Benchmarking perturbation predictions

Wolfgang Huber
Constantin Ahlmann-Eltze

21.11.2025

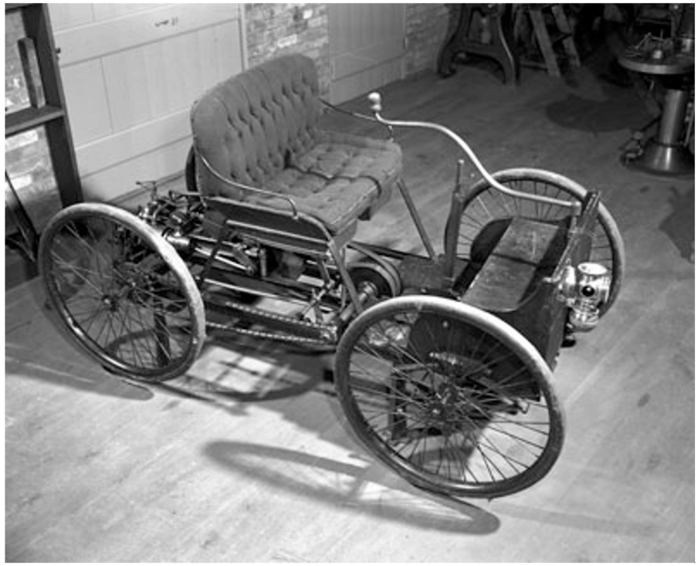
@wkhuber.bsky.social
@const-ae.bsky.social

github.com/const-ae/

What *is* quality?

Many definitions. E.g.:

