BRIDGING SEQUENCE AND KINETICS: UTILIZING MULTI-SCALE REPRESENTATIONS FOR GENOME-SCALE METABOLIC MODELS

Rana A. Barghout, Lya Chinas Serano, Zhiqing Xu,
Benjamin Sanchez-Lengeling, **Radhakrishnan Mahadevan**
Department of Chemical Engineering & Applied Chemistry
University of Toronto
Toronto, ON, CA
{rana.barghout}@mail.utoronto.ca

ABSTRACT

The construction of accurate enzyme-constrained genome-scale models (ecGEMs) remains a critical challenge in systems biology, limited by sparse kinetic data and the need for biologically meaningful representations. This work presents an integrated framework combining CPI-Pred, a deep learning model to predict kinetic parameters (k_{cat}, K_M, K_I, and k_{cat}/K_M) from sequence and compound embeddings, with kinGEMs, a pipeline to incorporate these parameters into ecGEMs for metabolic optimization. By leveraging representations at multiple scales, the approach captures sequence, structure, and kinetic data to enhance model generalizability and accuracy. Rigorous benchmarking demonstrates the framework's capability to predict growth rates and fluxes that are consistent with experimental observations, reduce median flux variability by 3 fold, and enable better-defined predictive and interpretable metabolic models. These innovations open new avenues for metabolic engineering and synthetic biology, offering robust tools to explore biological perturbations and guiding experimental designs.

1 INTRODUCTION

The rapid expansion of biological datasets, ranging from genomic sequences to metabolic profiles, has highlighted the need for computational frameworks capable of extracting meaningful insights from these complex systems. Among these challenges, the construction of enzyme-constrained genome-scale models (ecGEMs) has emerged as a crucial task for understanding cellular metabolism (Figure 1). These models enable the simulation of metabolic fluxes and growth rates under varying environmental and genetic conditions, offering valuable findings for metabolic engineering and synthetic biology through an interpretable lens, which is often not possible with deep learning-based methods of sequence-to-function prediction. However, the utility of ecGEMs is often hindered by the limited availability of accurate kinetic parameters, such as k_{cat} values, which are essential to constrain reaction rates.

Recent advances in machine learning have provided new avenues for addressing this bottleneck. Deep learning models, particularly those that leverage biological sequence embeddings, have shown promise in predicting enzyme kinetics from sequence and structural data. For example, models such as ESM (Rives et al., 2021) and ProtTrans (Elnaggar et al., 2021) have successfully demonstrated the ability of transformer-based architectures to extract meaningful sequence representations for downstream tasks, including the prediction of enzyme functions. Similarly, compound representations such as d-MPNN outputs (Yang et al., 2019) and ECFP (Rogers & Hahn, 2010) have been instrumental in predicting molecular interactions. Despite these advances, there remain significant gaps in translating these predictions into actionable insights within genome-scale models (GEMs). Current approaches, such as sMOMENT (Bekiaris & Klamt, 2020) and GECKO (Chen et al., 2024), integrate enzyme kinetics but often rely on sparse or manually curated datasets. Even when deep learning models are employed to predict missing parameters, these methods often face issues such

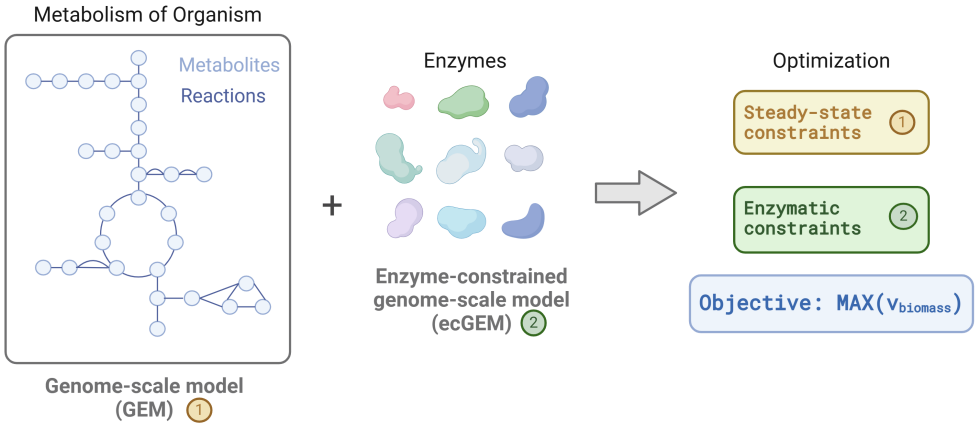

Figure 1: **Schematic representation of genome-scale metabolic modeling with enzyme constraints (ecGEMs)**. Integrating enzyme turnover rates (k_{cat}) into reaction flux constraints generates an ecGEM from a baseline GEM.

as unrealistic enzyme constraints, limiting their scalability and generalizability (Chen et al., 2024). Another critical limitation lies in the evaluation of these models. While some studies focus on metrics like prediction accuracy for kinetic parameters (Li et al., 2022), there is a lack of standardized benchmarks that link these predictions to downstream tasks (Sánchez et al., 2017; Kroll et al., 2023; Li et al., 2022), such as constraint-based modeling (CBM) for phenotypic prediction stemming from genetic perturbations (Orth et al., 2010). This inconsistency hampers the ability to compare models and assess their utility in biological contexts.

Moreover, the interpretability of learned representations remains an open question, particularly when integrating multi-scale data spanning sequence and metabolic pathway kinetics. In this work, we present an integrated framework that combines CPI-Pred (Xu et al., 2025), a deep learning model for **C**ompound-**P**rotein **I**nteraction **Pred**iction, and more specifically, for kinetic parameter prediction, with kinGEMs (**kin**etically-constrained **GE**nome-scale **M**odel**s**), a computational pipeline for constructing and optimizing ecGEMs. CPI-Pred leverages advanced sequence embeddings and compound encodings to predict k_{cat}, K_M, K_I, and k_{cat}/K_M values (although we will only focus on k_{cat} results in this paper), while kinGEMs integrates these parameters into GEMs to simulate metabolic behaviors. This study makes several key contributions:

1. **Accurate and Generalizable Kinetic Predictions:** Demonstrate the ability of CPI-Pred to generate accurate and generalizable kinetic parameters.

2. **Enhancing GEMs with Predicted Kinetics:** Improve the performance and interpretability of GEMs by integrating CPI-Pred predictions through the kinGEMs pipeline. Develop an optimization framework that tunes kinetic parameters using experimental growth rates, with potential expansion to metabolic fluxes & other omics data.

3. **Benchmarking Framework for Model Evaluation:** Establish a robust benchmarking framework to evaluate model performance against experimental genetic perturbation datasets.

