidence of the generators but the fastest dynamic possible of the models generated by a given generator. Thus, the reward was changed suitably as follows,

$$r = 0.5e^{-0.1\frac{\lambda_{mean}}{2}}$$

where λ_{mean} is the mean of the 10 fastest maximum eigenvalues (Supplementary figure 6) generated by a generator (out of 100 for this case study). This reward function ensured that the generators which generated models with more negative maximum eigenvalues (faster linearized dynamics, $\lambda_{\max}$) are rewarded more than the others.

Hyperparameter tuning of RENAISSANCE.

RENAISSANCE has several hyperparameters that can be tuned to achieve the desired objective (Supplementary Notes 3 and 4). In this study the hyperparameters used are as follows: the population size of the generator networks, $n = 20$, noise level in generating the agent population from the mean optimal weights in each generation, $\sigma = 10^{-2}$, learning rate of the gradient step, $\alpha = 10^{-3}$, and the decay rate of learning, $d = 5\%$. In addition, the generated K_Ms were constrained strictly between $\{1.3 \times 10^{-11}, 20\}$ to accurately represent experimentally measured K_M values as curated in the BRENDA database[50]. The hyperparameters of the neural networks are listed below.

Neural network implementation.

All software programs were implemented in Python (v3.8.3). Neural networks were implemented using TensorFlow library[59] (v2.3.0). The generator neural networks were composed of three layers that have a total of 1,076,352 parameters: layer 1, Dense with 256 units, BatchNormalization, Dropout (0.5); layer 2, Dense with 512 units, BatchNormalization, Dropout (0.5); layer 3, Dense with 1024 units, BatchNormalization, Dropout (0.5).

Dimension reduction and visualization of steady states

For generating Fig. 3 a, d, f (left) the following steps were followed: I) the steady state matrix (consisting of 1127 features) was subjected to principal component analysis (PCA)[48]. II) The components of PCA which contributed to over 99% of the total expected variance were reduced to 2 dimensions using t-SNE[49]. III) The t-SNE components $\{x_q, x_p\}$ were then subjected to polar coordinate transformation as follows,

$$x_1 = \sqrt{x_p^2 + x_q^2}$$

$$x_2 = arctan2(x_q, x_p).$$

$\{x_1, x_2\}$ were then plotted to generate the figures.

Curation of experimentally kinetic parameters of *E. coli* from BRENDA

Out of the 384 K_Ms in the metabolic model used in this study, 108 had associated experimentally measured values for *E. coli* in BRENDA database. They belong to the following metabolic subsystems: **i)** Pyruvate metabolism (9), **ii)** Citric acid cycle (17), **iii)** Nucleotide salvage pathway (3), **iv)** Tyrosine, Tryptophan and Phenylalanine metabolism (22), **v)** Glycolsis (14), **vi)** Glutamate metabolism (6), **vii)** Pentose phosphate pathway (11), **viii)** Anaplerotic reactions (15), **ix)** Glycine & Serine metabolism (3), **x)** Histidine metabolism (3), **xi)** Oxidative phosphorylation (2), **xii)** Not assigned (3). For the studies in Figure 3, they were integrated in RENAISSANCE as follows, **Case 1:** $K_{M,acon-C}^{ACONTa}$, $K_{M,cit}^{ACONTa}$, $K_{M,acon-C}^{ACONTb}$, $K_{M,icit}^{ACONTb}$ (total 4) **Case 2:** subsystem **ii** (total 17)**, Case 3:** subsystem **ii, vii** (total 28), **Case 4:** subsystem **ii, vii, i, viii** (Total 48)**, Case 5:** subsystem **ii, vii, i, viii, v, vi, iv** (Total 94). The integrated data is available in the provided supplementary data.

