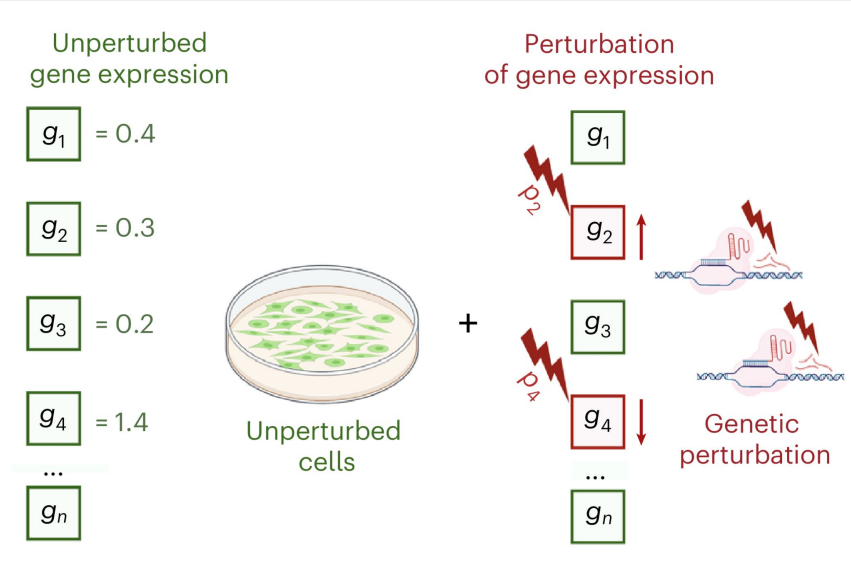
Adaptive Data-Knowledge Alignment in Genetic Perturbation Prediction

Yuanfang Xiang
Nanjing University

Lun Ai
EMBL-EBI

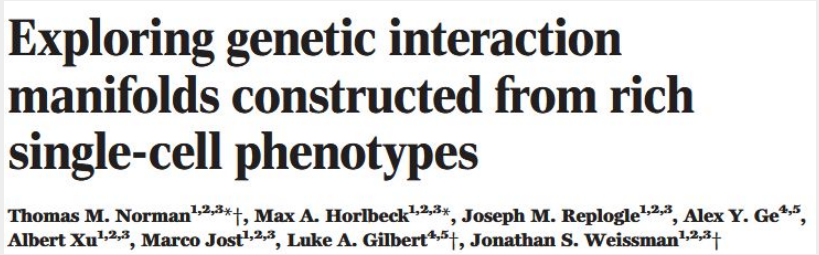
Genetic Perturbation Response



(from Roohani et al. 2024)

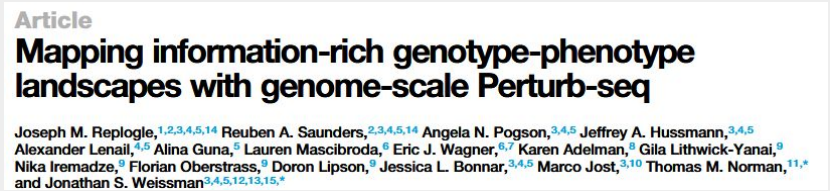
1. **Understand biological complexity**
2. **Biomedical importance**

Unprecedented experimental capabilities


Exploring genetic interaction manifolds constructed from rich single-cell phenotypes

Thomas M. Norman[1,2,3]*†, Max A. Horlbeck[1,2,3]*, Joseph M. Replogle[1,2,3], Alex Y. Ge[4,5], Albert Xu[1,2,3], Marco Jost[1,2,3], Luke A. Gilbert[4,5]†, Jonathan S. Weissman[1,2,3]†

89,000 samples, 102 perturbed genes


Article

Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq

Joseph M. Replogle,[1,2,3,4,5,14] Reuben A. Saunders,[2,3,4,5,14] Angela N. Pogson,[3,4,5] Jeffrey A. Hussmann,[3,4,5] Alexander Lenail,[4,5] Alina Guna,[5] Lauren Mascibroda,[6] Eric J. Wagner,[6,7] Karen Adelman,[8] Gila Lithwick-Yanai,[9] Nika Iremadze,[9] Florian Oberstrass,[9] Doron Lipson,[9] Jessica L. Bonnar,[3,4,5] Marco Jost,[3,10] Thomas M. Norman,[11,*] and Jonathan S. Weissman[3,4,5,12,13,15,*]

