Yicheng Gao, *Tongji University* | Research Statement

Towards Virtual Cell Construction by AI

My research focuses on the construction of virtual cells[1]: computational models that can represent cellular state, predict responses to perturbations, and enable mechanistic and causal reasoning across molecular and cellular scales. **I approach virtual cell construction as a layered problem, where molecular interactions, multimodal cellular states, perturbation dynamics, and causal dependencies must be modeled in a coherent and interpretable way.** Across my work, I have tried to addressed each of these layers through dedicated methodological contributions, and I am now working toward **unifying them through a natural-language-centered abstraction**, represented by CellHermes [2].

Advances in single-cell sequencing, immune repertoire profiling, and perturbation assays have made it possible to observe cells at unprecedented resolution. However, these technologies have also exposed fundamental computational challenges. Biological data are high-dimensional, noisy, sparsely sampled, and often biased toward limited experimental conditions. Many existing computational approaches rely on task-specific architectures or black-box representation learning, which limits interpretability, generalization, and the ability to perform reliable counterfactual reasoning.

I view virtual cell construction as an alternative paradigm that emphasizes representation, mechanism, and controllability. Rather than optimizing isolated predictive tasks, virtual cell models should encode how cells sense molecular signals, integrate multimodal information, respond to perturbations, and obey causal biological constraints. My research trajectory reflects a deliberate effort to **decompose this problem into its essential components, solve each with principled methods, and then seek a unifying framework that can scale across tasks and modalities**.

Throughout the past five and a half years, I have published **4 first-author papers in Nature Machine Intelligence, Nature Computational Science, Nature Communcations and Cell Genomics [3, 4, 5, 6] with 4 Research Highlight and 1 ESI highly cited mention**. Additionally, I have also published 4 co-first-author papers in Nature Methods, Genome Biology, Cell Systems and Nucleic Acid Research[7, 8, 9, 10].

A. Research Overview

I aim to use AI techniques to construct **virtual cells**. I first decompse this task into different layers, including molecular interactions, multimodal integration of cellular states, perturbation response, and causal disentanglement and counterfacutal generation. Then, I am searching for a unification way that can solve them within a single model, which is "**Natural language**"

Molecular interactions

- PanPep
  Gao et al. Nat. Mach. Intell. 2023
  ESI highly cited, Research Highlight

Multimodal integration

- UniTCR
  Gao et al. Cell. Genom. 2024
  Research Highlight

Perturbation prediction

- STAMP
  Gao et al. Nat. Comp. Sci. 2024
  Research Highlight

Causal disentanglement

- CausCell
  Gao et al. Nat. Commun. 2025
  Research Highlight

Unification

- CellHermes
  Gao et al. BioRxiv. 2025
  Under review

Molecular Interaction Modeling

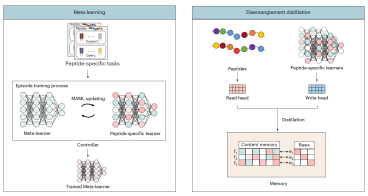
Cellular function ultimately emerges from molecular recognition events that initiate downstream signaling and transcriptional programs. In the adaptive immune system, peptide–T cell receptor (TCR) binding represents one of the most critical such interactions, determining whether a T cell becomes activated[11]. To model this molecular interaction layer of virtual cell construction, I developed PanPep[3], a framework for peptide–TCR binding prediction under few-shot and zero-shot settings.

PanPep departs from conventional supervised interaction prediction models by explicitly framing peptide–TCR recognition as a meta-learning problem. Instead of learning epitope-specific classifiers that require abundant labeled data, PanPep is trained to rapidly adapt to new peptides with only a handful of known binding TCRs, or even none at all. This design reflects the realistic immunological setting, where most epitopes are rare and experimental binding data are extremely limited. By learning transferable interaction patterns across peptides, PanPep captures general principles of molecular recognition rather than memorizing peptide-specific signatures.

Generalization across molecular contexts is more important than maximizing performance on well-sampled epitopes. PanPep explicitly evaluates performance under majority, few-shot, and zero-shot testing regimes, revealing that models optimized only for data-rich epitopes fail to generalize to the long tail of biologically relevant antigens. Through this lens, PanPep provides a more faithful assessment and solution for real-world peptide–TCR interaction prediction.

B. PanPep architecture

Combining meta learning and Neural Turing Machine (NTM), we developed a novel framework called PanPep, which **overcomes the limitation of zero-shotgeneration in meta learning.** We applied this model into peptide-tcr binding prediction task and showed its superior in the zero-shot setting.



In application stage, PanPep can facilitate the T cell sorting, identify the key markers and discover some interaction residues by attention mechanism.