Our findings have significant implications for the broader field of systems biology and representation learning. By combining machine learning with metabolic modeling, this work highlights the potential for developing scalable and interpretable frameworks that bridge molecular-level data, whole cell modeling, and organismal-level predictions.

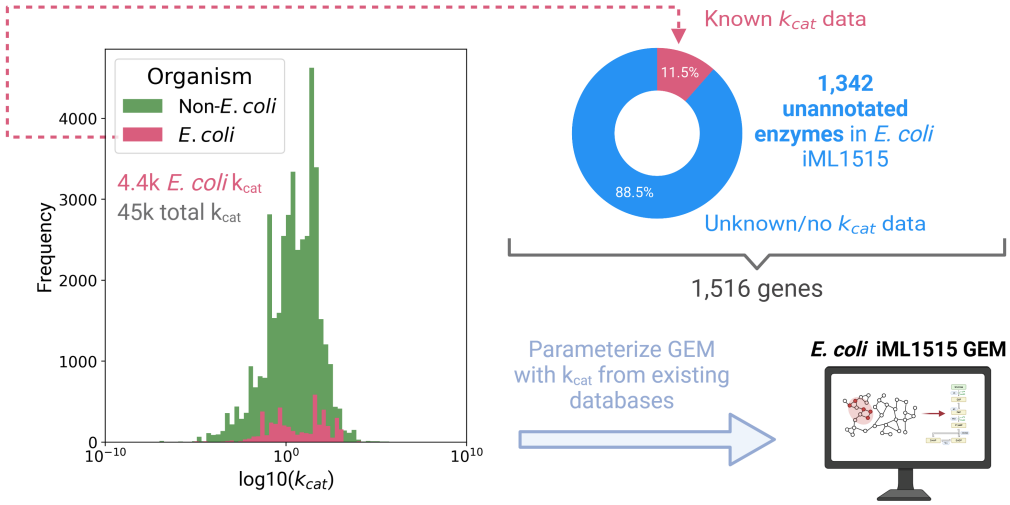

Figure 2: **Distribution of available kinetic parameter (k_{cat}) data from the BRENDA database**. The histogram highlights the abundance of k_{cat} values across diverse organisms (green) compared to those specifically available for *E. coli* (pink). Despite *E. coli* being a well-studied organism, the pie chart reveals that when utilizing k_{cat} values from BRENDA, an overwhelming majority of enzymes in the iML1515 GEM remain unannotated, underscoring a critical gap in kinetic parameter coverage.

2 METHODOLOGY

KINETIC PARAMETER TRAINING DATASET

Accurate kinetic parameter data is essential for constructing ecGEMs, yet a significant portion of kinetic parameter entries in BRENDA (Schomburg et al., 2003)—approximately 60%—lack annotated UniProt IDs, creating a disconnect between enzyme sequence and functional data. To bridge this gap, we utilize the KinMod database (Haddadi et al., 2022), which provides a hierarchical representation of metabolic regulatory networks across 9,814 organisms. This structure captures relationships between proteins and kinetic information derived from from *in vitro* experiments, with a particular focus on the small regulatory network (SMRN). Using KinMod, we assign protein sequences to BRENDA entries with missing UniProt IDs based on EC number and species. This approach enables the creation of "core" datasets, containing entries with complete UniProt annotations, and "pangenomic" datasets, which augment the core data with the KinMod-assigned sequences for entries lacking UniProt IDs. This integration of KinMod and BRENDA enhances dataset completeness and strengthens the foundation for downstream modeling.

CPI-PRED ARCHITECTURE

The CPI-Pred model employs a deep learning architecture to predict compound-protein interactions to gap-fill for sparse k_{cat} data in ecGEMs (Figure 2). It begins by processing protein sequences and compound SMILES as inputs into a multi-expert ensemble model and outputs the predicted kinetic parameter as a functional descriptor. First, the protein sequences are processed by generating embeddings using pre-trained protein language models (pLMs), including ESM-2 (Lin et al., 2022), ProtTrans (Elnaggar et al., 2021), Ankh (Elnaggar et al., 2023), and CARP (Yang et al., 2024). These models extract robust functional and structural representations, with ESM-2 embeddings primarily used for performance evaluation and other pLMs validating robustness across sequence-to-function tasks. To manage the high-dimensionality of sequence embeddings, CPI-Pred integrates multiple independent dimensionality reduction techniques. Self-attention pooling, LSTM variational encoder pooling, and 1D convolutional pooling are all utilized (independently in different models) as methods for reducing the dimension of the pLM embeddings and converting them to size-consistent 1D

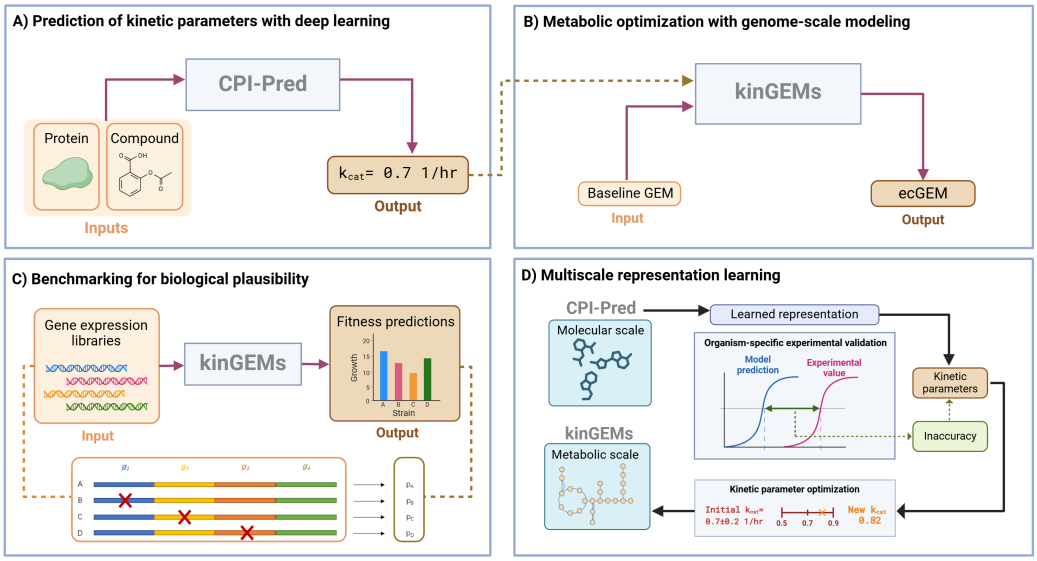

Figure 3: **Overview of the CPI-Pred and kinGEMs framework for multi-scale modeling of enzyme kinetics and genome-scale metabolism**. Panel A illustrates the CPI-Pred workflow, which predicts kinetic parameters (k_{cat}) by combining protein sequence embeddings (e.g., ESM-2) and compound SMILES representations through advanced neural network architectures. These predicted parameters are integrated into GEMs as shown in panel B, where kinGEMs refines the GEM to generate ecGEMs. Panel C highlights the application of kinGEMs in predicting fitness and growth across gene expression libraries, enabling the simulation of metabolic outcomes under diverse conditions. Finally, panel D demonstrates the iterative learning process for refining kinetic parameters, incorporating model predictions, experimental validation, and parameter optimization to enhance the accuracy and utility of ecGEMs. This multi-scale pipeline bridges molecular-level predictions with metabolic-scale applications, providing a robust framework for enzyme and strain design.