2.5+ million cells

But, existing methods:

1. Black-box learners

e.g. CPA, Geneformer, scGPT, scFoundation

2. Using static knowledge

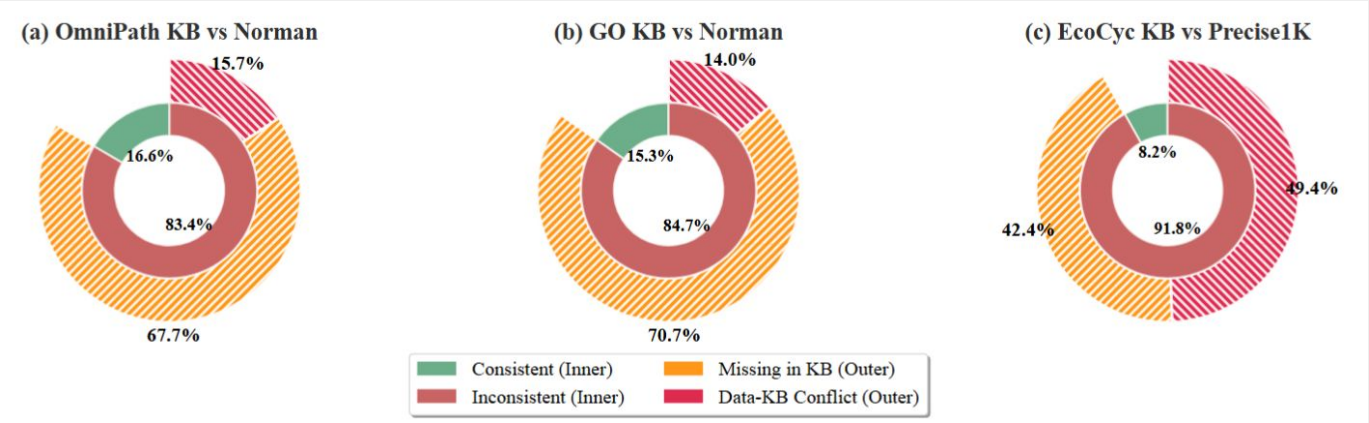
e.g. GEARS, TxPert, scLAMBDA, PRESAGE

“Deep learning-based predictors cannot help us better understand cells” (Peidli et al. 2024),

Mechanistic interpretability of predictions &

Continual refinement of established knowledge?

Challenge: data-knowledge inconsistency



- Perturbation datasets: *Norman et al. 2019 & Lamoureux et al. 2023*
- Prior knowledge bases: *OmniPath, Gene Ontology (GO) and EcoCyc*

Simultaneously imperfect sources

We cannot find the ground-truth perturbation response

To overcome this challenge, we need:

- **Balanced evaluation as neither sides represent *the ground-truth***
- **Align inconsistent data- and knowledge-driven predictions**
- **Select reliable signal to refine imperfect prior knowledge**

To overcome this challenge, we need:

- **Balanced evaluation**
- **Align data- and knowledge-driven predictions**
- **Refine imperfect prior knowledge**

Balancing consistency between data and knowledge

Balanced cons. ***Data cons.*** ***Knowledge cons.***

$$F_{1\,\text{balance}}(f(x), x, y, \mathcal{KB}) = \left(\frac{1}{2} F_1(y, f(x))^{-\gamma} + \frac{1}{2} F_1(\delta_{\mathcal{KB}}(x), f(x))^{-\gamma}\right)^{-1/\gamma}$$