Multimodal Integration of Cellular State

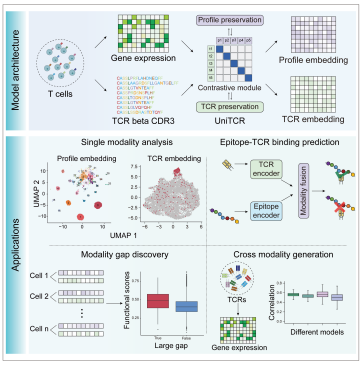
Cells operate through the integration of heterogeneous biological signals that jointly define cellular identity and function. In the adaptive immune system, this integration is exemplified by the coupling between T cell receptor (TCR) sequences and cellular transcriptional states[12]. To model this integration, I developed UniTCR[6], a unified framework for integrating scRNA-seq and TCR-seq data in settings where paired multimodal data are scarce and noisy.

UniTCR departs from standard contrastive learning by explicitly accounting for low-resource characteristics. Through a dual-modality contrastive objective combined with single-modality preservation constraints, the model learns a shared latent space without collapsing modality-specific structure. This enables a wide range of downstream tasks, including refined single-modality analysis, modality-gap-based discovery of functional T cells, epitope–TCR binding prediction, and cross-modality generation of gene expression profiles from TCR sequences.

A key insight from this work is the identification of the modality gap as a biologically meaningful signal rather than a training artifact. We showed that cells with large modality gaps are enriched for functionally important immune states, such as alloreactive or tumor-reactive T cells, highlighting how representation geometry itself can drive biological discovery. UniTCR advances virtual cell modeling by providing a principled representation of integrated cellular state under realistic data constraints.

C. UniTCR architecture

Due to the scarcity of paired TCR-seq and RNA-seq, aligning two modaltity without loss of single-modality information could be challenging. We proposed UniTCR, a model **built upon CLIP framework while preserving the geometry property in each modality** by designing the single-modality preservation module.



After integrating the information of each modality, we can explore the single-modality embedding which has processed the information from the other modality. It also illustrate the efffectiveness of integration on peptide-TCR binding prediction task. **The special modality gap analysis can help discovery potentially functional T cells.** Finally, it can also facilitate the cross modality generation task, which is from TCR sequence to gene expression profile.

Perturbation Response and Cellular Dynamics

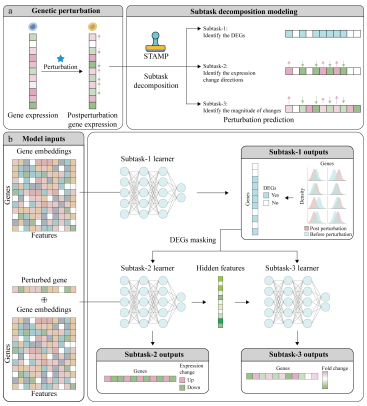
Cells respond to genetic perturbations by propagating gene-level interventions through regulatory programs, leading to systematic changes in transcriptional state[13]. To model this process, I developed STAMP[4], a framework for predicting single-cell transcriptional responses to genetic perturbations, such as gene knockouts and combinatorial gene perturbations. This setting captures a core mechanism by which cellular function is experimentally probed and manipulated, and it is a critical component of virtual cell construction.

STAMP decomposes genetic perturbation modeling into three subtasks through: (1) differentially expressed gene (DEG) prediction, identifying which genes are affected by a perturbation; (2) expression change direction prediction, determining whether affected genes are upregulated or downregulated; and (3) quantitative response modeling, capturing the magnitude of transcriptional changes. This decomposition allows the model to learn perturbation responses at multiple levels of granularity, rather than collapsing all information into a single regression target.

Based on this framework, genetic perturbation responses are more faithfully captured. By aligning modeling objectives with biologically meaningful subtasks, STAMP achieves improved interpretability and robustness under distribution shift.

D. STAMP architecture

The task of perturbation prediciton is a challenging task, due to the noisy and high dimensional data. STAMP decomposed this task into three subtasks for decreasing the difficulty of modeling. The model input is gene embedding which can be obtained from foundation model and output is three-level outcomes.



Causal Disentanglement and Counterfactual Generation

While multimodal integration and perturbation prediction enable rich descriptions of cellular behavior, reliable virtual cells require models that explicitly encode causal structure between biological factors and support counterfactual reasoning. To address this, I introduced CausCell[5], the first framework to explicitly integrate structural causal models with diffusion-based generative modeling for single-cell omics. Unlike prior VAE-based disentanglement approaches that assume independence between latent factors[14], CausCell models biologically meaningful concepts as nodes in a causal graph and enforces their relationships during representation learning and generation.