vectors. These methods enhance prediction accuracy across diverse datasets. On the compound side, CPI-Pred uses a message passing neural network (MPNN) to encode molecular representations, leveraging the full covalent graph view of a molecule. This integration strengthens the model's ability to capture nuanced compound-protein interactions, even with smaller datasets. CPI-Pred incorporates a cross-attention mechanism to dynamically align and integrate protein and compound representations. Unlike conventional two-stage pipelines (Li et al., 2022; Kroll et al., 2023; Boorla & Maranas, 2024), where protein and compound representations are concatenated and processed through a feed-forward neural network, this approach aims to incorporate a learnable component for multi-modal interaction modeling. Finally, an ensemble strategy aggregates outputs from four models, combining self-attention, LSTM, convolutional pooling, and cross-attention, maximizing robustness and accuracy in predictions.

KINGEMS FRAMEWORK

kinGEMs builds upon established ecGEM methodologies that incorporate quantitative proteomics data (Adadi et al., 2012; Sánchez et al., 2017; Bekiaris & Klamt, 2020) with the constraints presented in Supplementary Figure S1. GEMs can be represented as a system of linear equations that describe the flux of the reactions and metabolites that compromise the metabolic network of an organism. This formulation embedded into kinGEMs accounts for isoenzymes, protein complexes and functional enzyme activities using Boolean rules (further described in Supplementary Information section A.1). The kinGEMs framework begins by identifying the substrates and enzymes in a pre-existing GEM, by looking at the annotated reactions and its gene-protein-reaction (GPR) associations. Subsequently, k_{cat} values for the enzymes are predicted using CPI-Pred to address the lack of data (Figure 2) and integrated back into the ecGEM to fulfill its constraints (panels A and B in Figure 3).

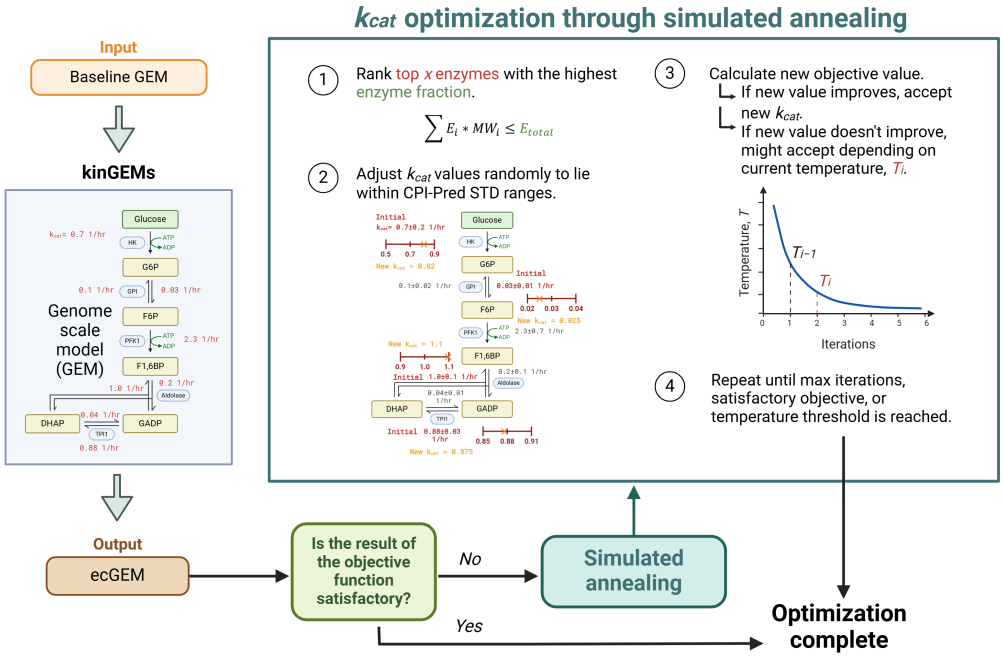

Figure 4: **Overview of the simulated annealing algorithm for optimizing the k_{cat} values that parametrize kinGEMs**. The framework starts by using CPI-Pred k_{cat} values and solving for the objective function. If the objective function is not satisfactory, simulated annealing begins.

Simulating biological perturbations is performed through CBM (Orth et al., 2010; Sánchez et al., 2017; Bekiaris & Klamt, 2020). CBM incorporates available stoichiometric information into a matrix and the metabolic network is assumed to operate under quasi-steady state conditions (Supplementary Figure S1). This assumption allows the formulation of an optimization problem to maximize or minimize a specified biological objective, such as biomass production. To evaluate the predictive accuracy of kinGEMs, we compared the predicted cell growth rates against experimental data. In this case, biomass formation was set as the objective function to maximize for the *E. coli* iML1515 model (Monk et al., 2017). This model was selected for its extensive documentation and the abundance of experimental data available in the literature for evaluation purposes. The optimization problem was implemented using the Pyomo library (Bynum et al., 2021), with iPOPT (Wächter & Biegler, 2006) as the solver of choice. If substantial deviations occur between the simulation outcomes and experimental data, a simulated annealing algorithm is applied to optimize the k_{cat} values of the metabolic enzymes, as described in Figure 4.

To evaluate kinGEMs' ability to predict genetic perturbations, we benchmarked its performance on the *E. coli* iML1515 model against the baseline GEM using the protocol described by Bernstein et al. (2023). The dataset utilized for this analysis, derived from RB-TnSeq experiments, measured fitness values across 25 carbon sources for 3,985 genes in *E. coli* BW25113 (Wetmore et al., 2015; Price et al., 2018). Of these, 1,332 genes were matched to the iML1515 model, resulting in a 1,332$\times$ 25 dataset. In this dataset, fitness values are compared to growth rates, where a fitness score of 0 indicates that mutants grew as well as the wild-type, while scores below -2 signify a strong fitness effect, indicative of no growth. Using this dataset, we simulated genetic perturbations along with the carbon source conditions and predicted whether *E. coli* grew or not (Figure 3D).