Integrating known kinetic parameters in BRENDA

If there were multiple experimentally measured values for a single K_M in BRENDA, we took the geometric mean ($K_{M,exp}$) of the different values and added an experimental error rate of ± 20% to $K_{M,exp}$. The same error rate was applied if there was only 1 recorded experimental value, $K_{M,exp}$. Then the value of an integrated K_M was sampled uniformly from the range $K_{M,exp}$ ± 20% when integrated into RENAISSANCE for the training process and for generation.

Data availability:

The data that support the findings of this study are publicly available in the Zenodo repository (https://doi.org/10.5281/zenodo.7628650 and the links therein).

Code availability:

A Python implementation of the RENAISSANCE workflow is publicly available at https://github.com/EPFL-LCSB/renaissance and https://gitlab.com/EPFL-LCSB/renaissance. The ORACLE framework is implemented in the SKimPy (Symbolic Kinetic models in Python)[60] toolbox, available at https://github.com/EPFL-LCSB/skimpy.

Acknowledgements

This work was supported by funding from the Swiss National Science Foundation grant 315230_163423, the European Union's Horizon 2020 research and innovation programme under grant agreement 814408, Swedish Research Council Vetenskapsradet grant 2016-06160, and the Ecole Polytechnique Fédérale de Lausanne (EPFL).

Author contributions

S.C., M.M., and L.M. designed the overall method and approach. V.H. and L.M. supervised the research. S.C. and L.M. developed the RENAISSANCE method. S.C., B.N, and M.M. designed the code. S.C., B.N. and L.M. analyzed the data. S.C. and L.M. wrote the manuscript. All authors read and commented on the manuscript.

Conflict of interest

The authors declare no financial or commercial conflict of interest.

Supplementary Information

Supplementary Note 1: Natural Evolution Strategies for Neural Networks

Supplementary Note 2: Comparison between RL and ES

Supplementary Note 3: RENAISSANCE hyperparameters

Supplementary Note 4: RENAISSANCE hyperparameter tuning

Supplementary Figure. 1: Overview of the hyperparameters of RENAISSANCE

Supplementary Figure. 2,3: Hyperparameter tuning of RENAISSANCE

Supplementary Figure. 4: Network map of the *E. coli* metabolic model

Supplementary Figure. 5: The time evolution of (i) Tryptophan concentration (in g/L) (ii) Phenyalanine concentration (in g/l) (iii) Tyrosine concentration (in g/l) in the bioreactor simulation.

Supplementary Figure. 6: Hyperparameter tuning of RENAISSANCE for finding the fastest dynamic

Supplementary Figure. 7: Concentrations of cytosolic metabolites in the steady state versus the maximum eigenvalue

Supplementary Figure. 8: Mean incidence of relevant models over generations when 0, 4, 17, 28, 48 and 94 kinetic parameters are integrated in RENAISSANCE

References

1. Bui, A. A. T., Horn, J. D. V. & Consortium, the N. B. C. Envisioning the future of 'big data' biomedicine. *J Biomed Inform* 69, 115–117 (2017).

2. Monk, J. M. *et al.* iML1515, a knowledgebase that computes Escherichia coli traits. *Nat Biotechnol* 35, 904–908 (2017).

3. Brunk, E. *et al.* Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol* 36, 272–281 (2018).

4. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* 161, 971–987 (2015).

5. Fang, X., Lloyd, C. J. & Palsson, B. O. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat Rev Microbiol* 18, 731–743 (2020).

6. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10, 291–305 (2012).

7. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 15, 107–120 (2014).

8. Lerman, J. A. *et al.* In silico method for modelling metabolism and gene product expression at genome scale. *Nat Commun* 3, 929 (2012).

9. Salvy, P. & Hatzimanikatis, V. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models. *Nat Commun* 11, 30 (2020).

10. Sánchez, B. J. *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol* 13, 935 (2017).

11. Beard, D. A., Liang, S. & Qian, H. Energy Balance for Analysis of Complex Metabolic Networks. *Biophys J* 83, 79–86 (2002).