- adherence to specifications
- fitness for purpose

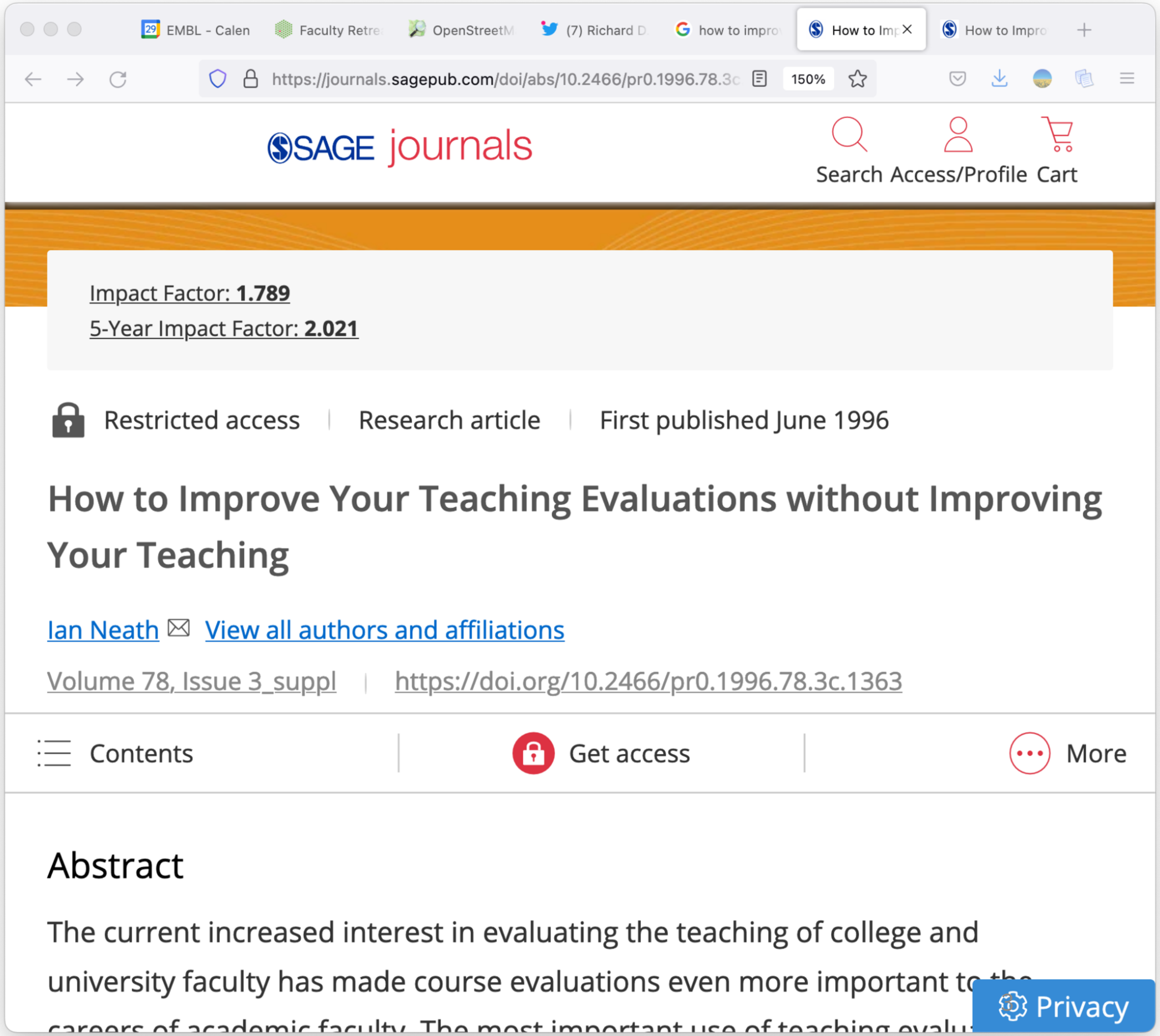


Henry Ford (possibly apocryphal):

''If I had asked people what they wanted, they would have said faster horses.''

Goodhart's law

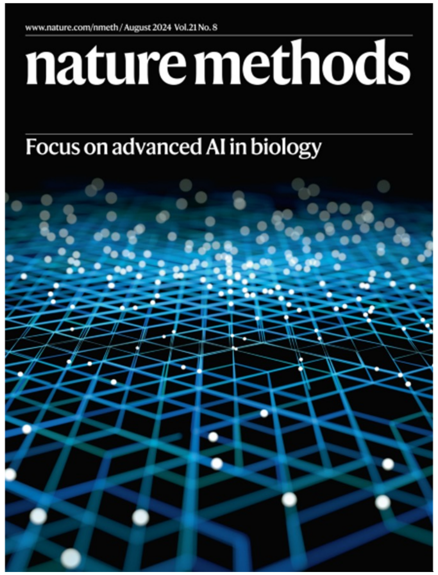
when a measure becomes a target, it ceases to be a good measure



@const-ae.bsky.social | @wkhuber.bsky.social

Benchmarking is really difficult

- Not even a matter of “ground truth”
- Usefulness (*’All models are wrong but some are useful’*)
- But what is *useful*?



nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature methods > articles > article

Article | Published: 06 June 2024

Large-scale foundation model on single-cell transcriptomics

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma ✉, Xuegong Zhang ✉ & Le Song ✉

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature methods > articles > article

Article | Published: 26 February 2024

scGPT: toward building a foundation model for single-cell multi-omics using generative AI

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan & Bo Wang ✉

nature biotechnology

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature biotechnology > articles > article

Article | Open access | Published: 17 August 2023

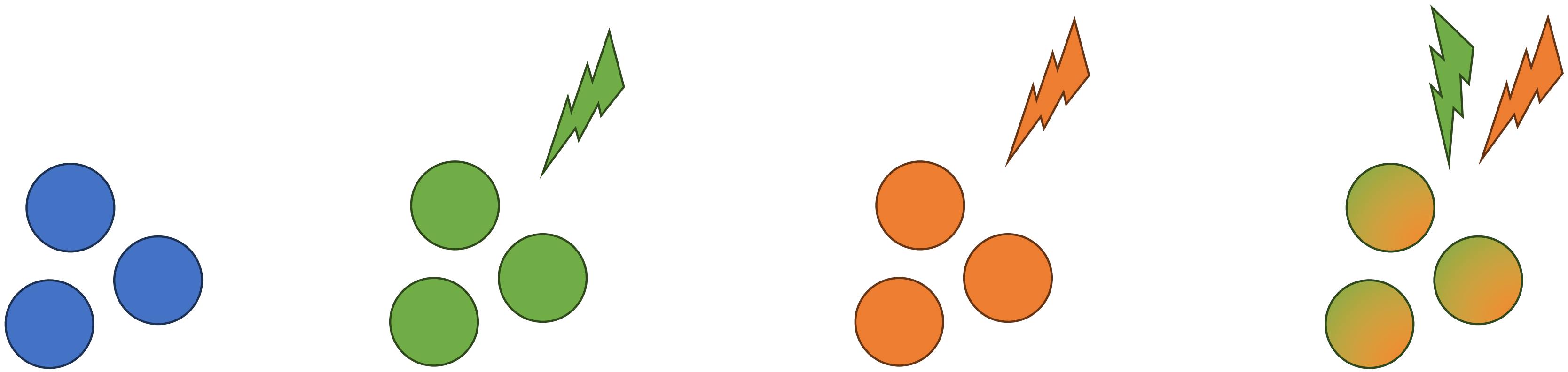
Predicting transcriptional outcomes of novel multigene perturbations with GEARS

Yusuf Roohani, Kexin Huang & Jure Leskovec ✉

How to predict the effect of unseen perturbations?