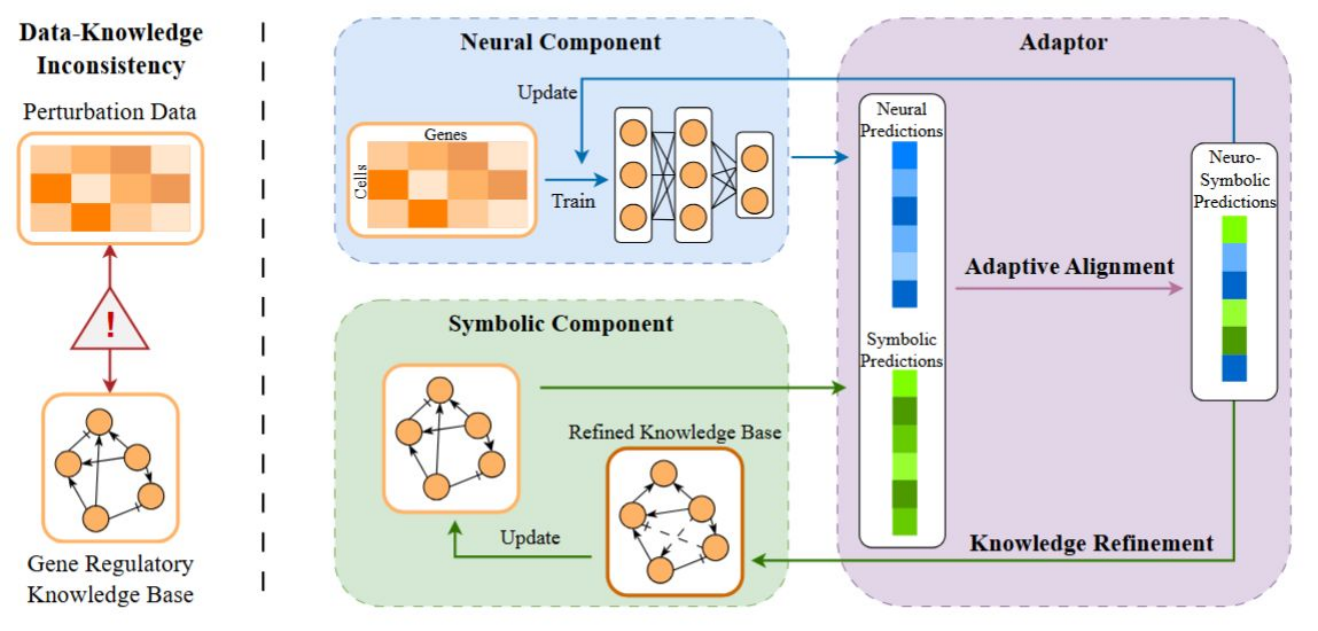
γ penalises low, unequal data and knowledge cons.

- **Data cons. - predictions' generalization on test data**
- **Knowledge cons. - extent of grounded in knowledge**

To overcome this challenge, we need:

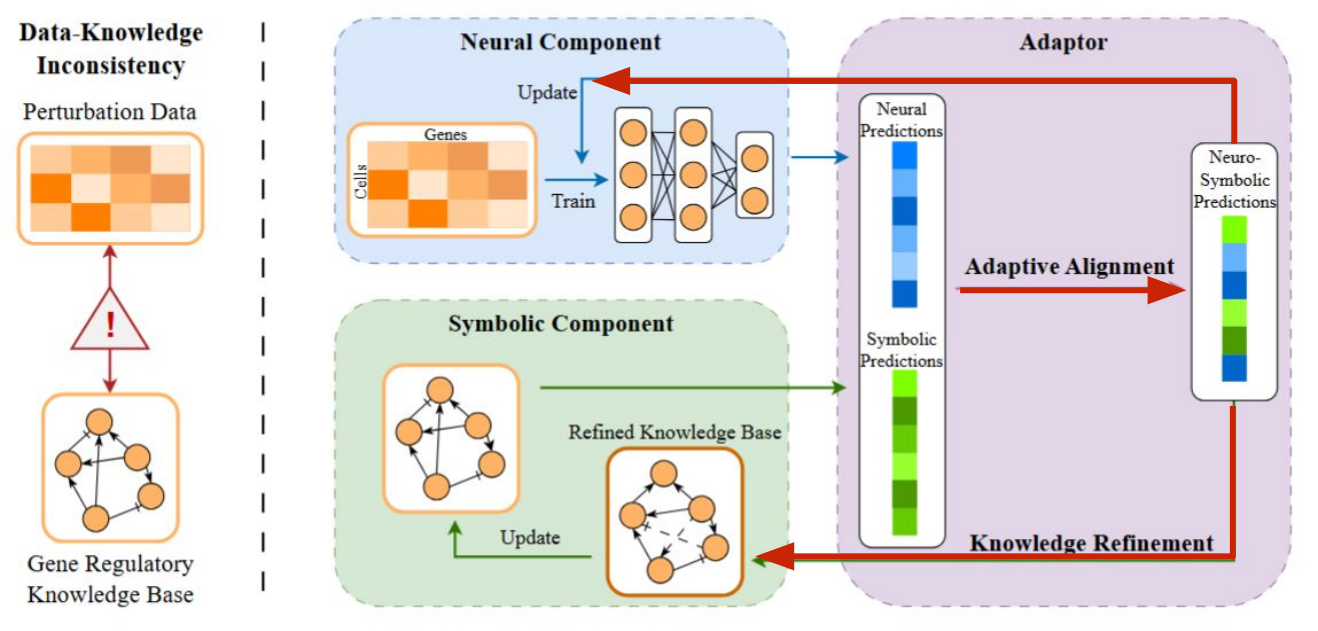
- **Balanced evaluation**
- **Align data- and knowledge-driven predictions**
- **Refine imperfect prior knowledge**