12. Kümmel, A., Panke, S. & Heinemann, M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol* 2, 2006.0034-2006.0034 (2006).

13. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-Based Metabolic Flux Analysis. *Biophys J* 92, 1792–1805 (2007).

14. Oftadeh, O. *et al.* A genome-scale metabolic model of Saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics. *Nat Commun* 12, 4790 (2021).

15. Miskovic, L., Tokic, M., Fengos, G. & Hatzimanikatis, V. Rites of passage: requirements and standards for building kinetic models of metabolic phenotypes. *Curr Opin Biotech* 36, 146–153 (2015).

16. Saa, P. A. & Nielsen, L. K. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. *Biotechnol Adv* 35, 981–1003 (2017).

17. DeBerardinis, R. J. & Chandel, N. S. Fundamentals of cancer metabolism. *Sci Adv* 2, e1600200 (2016).

18. Munger, J. *et al.* Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy. *Nat Biotechnol* 26, 1179–1186 (2008).

19. DeBerardinis, R. J. & Keshari, K. R. Metabolic analysis as a driver for discovery, diagnosis, and therapy. *Cell* 185, 2678–2689 (2022).

20. Masri, S. & Sassone-Corsi, P. The emerging link between cancer, metabolism, and circadian rhythms. *Nat Med* 24, 1795–1803 (2018).

21. Cascante, M. *et al.* Metabolic control analysis in drug discovery and disease. *Nat Biotechnol* 20, 243–249 (2002).

22. Na, D. *et al.* Metabolic engineering of Escherichia coli using synthetic small regulatory RNAs. *Nat Biotechnol* 31, 170–174 (2013).

23. Gupta, A., Reizman, I. M. B., Reisch, C. R. & Prather, K. L. J. Dynamic regulation of metabolic flux in engineered bacteria using a pathway-independent quorum-sensing circuit. *Nat Biotechnol* 35, 273–279 (2017).

24. Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E. & Siuzdak, G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat Biotechnol* 36, 316–320 (2018).

25. Khodayari, A. & Maranas, C. D. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat Commun* 7, 13806 (2016).

26. Foster, C. J., Gopalakrishnan, S., Antoniewicz, M. R. & Maranas, C. D. From Escherichia coli mutant 13C labeling data to a core kinetic model: A kinetic model parameterization pipeline. *Plos Comput Biol* 15, e1007319 (2019).

27. Gopalakrishnan, S., Dash, S. & Maranas, C. K-FIT: An accelerated kinetic parameterization algorithm using steady-state fluxomic data. *Metab Eng* 61, 197–205 (2020).

28. Hameri, T., Fengos, G., Ataman, M., Miskovic, L. & Hatzimanikatis, V. Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations. *Metab Eng* 52, 29–41 (2019).

29. John, P. C. St., Strutz, J., Broadbelt, L. J., Tyo, K. E. J. & Bomble, Y. J. Bayesian inference of metabolic kinetics from genome-scale multiomics data. *Plos Comput Biol* 15, e1007424 (2019).

30. Haiman, Z. B., Zielinski, D. C., Koike, Y., Yurkovich, J. T. & Palsson, B. O. MASSpy: Building, simulating, and visualizing dynamic biological models in Python using mass action kinetics. *Plos Comput Biol* 17, e1008208 (2021).

31. Bordbar, A. *et al.* Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics. *Cell Syst* 1, 283–292 (2015).

32. Choudhury, S. *et al.* Reconstructing Kinetic Models for Dynamical Studies of Metabolism using Generative Adversarial Networks. *Nat Mach Intell* 4, 710–719 (2022).

33. Miskovic, L., Béal, J., Moret, M. & Hatzimanikatis, V. Uncertainty reduction in biochemical kinetic models: Enforcing desired model properties. *Plos Comput Biol* 15, e1007242 (2019).