- Cell fate
- Cell morphology
- Metabolome
- **Transcriptome**

- Unseen drugs
- Unseen cell types
- **Unseen single perturbations**
- **Double perturbations**



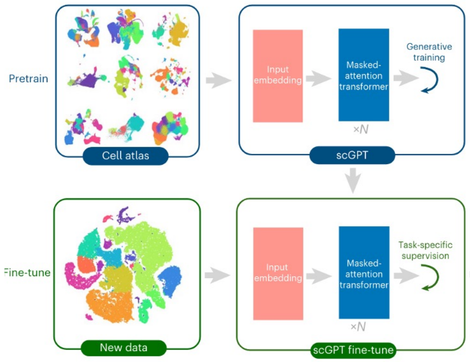
GENOMICS

Exploring genetic interaction manifolds constructed from rich single-cell phenotypes

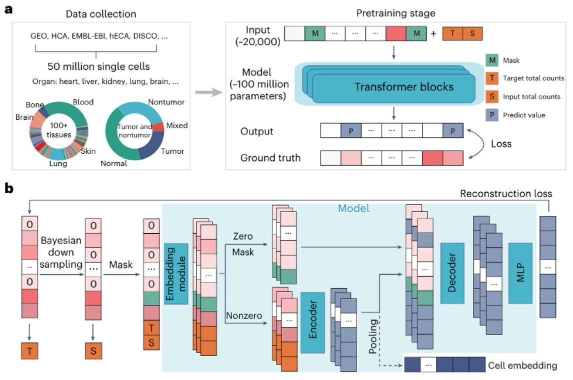
Thomas M. Norman[1,2,3*†], Max A. Horlbeck[1,2,3*], Joseph M. Replogle[1,2,3], Alex Y. Ge[4,5], Albert Xu[1,2,3], Marco Jost[1,2,3], Luke A. Gilbert[4,5†], Jonathan S. Weissman[1,2,3†]

- K562 cell line
- CRISPR activation
- 101 single perturbations + 62 double perturbations
- 110,000 cells

