Revolutionize drug discovery with dense PPI data

Wei Lu
Aureka Biotechnologies
luwei@aurekabio.com

Lixia Yi
Fudan University
lxyi22@m.fudan.edu.cn

Jixian Zhang
Aureka Biotechnologies
jixian@aurekabio.com

Ming Gu
Aureka Biotechnologies
guming@aurekabio.com

Zhongyue Zhang
Shanghai Jiao Tong University
zhongyuezhang@sjtu.edu.cn

Jiahua Rao
Sun Yat-sen University
raojh7@mail.sysu.edu.cn

Shuangjia Zheng
Shanghai Jiao Tong University
shuangjia.zheng@sjtu.edu.cn

1 Overview

Drug development is costly, and the inherent tradeoff between efficacy, safety, and developability demands exploration of a vast sequence space to identify candidates that excel across all three dimensions [1]. Yet this search is limited by the absence of an appropriate foundation model.

AlphaFold [2, 3] has advanced the field, but only partially: it can show how two proteins interact once binding occurs, but it cannot reliably predict whether they will bind in the first place [4]. This distinction is critical for drug discovery, where the central challenge is to design molecules that bind strongly to a desired epitope on a target antigen while avoiding unwanted off-target interactions. Protein language models [5, 6] capture the sequence landscape, but their focus is on intrinsic properties such as stability rather than extrinsic properties like protein–protein interactions (PPI). The field therefore urgently needs a model that can accurately capture PPI, but existing datasets are too *sparse* [7] or too small [8] to support training at scale. A schematic illustrating the difference between *sparse* and **dense** PPI data is shown in Fig. 1.

We propose a solution to a fundamental challenge in drug discovery by creating a new kind of dataset: **a dense PPI dataset**, which can be produced rapidly, at scale, and at a low cost. Instead of sparse, memorization-prone data from natural protein pairs, our dense datasets will systematically sample millions of mutated protein pairs. **This isn't just more data; it's a new way of thinking**. Our dense datasets will enable the training of a PPI-specific foundation model that can learn the transferable physics of PPIs and explore a much larger, interaction-aware sequence landscape. By making such exploration possible, the model could help overcome the long-standing tradeoffs between efficacy, developability, and safety, ultimately transforming drug development.

2 AI task definition

We frame the problem as a **prediction** task. In protein design, sequence generation faces the fundamental challenge of lacking a ground-truth oracle, so benchmarking relies on indirect metrics such as recovery rate, diversity, or surrogate model scores, which are intrinsically biased [9]. A model with 100 percent recovery is not necessarily better, and very high diversity does not guarantee quality. The only fair and informative way to evaluate generation is through prospective experiments, but these are slow, costly, and introduce bias if reused for newer models. In contrast, a supervised prediction task relies on fixed held-out labels and standard metrics, enabling clear and reproducible comparisons without additional experiments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

Task. Given the amino acid sequences of two proteins, for example an antibody and its antigen, predict their **binding affinity**.

3 Dataset Rationale

Existing high-throughput PPI resources such as STRING [7] provide valuable breadth but are too *sparse* to support learning transferable interaction physics. These datasets largely reflect natural proteins paired with their native partners, offering limited signal for model training. A natural protein typically interacts with only a few partners, and even those can differ substantially by binding to different sites. As a result, the data is inherently *sparse*. **In the extreme case where a protein binds to only one partner, a model can appear to perform well by simply memorizing that sequence rather than learning the underlying physics of protein–protein interactions.** Binding affinity measurements from such pairs thus contribute little to generalizable models, as the learning signal is dominated by pair-specific identity rather than transferable interaction principles.

In contrast, dense PPI datasets systematically sample large mutational neighborhoods around a given protein–protein pair, generating millions of related variants. Many mutations are benign and cause little change to the protein, but some, even a single substitution, can alter affinity by hundreds of folds and lead to unwanted, disease-causing interactions [10]. Dense PPI forces models to go beyond memorizing pair identities and instead learn how sequence variation affects binding affinity. Such datasets provide the foundation for building generalizable models of protein–protein interactions, enabling more accurate affinity prediction and better modeling of the interaction-aware protein sequence landscape. This is essential for drug design, since the therapeutic value of a molecule lies not only in its intrinsic properties, but also in how it influences other proteins.

Because existing dense PPI data is so limited, no model can yet accurately predict binding affinity between two proteins [11, 12]. AlphaFold's confidence score and certain force fields show some correlation with affinity, but the signal is not strong enough to support efficient exploration of large sequence spaces. Training on dense PPI data could bridge this gap and deliver the long-awaited capability for affinity prediction. As an added benefit, it may also enhance complex structure prediction by providing a new data source, in the same way that sequence databases improved structural modeling through multiple sequence alignments, since affinity and structure are inherently intertwined.

4 Data Creation Pathway

The dataset will be created through new experiments. A non-profit can coordinate generation and make the data openly available, prioritizing protein pairs of broad community interest. A common protocol is desirable, although strict standardization is not essential because supervision relies primarily on intra-assay comparisons among variants of the same pair.