ALIGNED (Adaptive aLignment for Inconsistent Genetic kNowledgE and Data)



Neural predictor & symbolic knowledge simulator, integrated via the adaptor

ALIGNED (Adaptive aLignment for Inconsistent Genetic kNowledgE and Data)



Neuro-symbolic alignment & Bidirectional update

Adaptively aligning neural & symbolic predictions

The aligned prediction should:

- **Be more consistent with the knowledge base**
- **Use more data-derived information**
- **Use more well represented interactions in knowledge**

$$\min_{f_y, f_a} \mathcal{L} = \frac{1}{|D_l|} \sum_{(x,y)\in D_l} CE(f_y(x), y)$$
$$+C\frac{1}{|D_l \cup D_u|} \sum_{x\in D_l \cup D_u} L_a(a, x, \hat{y}) \log f_a(x)$$

The adaptor loss combines neural and symbolic predictions (Williams, 1992; Hu et al., 2025)

Efficiently updating knowledge via sparse regularization

Continuous approximation: computationally efficient, gradient-based update

Boolean matrix approximation (Ravanbakhsh et al., 2016)

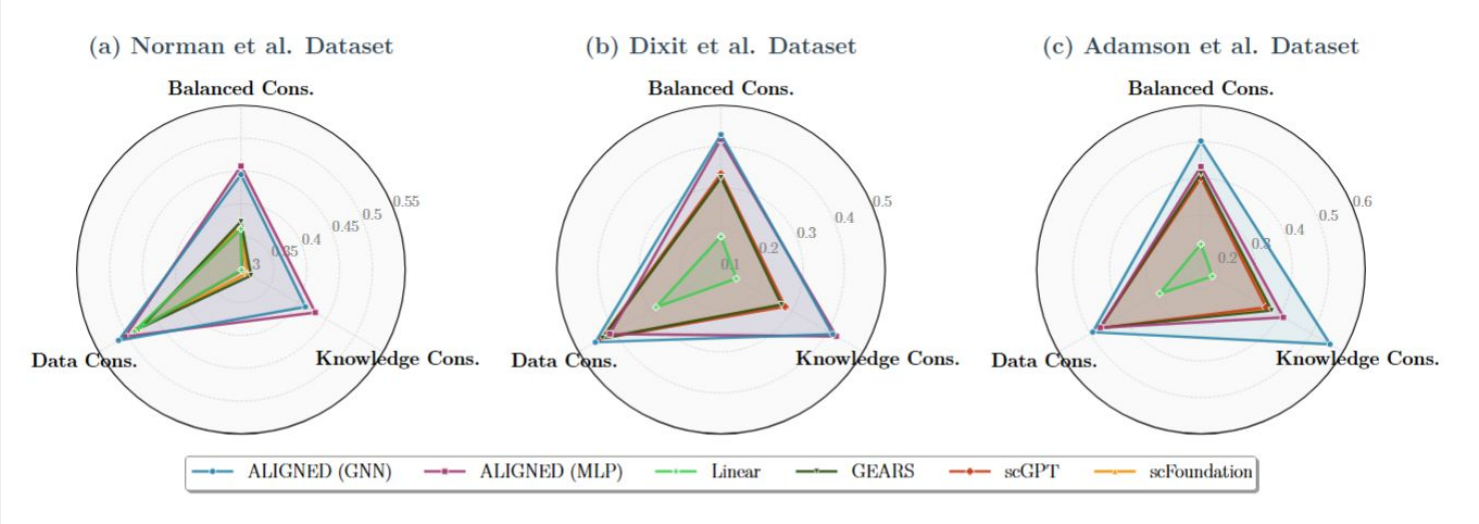
$$\varepsilon_t(X)_{i,j} = 1 - \exp(-tX_{i,j}), X_{i,j} \geq 0$$