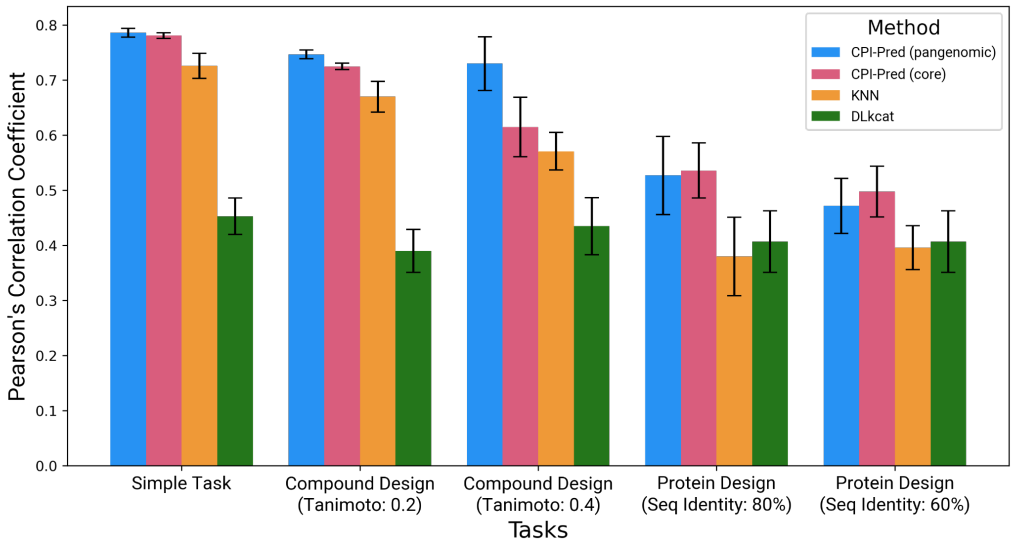

Figure 5: **Prediction performance (Pearson's correlation coefficient) of CPI-Pred on compound and protein design tasks for** k_{cat}. KNN and DLkcat were used for benchmarking against the same (core) dataset splits. The results correspond to averages and standard deviations across models trained using 5-fold cross validation.

BENCHMARKING METRICS

To evaluate prediction accuracy, generalizability, and biological plausibility across our multi-scale pipeline, we employed rigorous benchmarking metrics tailored to each stage. For CPI-Pred, we used Pearson's correlation with error bars derived from 5-fold cross-validation to ensure robust performance evaluation. CPI-Pred's performance was benchmarked against DLkcat (Li et al., 2022) and a k-Nearest Neighbors (KNN) model. To test generalizability, CPI-Pred was further evaluated using two design tasks: the protein design task and the compound design task. Protein sequences were clustered with CD-HIT (Li & Godzik, 2006) at 80% and 60% similarity thresholds to introduce varying levels of difficulty, with 5-fold cross-validation ensuring diverse and non-overlapping clusters across folds. Similarly, compounds were clustered based on Tanimoto similarity (Tanimoto, 1958) at thresholds of 0.2 and 0.4, creating structurally distinct clusters. These tasks rigorously tested the model's ability to predict kinetic parameters for novel proteins and compounds, reinforcing its applicability in enzyme engineering and discovery.

For kinGEMs, we assessed biological plausibility and modeling precision through flux variability analysis (FVA) (Mahadevan & Schilling, 2003) and genetic lethality predictions. FVA was used to evaluate the cumulative distribution of flux variability across all reactions in the *E. coli* iML1515 GEM, with median flux variability serving as an indicator of model constraint precision. Additionally, a kinGEMs-enhanced ecGEM (*E. coli* iML1515) was benchmarked against experimental genetic lethality datasets as described by Bernstein et al. (2023), measuring their ability to predict growth outcomes for gene knockouts under varying carbon sources using metrics like AUC-ROC and accuracy. These evaluations demonstrated the framework's capacity to reduce uncertainty and improve biological relevance in genome-scale models.

3 RESULTS AND DISCUSSION

3.1 CPI-PRED DEMONSTRATES SUPERIOR GENERALIZATION TO NOVEL PROTEINS AND COMPOUNDS COMPARED TO BASELINE MODELS

The performance of CPI-Pred in predicting kinetic parameters highlights its ability to generalize effectively across datasets containing novel proteins and compounds. Figure 5 demonstrates that CPI-Pred, particularly the one trained on the pangenomic variant, consistently outperforms baseline

models such as DLkcat and KNN across key evaluation metrics (Pearson's correlation, with more metrics reported in Supplementary Table S2). The pangenomic variant's incorporation of additional KinMod-augmented sequences provides a broader representation of enzyme diversity, contributing to improved generalization. This superior performance is particularly evident in the context of the protein and compound design tasks. By leveraging clustered protein sequences with CD-HIT at 80% and 60% similarity thresholds and clustering compounds based on Tanimoto similarity at thresholds of 0.2 and 0.4, these tasks introduced varying levels of difficulty. CPI-Pred's ability to maintain high predictive accuracy across these conditions underscores its robustness and versatility in handling structurally diverse proteins and compounds. The results confirm that CPI-Pred excels in capturing the intricate relationships between sequences and molecular interactions even when presented with novel datasets. In comparison, DLkcat and KNN exhibited lower performance, particularly under the more challenging conditions of lower sequence and compound similarity. These results emphasize the limitations of traditional approaches in capturing complex protein-compound interactions and their inability to generalize to novel datasets.

3.2 kinGEMs reduces flux variability and constrains GEM solution space

An FVA was performed to evaluate the degree of flexibility across the model's solution space for a given objective function, with results presented in Figure 6. FVA determines the range of possible flux distributions in both optimal and suboptimal states, identifying redundancies and alternative pathways in metabolic networks (Mahadevan & Schilling, 2003) (further discussed in Supplementary Information section A.3). Our findings reveal that incorporating k_{cat} values, followed by optimization through the simulated annealing algorithm, significantly reduces variability in the solution space. Fewer variable reactions in a metabolic network minimize the uncertainty and enhance model predictability, further validating the biological relevance of CPI-Pred predictions.

3.3 Predicting genetic perturbations for *E. coli* using kinGEMs

Understanding the impact of integrating machine learning-predicted kinetic parameters into genome-scale models is crucial for improving predictions of biological responses to genetic perturbations. In this section, we evaluate the predictive performance of the kinGEMs-enhanced *E. coli* iML1515 model compared to its baseline counterpart in a single-gene mutation prediction problem, where growth serves as the phenotype. Our evaluation focuses on precision-recall and accuracy metrics to assess the model's ability to capture the effects of genetic perturbations.

Incorporating the kinGEMs' produced model improved the AUC-ROC from **0.613 to 0.633** and the accuracy from **0.938 to 0.944**. These results, though modest, demonstrate kinGEMs' capability to enhance predictive performance for genetic perturbation tasks, even for a well-established GEM like *E. coli* iML1515. The slight improvement is likely attributable to iML1515's highly refined and extensively studied nature, leaving limited room for additional enhancement under the constraint-based set-up.