This design enables three key advances. First, it produces interpretable concept-level embeddings aligned with known biological factors such as cell type, treatment, and temporal state. Second, by conditioning a diffusion model on causally structured latent variables, it achieves strong generalization and high-fidelity reconstruction. Third, it sup-

ports causally consistent counterfactual generation, allowing in silico interventions that respect biological constraints. We further established the first comprehensive benchmarking framework for single-cell disentanglement, demonstrating that causal structure is essential for reliable virtual cell simulation, especially in out-of-distribution and small-sample regimes.

CausCell provides a potential solution to one of the most challenging aspects of virtual cell construction: generating realistic and interpretable cellular states under hypothetical interventions.

Toward Unification via Natural Language: CellHermes

Although PanPep, UniTCR, STAMP, and CausCell address complementary aspects of virtual cell construction, they remain separate systems with task-specific architectures. CellHermes is motivated by the need to unify these components under a common abstraction. Rather than directly incorporating molecular interaction modeling, multimodal integration, perturbation dynamics, and causal graphs into a single model, CellHermes proposes a potential unifying solution based on natural language reformulation[15, 16].

In CellHermes, diverse computational biology problems are expressed as natural-language question–answer pairs that emulate modality-specific self-supervised learning objectives. This approach allows pretrained large language models to serve as a general reasoning backbone, reducing reliance on bespoke architectures and enabling flexible extension to new biological tasks. At its current stage, CellHermes does not yet fully integrate all components of virtual cell construction. Instead, it provides a framework for reformulating these components into a shared linguistic interface, offering a scalable pathway toward unification.

Future Research Agenda

My future research will continue to advance virtual cell construction by moving from specialized, task-specific models toward unified, interpretable, and extensible frameworks for computational biology. Building on my prior work in molecular interaction modeling, multimodal integration, genetic perturbation prediction, and causal disentanglement, I aim to reformulate these complementary components within a shared abstraction that supports flexible reasoning, generalization, and biological interpretation.

Therefore, a central direction of my future work is to further develop CellHermes as a unifying paradigm for virtual cell construction. Rather than directly merging molecular interaction models, perturbation predictors, and causal generators into a single architecture, I aim to reformulate these components through natural language as a common interface. By expressing molecular interactions, perturbation objectives, and causal constraints as language-based instructions and questions, CellHermes provides a scalable pathway to integrate heterogeneous biological tasks without redesigning task-specific architectures. Future efforts will explore how molecular interaction reasoning, genetic perturbation effects, and causal interventions can be progressively incorporated into this language-mediated framework.

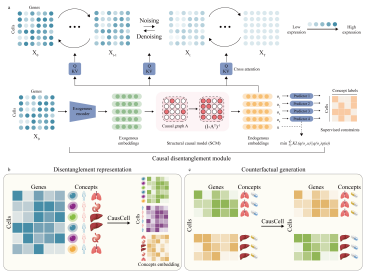
Integrating Causal Reasoning with Perturbation Modeling

One key direction is the deeper integration of causal disentanglement and genetic perturbation modeling. While STAMP models how genetic perturbations transform cellular states and CausCell enables causal counterfactual generation, these components are currently developed separately. I plan to bridge them by embedding causal structure into perturbation response learning, enabling models that not only predict perturbation outcomes but also explain why specific genetic interventions lead to observed transcriptional changes. This integration will support causally grounded virtual perturbation experiments that better reflect biological mechanisms.

Expanding Multimodal and Multi-Scale Virtual Cells

E. CausCell architecture

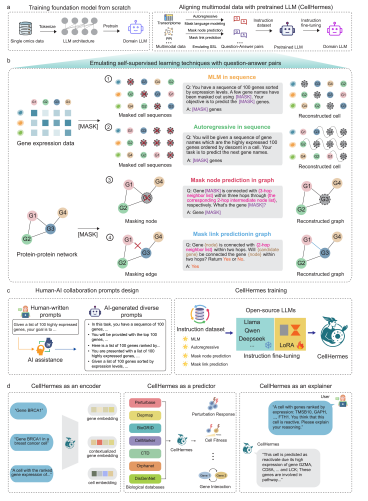
Combing diffusion model and structural causal model (SCM), we designed a novel causal disentanglment framework, which we also derived its optimization ELBO. This can **overcome the trade-off between reconstruction and disentanglement**, to a certain extent.