34. Andreozzi, S. *et al.* Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant E. coli using large-scale kinetic models. *Metab Eng* 35, 148–159 (2016).

35. Vent, W. Rechenberg, Ingo, Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. 170 S. mit 36 Abb. Frommann-Holzboog-Verlag. Stuttgart 1973. Broschiert. *Feddes Repert* 86, 337–337 (1975).

36. Salimans, T., Ho, J., Chen, X., Sidor, S. & Sutskever, I. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *Arxiv* (2017) doi:10.48550/arxiv.1703.03864.

37. Wang, L., Birol, I. & Hatzimanikatis, V. Metabolic Control Analysis under Uncertainty: Framework Development and Case Studies. *Biophys J* 87, 3750–3763 (2004).

38. Miskovic, L. & Hatzimanikatis, V. Production of biofuels and biochemicals: in need of an ORACLE. *Trends Biotechnol* 28, 391–397 (2010).

39. Miskovic, L. *et al.* A design–build–test cycle using modeling and experiments reveals interdependencies between upper glycolysis and xylose uptake in recombinant S. cerevisiae and improves predictive capabilities of large-scale kinetic models. *Biotechnol Biofuels* 10, 166 (2017).

40. Tokic, M., Hatzimanikatis, V. & Miskovic, L. Large-scale kinetic metabolic models of Pseudomonas putida KT2440 for consistent design of metabolic engineering strategies. *Biotechnol Biofuels* 13, 33 (2020).

41. Shamir, M., Bar-On, Y., Phillips, R. & Milo, R. SnapShot: Timescales in Cell Biology. *Cell* 164, 1302-1302.e1 (2016).

42. Balderas-Hernández, V. E. *et al.* Metabolic engineering for improving anthranilate synthesis from glucose in Escherichia coli. *Microb Cell Fact* 8, 19 (2009).

43. Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. & Doyle, J. Robustness of Cellular Functions. *Cell* 118, 675–685 (2004).

44. Narayanan, B., Weilandt, D., Masid, M., Miskovic, L. & Hatzimanikatis, V. Rational strain design with minimal phenotype perturbation. *Biorxiv* 2022.11.14.516382 (2022) doi:10.1101/2022.11.14.516382.

45. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys J* 95, 1487–1499 (2008).

46. Mavrovouniotis, M. L. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* 36, 1070–1082 (1990).

47. Park, J. O. *et al.* Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat Chem Biol* 12, 482–489 (2016).

48. Dunteman, G. Principal Components Analysis. (1989) doi:10.4135/9781412985475.

49. Hinton, L. van der M. and G. E. Visualizing Data using t-SNE. *JMLR* 9, 2579–2605 (2008).

50. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* 49, D498–D508 (2020).

51. Ezgu, F. Chapter Seven Inborn Errors of Metabolism. *Adv Clin Chem* 73, 195–250 (2016).

52. Goodfellow, I. *et al.* Generative adversarial networks. *Commun Acm* 63, 139–144 (2020).

53. Wittig, U. *et al.* SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Res* 40, D790–D796 (2012).

54. Li, F. *et al.* Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat Catal* 1–11 (2022) doi:10.1038/s41929-022-00798-z.

55. Boorla, V. S., Upadhyay, V. & Maranas, C. D. ML helps predict enzyme turnover rates. *Nat Catal* 5, 655–657 (2022).

56. Heckmann, D. *et al.* Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat Commun* 9, 5252 (2018).

57. Ataman, M. & Hatzimanikatis, V. lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *Plos Comput Biol* 13, e1005513 (2017).

58. Salvy, P. *et al.* pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. *Bioinformatics* 35, 167–169 (2019).

59. Abadi, M. *et al.* TensorFlow: A system for large-scale machine learning. *Arxiv* (2016).

60. Weilandt, D. R. *et al.* Symbolic kinetic models in python (SKiMpy): intuitive modeling of large-scale biological kinetic models. *Bioinformatics* 39, btac787 (2022).