This analysis highlights kinGEMs' potential for modeling biological perturbations and predicting organismal responses. More substantial gains are anticipated for less-defined GEMs of under-studied organisms, where the integration of kinetic parameters through kinGEMs could fill critical knowledge gaps. This application and its implications for expanding kinGEMs to other organisms will be explored further in the next section.

4 Conclusion and future directions

In this work, we present a novel framework that integrates ML-predicted kinetic parameters into ecGEMs, addressing critical gaps in metabolic modeling. Unlike regular GEMs, which rely solely on stoichiometric constraints and often yield a broad solution space, the incorporation of enzyme kinetics improves model specificity. CPI-Pred's use of advanced sequence embeddings, dimensionality reduction techniques, and molecular representation positions makes it a robust tool for addressing the challenges of generalization in kinetic parameter prediction. By combining CPI-Pred's predictive insights with the kinGEMs' mechanistic modeling, we establish a robust framework for assessing prediction accuracy, model generalizability, and biological plausibility. This approach not only enhances the precision and interpretability of ecGEMs but also demonstrates utility in real-

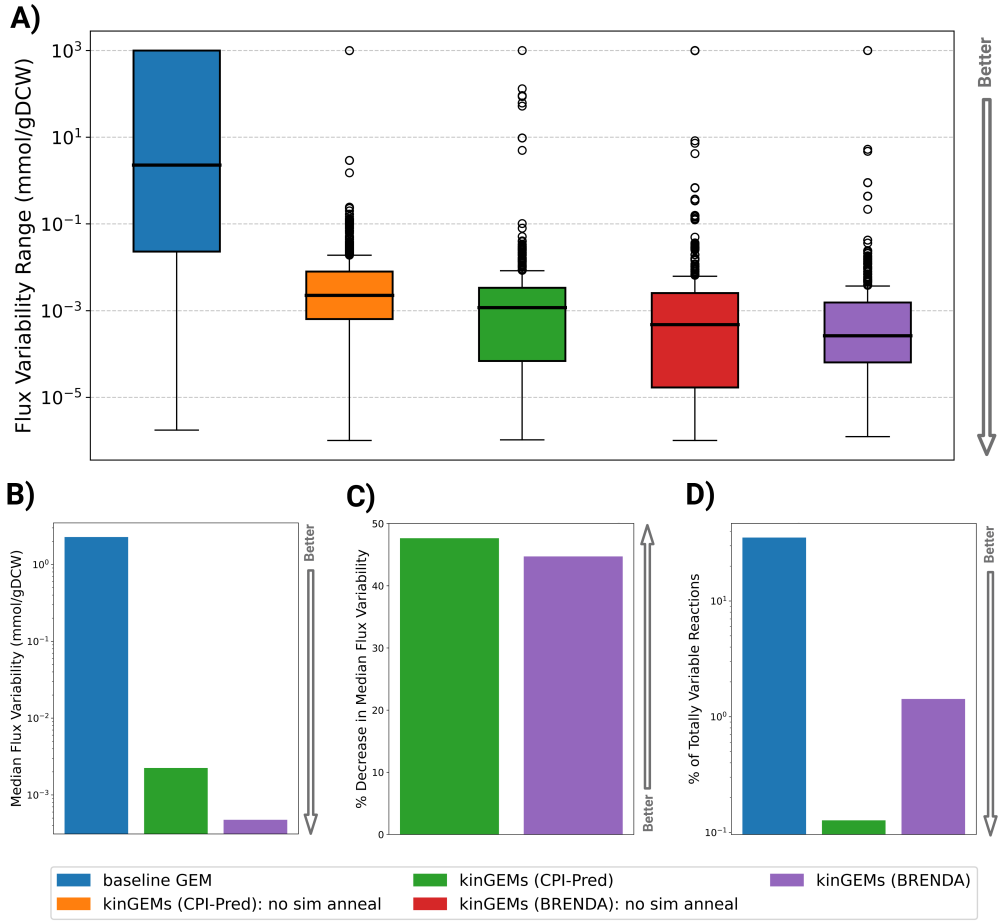

Figure 6: **FVA results for *E. coli* iML1515 ecGEMs constructed with kinGEMs with biomass maximization set as the objective**. Panel A shows a box plot of the flux variability ranges for a baseline GEM as well as kinGEMs-produced GEMs with CPI-Pred and BRENDA sourced k_{cat} values, with and without simulated annealing. Panel B depicts the median flux variability range for the different GEMs, with a lower value signifying less model uncertainty. Panel C depicts the % decrease in median flux variability for the kinGEMs models when applying simulated annealing, where higher values signify better performance. Panel D depicts the % of totally variable reactions—those that operate across the entire range of their predefined lower and upper bounds—with lower percentages signifying better models with more defined operations.

world applications, such as predicting genetic perturbations and improving biological relevance of metabolic networks. Looking ahead, several avenues for future work could further enhance this framework. First, feedback from kinGEMs' perturbation results can be used to refine CPI-Pred's kinetic predictions, creating an iterative loop that improves accuracy over time. Additionally, integrating causal and multi-modal representation learning into kinGEMs could deepen its ability to model complex biological interactions and improve generalizability to under-studied organisms. Furthermore, the practical impact of kinGEMs can be assessed through simulations that explicitly depend on kinetic parameters, such as estimating enzyme costs when screening pathway efficiencies for target molecule production (Noor et al., 2016). Finally, fostering collaborations for dataset sharing and benchmark standardization will enable broader adoption and validation of this approach, facilitating advancements in both computational modeling and experimental biology. Together, these directions hold promise for extending the applicability and impact of kinGEMs in the fields of metabolic engineering and systems biology.

MEANINGFULNESS STATEMENT

This research contributes to shaping a *meaningful representation of life* by leveraging foundation models capable of predicting enzyme kinetics, which are subsequently integrated into genome-scale metabolic models. Our framework establishes a connection between molecular-level data, given by kinetic parameters, and biological-level outcomes, represented by metabolic fluxes and growth rates. This enables multi-scale data utilization by effectively linking distinct levels of biological organization. Finally, the interpretability of these models is enhanced by linking predictions to downstream tasks like phenotypic prediction of genetic perturbations, enabling a more comprehensive and functional understanding of biological systems.

REFERENCES

Roy Adadi, Benjamin Volkmer, Ron Milo, Matthias Heinemann, and Tomer Shlomi. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Computational Biology*, 8(7):e1002575, 2012. doi: 10.1371/journal.pcbi.1002575. URL https://doi.org/10.1371/journal.pcbi.1002575.