Sparse regularization: keep the structure of refined knowledge base

Refinement loss

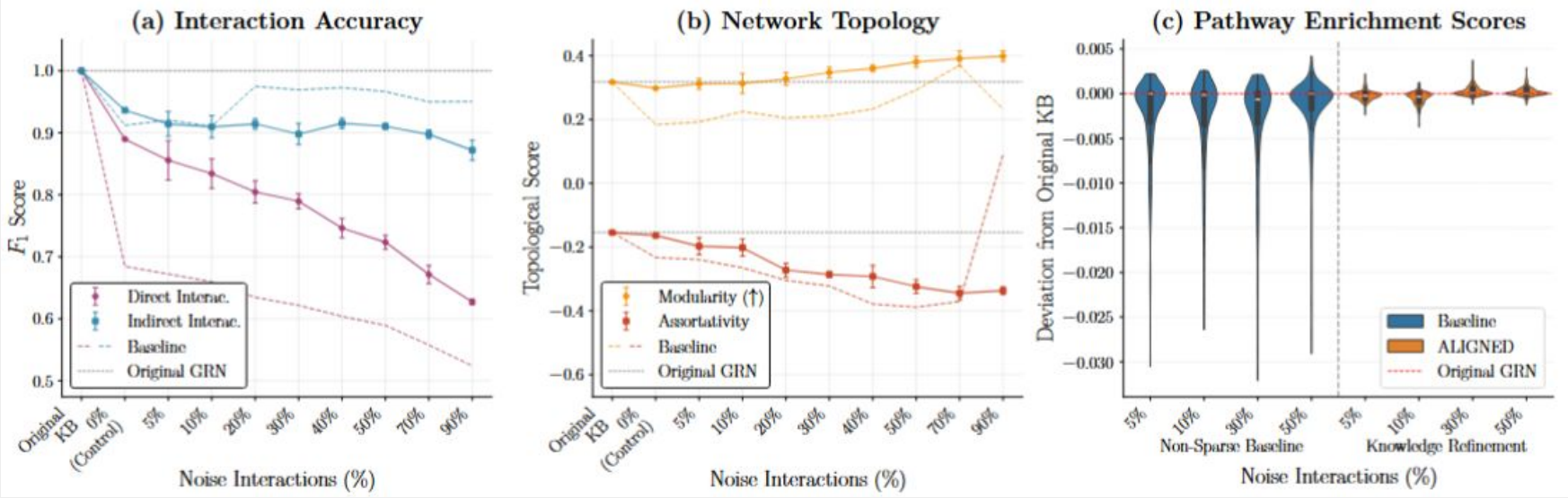
$$\min_{P_+^{(0)}, P_-^{(0)} \in R_+^{n \times n}} \mathcal{L}_{\text{refine}}(P_+^{(0)}, P_-^{(0)}, k) = \sum_{x,y \in \langle X_u, \bar{Y} \rangle} \|\varepsilon_{t_k}(P_+^{(k)} - P_-^{(k)})^\top x - y\|_2^2 + \lambda\left(\|\varepsilon_{t_0}(P_+^{(0)}) - R_+^{(0)}\|_1 + \|\varepsilon_{t_0}(P_-^{(0)}) - R_-^{(0)}\|_1\right)$$

ALIGNED Improves Balanced Consistency without Damaging Prediction Performance



Improved knowledge consistency while keeping / improving data consistency

ALIGNED Re-Discovers Biologically Meaningful Regulatory Knowledge from Synthetic Data



Accurately recovered corrupted knowledge base

Preserves **well-structured** regulatory interactions

Re-discovered **meaningful** cross-reference knowledge

Future work

- Integrate with self-driving lab workflow
- Differentiable knowledge modelling (Faure et al., 2023)
- Protein-protein interaction and metabolic networks