With this model, we can perform **counterfacutal generation consistent with the causal structure of different concepts.**

F. CellHermes architecture

By virtue of the powerful capability of existing LLMs, such as LLaMA, DeepSeek and Qwen, we proposed using these **open-source LLM to study omics data** and unify all of things we mentioned above. **Natural language is a flexible medium, which can be used to integrate heterogeneous biological information.**



Upon this framework, CellHermes can be used as encoder for gene, cell representation, as predictor for solving different biological tasks within one model, as explainer for interpreting its prediction results.

Future work will extend virtual cell models beyond current modalities by incorporating additional omics layers, such as spatial transcriptomics, proteomics, and protein–protein interaction networks. By grounding these modalities in shared representations and, where possible, natural language abstractions, I aim to build virtual cells that operate across molecular, cellular, and tissue scales. This direction will further strengthen the connection between molecular interaction modeling, multimodal cellular state representation, and system-level behavior.

Toward Interactive and Reasoning-Driven Virtual Cells

Ultimately, I aim to move virtual cell models from passive predictors toward interactive reasoning systems. By combining language-based interfaces with learned biological representations, future virtual cells could support hypothesis-driven queries, iterative intervention planning, and human-in-the-loop exploration. Such systems would enable researchers to ask complex biological questions, simulate interventions, and interpret results within a unified computational framework, narrowing the gap between data-driven modeling and biological insight.

In summary, through these directions, my long-term goal is to establish virtual cell systems that are interpretable, controllable, and adaptable, serving as reliable computational partners for biological discovery.

References

[1] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

[2] Yicheng Gao, Weixu Wang, Yuheng Zhao, Kejing Dong, Caihua Shan, Weizhong Zheng, Till Richter, Zekai Li, Siming Chen, Fabian J Theis, et al. Language may be all omics needs: Harmonizing multimodal data for omics understanding with cellhermes. *bioRxiv*, pages 2025–11, 2025.

[3] Yicheng Gao, Yuli Gao, Yuxiao Fan, Chengyu Zhu, Zhiting Wei, Chi Zhou, Guohui Chuai, Qinchang Chen, He Zhang, and Qi Liu. Pan-peptide meta learning for t-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, 5(3):236–249, 2023.

[4] Yicheng Gao, Zhiting Wei, Kejing Dong, Ke Chen, Jingya Yang, Guohui Chuai, and Qi Liu. Toward subtask-decomposition-based learning and benchmarking for predicting genetic perturbation outcomes and beyond. *Nature Computational Science*, 4(10):773–785, 2024.

[5] Yicheng Gao, Kejing Dong, Caihua Shan, Dongsheng Li, and Qi Liu. Causal disentanglement for single-cell representations and controllable counterfactual generation. *Nature communications*, 16(1):6775, 2025.

[6] Yicheng Gao, Kejing Dong, Yuli Gao, Xuan Jin, Jingya Yang, Gang Yan, and Qi Liu. Unified cross-modality integration and analysis of t cell receptors and t cell transcriptomes by low-resource-aware representation learning. *Cell Genomics*, 4(5), 2024.

[7] Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, pages 1–14, 2025.

[8] Kejing Dong, Yicheng Gao, Qi Zou, Yan Cui, Chuangyi Han, Senlin Lin, Zhikang Wang, Chen Tang, Xiaojie Cheng, Fangliangzi Meng, et al. Benchmarking multi-slice integration and downstream applications in spatial transcriptomics data analysis. *Genome Biology*, 26(1):318, 2025.

[9] Yuli Gao, Yicheng Gao, Siqi Wu, Danlu Li, Chi Zhou, Fangliangzi Meng, Kejing Dong, Xueying Zhao, Ping Li, Aibin Liang, et al. Weakly supervised peptide-tcr binding prediction facilitates neoantigen identification. *Cell Systems*, 16(10), 2025.

[10] Zhiting Wei, Duanmiao Si, Bin Duan, Yicheng Gao, Qian Yu, Zhenbo Zhang, Ling Guo, and Qi Liu. Perturbase: a comprehensive database for single-cell perturbation data analysis and visualization. *Nucleic Acids Research*, 53(D1):D1099–D1111, 2025.

[11] Antonio Lanzavecchia, Giandomenica Iezzi, and Antonella Viola. From tcr engagement to t cell activation: a kinetic view of t cell behavior. *Cell*, 96(1):1–4, 1999.

[12] David Zemmour, Rapolas Zilionis, Evgeny Kiner, Allon M Klein, Diane Mathis, and Christophe Benoist. Single-cell gene expression reveals a landscape of regulatory t cell phenotypes shaped by the tcr. *Nature immunology*, 19(3):291–301, 2018.

[13] Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023.

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[15] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pages 2023–09, 2024.

[16] Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. Language is all a graph needs. In *Findings of the association for computational linguistics: EACL 2024*, pages 1955–1973, 2024.