Pavlos Stephanos Bekiaris and Steffen Klamt. Automatic construction of metabolic models with enzyme constraints. *BMC Bioinformatics*, 21(1):19, January 2020. ISSN 1471-2105. doi: 10.1186/s12859-019-3329-9. URL https://doi.org/10.1186/s12859-019-3329-9.

David B. Bernstein, Bilge Akkas, Morgan N. Price, and Adam P. Arkin. Evaluating e. coli genome-scale metabolic model accuracy with high-throughput mutant fitness data. *Molecular Systems Biology*, 19(12), 2023. doi: 10.15252/msb.202311566. URL https://doi.org/10.15252/msb.202311566.

Veda Sheersh Boorla and Costas D. Maranas. Catpred: A comprehensive framework for deep learning in vitro enzyme kinetic parameters kcat, km and ki. *bioRxiv*, 2024. doi: 10.1101/2024.03.10.584340. URL https://www.biorxiv.org/content/early/2024/03/26/2024.03.10.584340.

Michael L Bynum, Gabriel A Hackebeil, William E Hart, Carl D Laird, Bethany L Nicholson, John D Siirola, Jean-Paul Watson, and David L Woodruff. *Pyomo - optimization modeling in Python, 3rd Edition*, volume 67. Springer, 2021. doi: 10.1007/978-3-030-68928-5.

Yu Chen, Johan Gustafsson, Albert Tafur Rangel, Mihail Anton, Iván Domenzain, Cheewin Kittikunapong, Feiran Li, Le Yuan, Jens Nielsen, and Eduard J. Kerkhoven. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. *Nature Protocols*, 19(3):629–667, March 2024. ISSN 1750-2799. doi: 10.1038/s41596-023-00931-7. URL https://www.nature.com/articles/s41596-023-00931-7. Publisher: Nature Publishing Group.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, W. Yu, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(08):1–16, 2021. doi: 10.1109/TPAMI.2021.3095381.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. (arXiv:2301.06568), 2023. doi: 10.48550/arXiv.2301.06568. URL http://arxiv.org/abs/2301.06568. Issue: arXiv:2301.06568.

Kiana Haddadi, Rana A. Barghout, and Radhakrishnan Mahadevan. Kinmod database: a tool for investigating metabolic regulation. *Database*, 2022, 2022. doi: 10.1093/database/baac081. URL https://doi.org/10.1093/database/baac081.

Alexander Kroll, Sahasra Ranjan, Martin K. M. Engqvist, and Martin J. Lercher. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications*, 14(1):2787, 05 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38347-2. URL https://doi.org/10.1038/s41467-023-38347-2.

Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin K. M. Engqvist, Eduard J. Kerkhoven, and Jens Nielsen. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8):662–672, 2022. ISSN 2520-1158. doi: 10.1038/s41929-022-00798-z. URL https://doi.org/10.1038/s41929-022-00798-z.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 05 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158. URL https://doi.org/10.1093/bioinformatics/btl158.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.

Radhakrishnan Mahadevan and Christopher Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–276, 2003. doi: 10.1016/j.ymben.2003.09.002. URL https://doi.org/10.1016/j.ymben.2003.09.002.

Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, Adam M Feist, and Bernhard O Palsson. iML1515, a knowledgebase that computes Escherichia coli traits. *Nature biotechnology*, 35(10):904–908, October 2017. ISSN 1087-0156. doi: 10.1038/nbt.3956. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6521705/.

Elad Noor, Avi Flamholz, Arren Bar-Even, Dan Davidi, Ron Milo, and Wolfram Liebermeister. The Protein Cost of Metabolic Fluxes: Prediction from Enzymatic Rate Laws and Cost Minimization. *PLoS computational biology*, 12(11):e1005167, November 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005167.

Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010. doi: 10.1038/nbt.1614.

Morgan N. Price, Kelly M. Wetmore, Rosalind J. Waters, Martha Callaghan, Jonathan Ray, Hualan Liu, Jennifer V. Kuehl, Ryan A. Melnyk, Joseph S. Lamson, Youn Suh, Hans K. Carlson, Zarayda Esquivel, Hariharan Sadeeshkumar, Romy Chakraborty, Grant M. Zane, Benjamin E. Rubin, Judy D. Wall, Axel Visel, James Bristow, and Adam M. Deutschbauer. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706):503–509, 2018. doi: 10.1038/s41586-018-0124-0. URL https://doi.org/10.1038/s41586-018-0124-0.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. PMID: 20426451.

Ida Schomburg, Antje Chang, Christian Ebeling, Marian Gremse, Carsten Heldt, Gabriele Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32(90001):431D–433, 2003. doi: 10.1093/nar/gkh081. URL https://doi.org/10.1093/nar/gkh081.

Benjamín J Sánchez, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J Kerkhoven, and Jens Nielsen. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology*, 13(8):935, August 2017. ISSN 1744-4292. doi: 10.15252/msb.20167411. URL https://www.embopress.org/doi/full/10.15252/msb.20167411. Publisher: John Wiley & Sons, Ltd.

T.T. Tanimoto. *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation, 1958. URL https://books.google.ca/books?id=yp34HAAACAAJ.

Kelly M. Wetmore, Morgan N. Price, Rosalind J. Waters, Joseph S. Lamson, Jian He, Cara A. Hoover, Matthew J. Blow, James Bristow, Gareth Butland, Adam P. Arkin, and Adam Deutschbauer. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *mBio*, 6(3), 2015. doi: 10.1128/mbio.00306-15. URL https://doi.org/10.1128/mbio.00306-15.

Andreas Wächter and Lorenz T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106 (1):25–57, 2006. Preprint.

Zhiqing Xu, Rana Ahmed Barghout, Jinghao Wu, Dhruv Garg, Yun S. Song, and Radhakrishnan Mahadevan. Cpi-pred: A deep learning framework for predicting functional parameters of compound-protein interactions. *bioRxiv*, 2025. doi: 10.1101/2025.01.16.633372. URL https://www.biorxiv.org/content/early/2025/01/21/2025.01.16.633372.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237. URL https://doi.org/10.1021/acs.jcim.9b00237. PMID: 31361484.

Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294.e2, 2024. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2024.01.008. URL https://www.sciencedirect.com/science/article/pii/S2405471224000292.

A SUPPLEMENTARY INFORMATION

A.1 GENOME-SCALE MODEL CONSTRAINTS & CONSTRUCTION

Genome-scale models, or GEMs, are mathematical representations of the metabolic network of an organism. Adding kinetic constraints to the reaction fluxes of the GEM turns it into an enzyme-constrained GEM, or ecGEM (Supplementary Figure S1). We also take into account multiple scenarios of enzymatic reactions, including those in which there is a single enzyme catalyzing a single reaction, ones where there are multiple enzymes catalyzing a single reaction, and those in which there are single enzymes catalyzing multiple reactions (Supplementary Table S1).