Data Modality and Resolution We will combine fluorescence-activated cell sorting (FACS) with long-read sequencing. Both are mature, widely accessible techniques. Libraries will introduce variation in both proteins using random mutagenesis and designed combinatorial schemes. We can build libraries on the order of 10^8 variants. Each FACS plus sequencing round can yield about 10^7 reads and $\sim 10^6$ labeled data points. In analyses of comparable assays [13], technical repeats achieve Spearman correlation > 0.8, indicating reliable resolution for learning.

Cost and Scalability Generating approximately 10^6 labeled data points is estimated to cost about $\$1{,}000$. The workflow relies on standard molecular biology and flow cytometry, with no specialized technical barriers to scaling. As experimental volume grows, the cost per million data could drop further, potentially to only a few hundred dollars. By running thousands of FACS sorting experiments across hundreds of unique protein-protein pairs, we aim to accumulate at least 10^9 labeled data points, sufficient to train PPI foundation models that can efficiently explore the interaction-aware sequence space. Each data point consists of two protein sequences paired with an affinity value inferred from sequencing results. The total cost for generating 10^9 labeled data points is expected to be within a few million dollars.

5 Expected Impact

A robust interaction-aware foundation model trained on dense PPI data could revolutionize therapeutic drug discovery (antibodies, peptides, etc.), from hit generation to lead optimization, while also advancing virtual-cell modeling [14] by linking sequence, interaction, and phenotype. Beyond drug development, such a resource would enable efficient design of interacting proteins and provide mechanistic insights into cellular pathways and protein function. Ultimately, these capabilities could reshape drug discovery, diagnostics, synthetic biology, and the life sciences more broadly.

Acknowledgments and Disclosure of Funding

The authors declare that there are no competing interests associated with this work. W.L, J.Z, M.G are employees of Aureka Biotechnologies.

References

[1] Raymond J Deshaies. How multispecific molecules are transforming pharmacotherapy. *Nature Reviews Drug Discovery*, pages 1–13, 2025.

[2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[4] Matthew J Styles, Joshua A Pixley, Tongyao Wei, Christopher Basile, Shannon S Lu, and Bryan C Dickinson. Pancs-binders: a rapid, high-throughput binder discovery platform. *Nature Methods*, 22(8):1720–1730, 2025.

[5] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature biotechnology*, 42(2):275–283, 2024.

[6] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[7] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

[8] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[9] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[10] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

[11] Wei Lu, Jixian Zhang, Jiahua Rao, Zhongyue Zhang, and Shuangjia Zheng. Alphafold3, a secret sauce for predicting mutational effects on protein-protein interactions. *bioRxiv*, pages 2024–05, 2024.

[12] Aerin Yang, Kevin M Jude, Ben Lai, Mason Minot, Anna M Kocyla, Caleb R Glassman, Daisuke Nishimiya, Yoon Seok Kim, Sai T Reddy, Aly A Khan, et al. Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science*, 381(6656):eadh1720, 2023.

[13] Mason Minot and Sai T Reddy. Meta learning addresses noisy and under-labeled data in machine learning-guided antibody engineering. *Cell Systems*, 15(1):4–18, 2024.

[14] Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025.

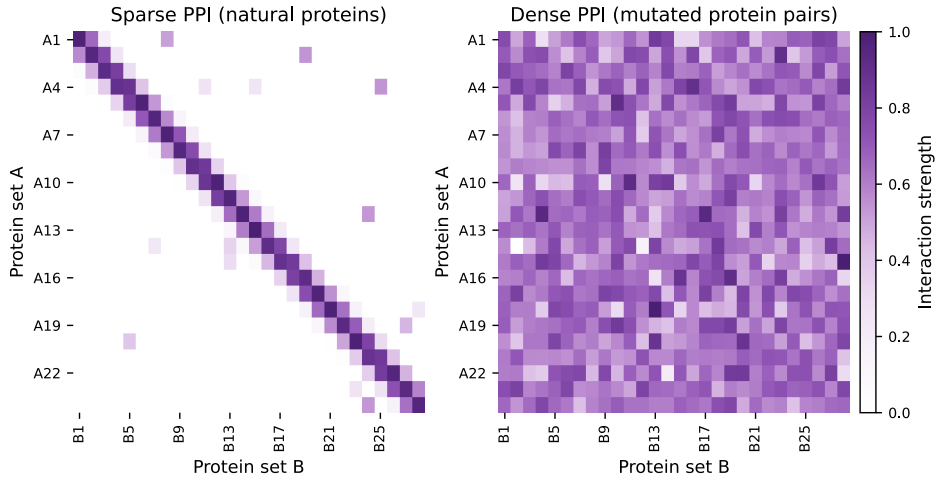

Figure 1: Comparison of sparse versus dense protein–protein interaction (PPI) datasets. Sparse datasets, such as those derived from natural PPIs, provide limited interaction information. Because each protein pair is typically unrelated to the others, models tend to memorize protein-pair identities rather than learn the underlying physics of interaction. In contrast, dense datasets systematically vary both partners, forcing models to learn mutational effects. This signal is more transferable and enables the training of generalizable models that capture the principles of protein–protein interactions.