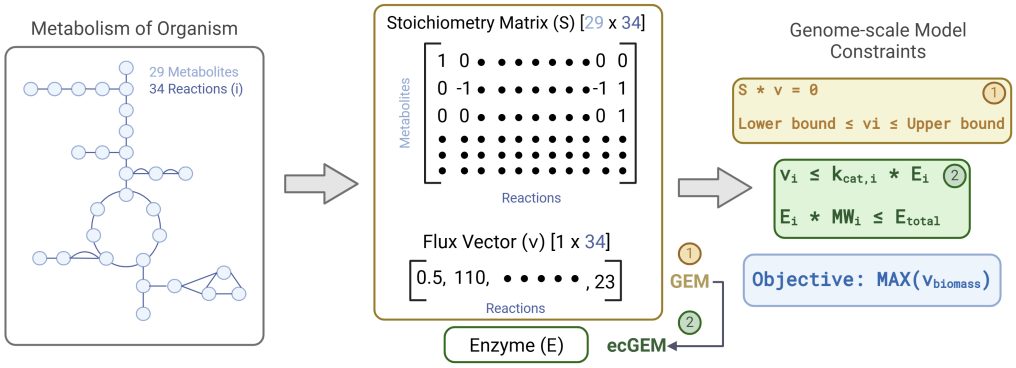

Figure S1: **Schematic representation of genome-scale metabolic modeling with ecGEMs**. Integrating enzyme turnover rates (k_{cat}) into reaction flux (v_j) bounds generates an ecGEM (1 + 2) from a baseline GEM (1).

Scenario	Enzymes	Reaction	Flux Formulation	k_{cat} value if multiple substrates
Baseline	single	single	$v_i \leq k_{cat,i}[E_i]$	$k_{cat,i} = \max(k_{cat,s})$
Isoenzymes (OR)	multiple	single	$v_i \leq \sum_j k_{cat,ij}[E_{ij}]$	$k_{cat,ij} = \max(k_{cat,s})$
Enzyme complex (AND)	multiple	single	$v_i \leq k_{cat,i} \min\left(\frac{[E_{ij}]}{s_j}\right)$	$k_{cat,i} = \max(k_{cat,s})$
Promiscuous Enzymes	single	multiple	$\sum_i \frac{v_i}{k_{cat,ij}} \leq [E_j]$	$k_{cat,ij} = \max(k_{cat,s})$

Table S1: Comparison of different enzyme scenarios and their flux constraints, where i, j, and s annotate reactions, enzymes, and substrates, respectively.

A.2 CPI-PRED

Supplementary Figure S2 shows the general architecture of CPI-Pred, a deep learning framework for compound-protein interaction prediction. Protein sequences are processed using ESM-2 to generate residue-level embeddings. To obtain fixed-length representations, self-attention pooling extracts global contextual dependencies, preserving essential residue-specific features. Conv-1D filtering captures local sequence motifs, refining embeddings by detecting short-range dependencies. LSTM variational encoder pooling compresses variable-length embeddings into consistent latent vectors, mitigating padding effects and improving computational efficiency. Substrate structures are encoded using a message-passing neural network (MPNN) to learn atom-level representations. The protein and substrate embeddings are then concatenated and passed through a feedforward neural network (FFNN) to predict k_{cat} values.

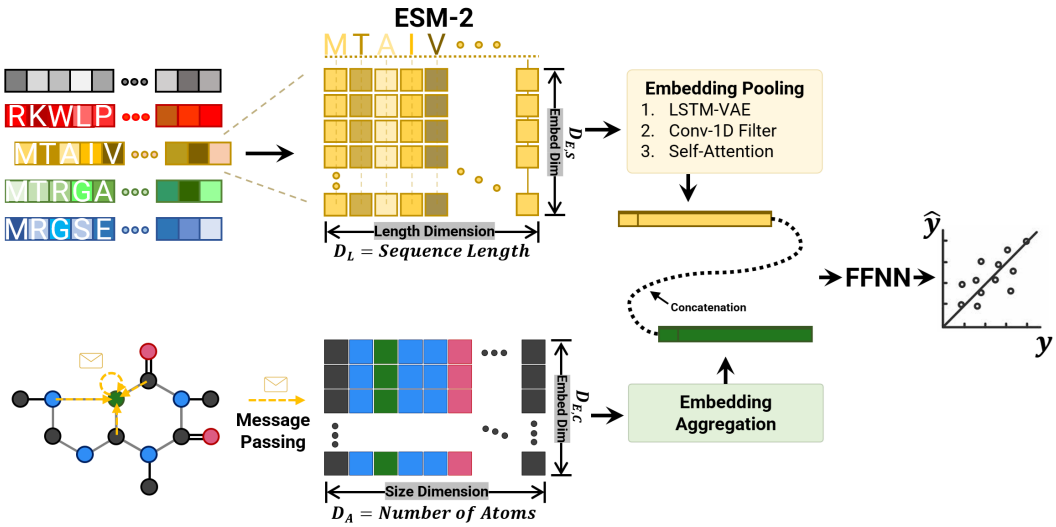

Figure S2: **Schematic of the CPI-Pred model for compound-protein interaction prediction**. Protein sequences are encoded using protein language models (i.e., ESM-2), followed by an embedding pooling block. Compound structures are processed through a message-passing neural network (MPNN) to generate atom-level embeddings. The resulting protein and compound representations are concatenated and passed through a feedforward neural network (FFNN) to predict interaction properties.

Supplementary Table S2 summarizes the prediction performance of CPI-Pred compared to baseline models (KNN and DLkcat) on various compound and protein design tasks involving k_{cat} prediction. Results show that CPI-Pred consistently outperforms baselines across all metrics (Pearson's R, Spearman's ρ, and R^2), particularly under challenging thresholds and sequence identity cutoffs, demonstrating its robustness and generalizability.

Table S2: Prediction performance of CPI-Pred on compound and protein design tasks of kinetic parameters for k_{cat}, with KNN and DLkcat used for benchmarking. Only core validation datasets were used for testing. The results correspond to averages and standard deviations across models trained using 5-fold cross-validation.

Metric	Model	Simple Task	Compound Design (Threshold: 0.2)	Compound Design (Threshold: 0.4)	Protein Design (Identity: 80%)	Protein Design (Identity: 60%)
Pearson's R	CPI-Pred (pangemonic)	0.786±0.008	0.747±0.008	0.730±0.049	0.527±0.071	0.472±0.050
	CPI-Pred (core)	0.781±0.005	0.725±0.006	0.615±0.054	0.536±0.050	0.498±0.046
	KNN	0.726±0.023	0.670±0.028	0.571±0.034	0.380±0.071	0.396±0.040
	DLkcat	0.453±0.033	0.390±0.039	0.435±0.052	0.407±0.056	0.407±0.056
Spearman's ρ	CPI-Pred (pangenomic)	0.767±0.010	0.733±0.009	0.705±0.042	0.511±0.065	0.468±0.060
	CPI-Pred (core)	0.758±0.008	0.718±0.011	0.592±0.062	0.513±0.061	0.487±0.052
	KNN	0.711±0.021	0.678±0.031	0.569±0.041	0.355±0.079	0.369±0.038
	DLkcat	0.380±0.022	0.360±0.035	0.391±0.045	0.371±0.051	0.381±0.043
R^2	CPI-Pred (pangenomic)	0.619±0.016	0.558±0.022	0.532±0.030	0.277±0.074	0.219±0.070
	CPI-Pred (core)	0.609±0.014	0.561±0.019	0.417±0.075	0.315±0.065	0.302±0.055
	KNN	0.526±0.034	0.448±0.037	0.338±0.043	0.144±0.069	0.152±0.058
	DLkcat	0.206±0.023	0.166±0.029	0.209±0.031	0.151±0.038	0.155±0.035

A.3 FLUX VARIABILITY ANALYSIS

Flux Variability Analysis (FVA) is a mathematical approach used in constraint-based metabolic modeling to determine the possible range of flux values for each reaction in a metabolic network while maintaining a given optimal objective function (e.g., biomass production). It helps identify essential and variable reactions under different conditions.

FVA solves two linear optimization problems for each reaction v_i in the system, subject to the mass balance constraints:

$$\mathbf{S} \cdot \mathbf{v} = 0$$

where $\mathbf{S}$ is the stoichiometric matrix, and $\mathbf{v}$ is the flux vector. The fluxes are constrained by lower and upper bounds:

$$v_{\text{min},i} \leq v_i \leq v_{\text{max},i}$$

For each reaction v_i, FVA computes the **minimum** and **maximum** flux values while keeping the objective function Z (e.g., biomass production) within a certain threshold (often its optimal value Z^*):

$$\min v_i, \quad \text{subject to } Z \geq \alpha Z^*$$

$$\max v_i, \quad \text{subject to } Z \geq \alpha Z^*$$

where α is typically set to 0.9 or 1 to ensure near-optimal or optimal growth. The resulting flux ranges provide insights into essential reactions, alternative pathways, and metabolic flexibility. For our results highlighted in this paper, we kept the the objective function at the maximum (optimum) value ($alpha = 1$).

A cumulative distribution plot of the FVA results is highlighted in Supplementary Figure S3. The point $\star$ (top right) denotes the low percentage of variable reactions in each case study constructed with kinGEMs; ranging from **1.42%** (using experimental k_{cat} values) as low as **0.13%** (using CPI-Pred k_{cat} values). This represents a large reduction from the **35.29%** variable reactions for the unconstrained baseline GEM (blue line). The baseline GEM achieves a high biomass formation (blue horizontal line) due to the minimal constraints applied to the model, in comparison to the kinGEMs models (other horizontal dashed lines). However, the simulated annealing process gives the user the opportunity to tune the k_{cat} values until a satisfactory growth rate is achieved, which was around $0.2/hr$ for the CPI-Pred kinGEMs model (green horizontal line) for the experiments presented in Supplementary Figure S3.

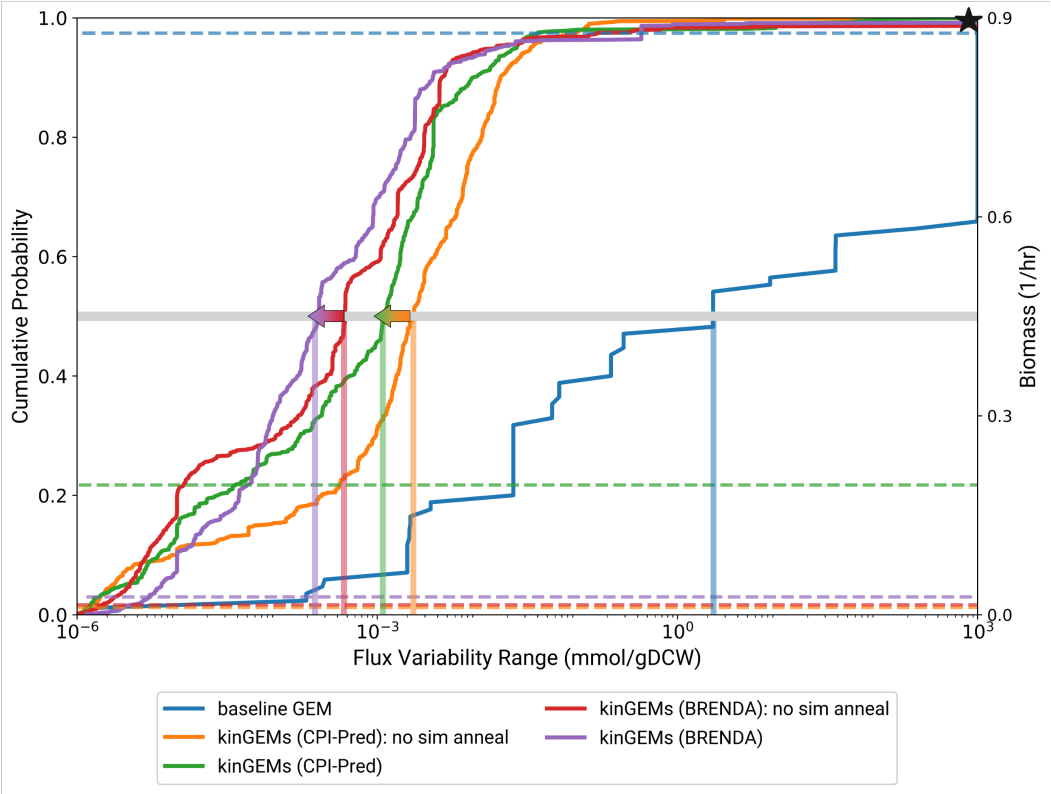

Figure S3: **FVA results for *E. coli* iML1515 ecGEMs constructed with kinGEMs with biomass maximization set as the objective**. The left y-axis shows the cumulative probability, where curves to the right of the plot suggest higher variability. The right y-axis represents the maximum biomass formation rates achieved in each case, marked by the dashed horizontal lines. Case studies with different k_{cat} values were conducted: experimental values from (Schomburg et al., 2003) (red and purple lines), and CPI-Pred values (yellow and green lines). The use of the simulated annealing algorithm showcases a significant reduction in the median variability (gray line) for each case; this reduction is particularly significant when comparing the results of the kinGEMs framework against the baseline GEM without constraints (blue lines).