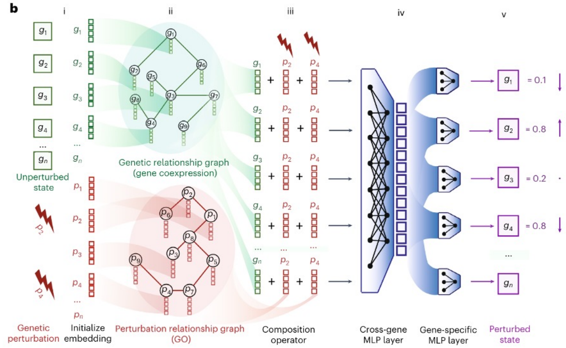
scGPT



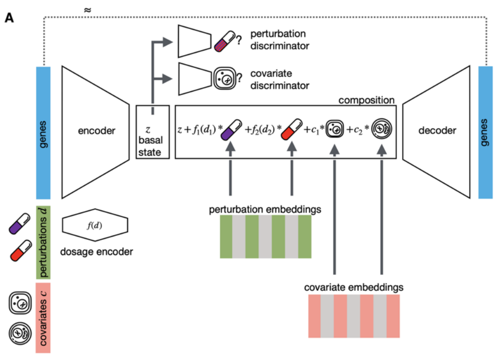
scFoundation



GEARS



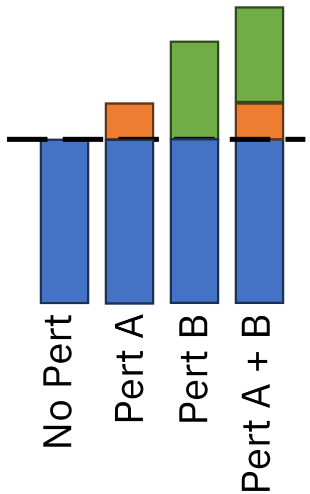
CPA



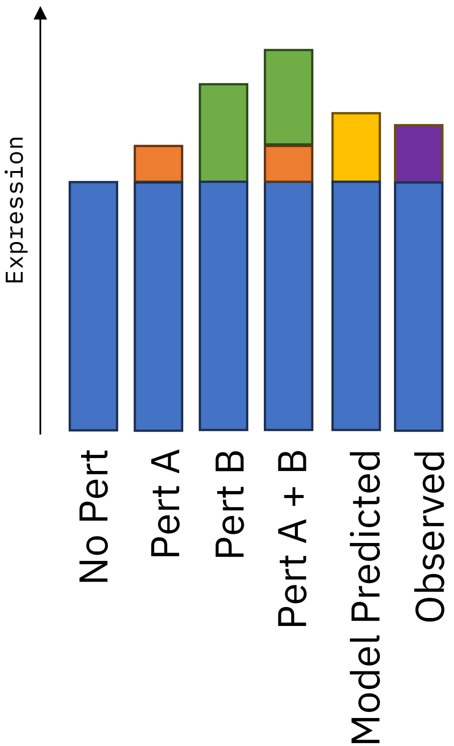
Additional models:

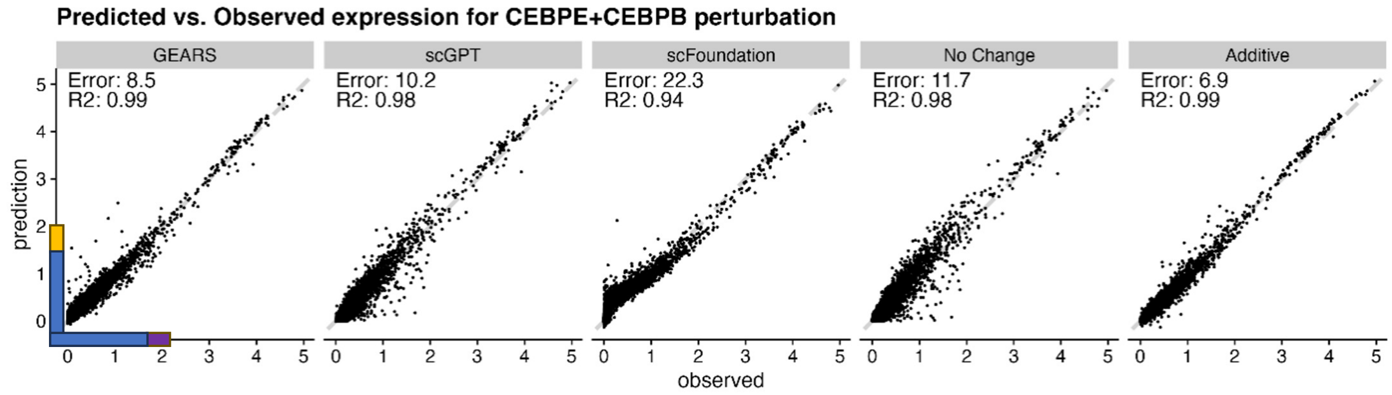
- Geneformer
- UCE
- scBert

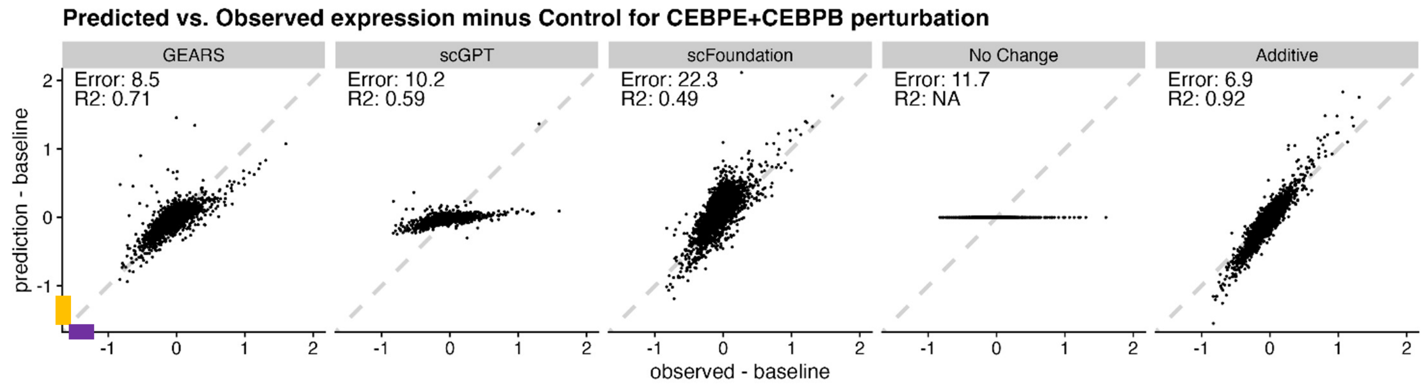
Baseline



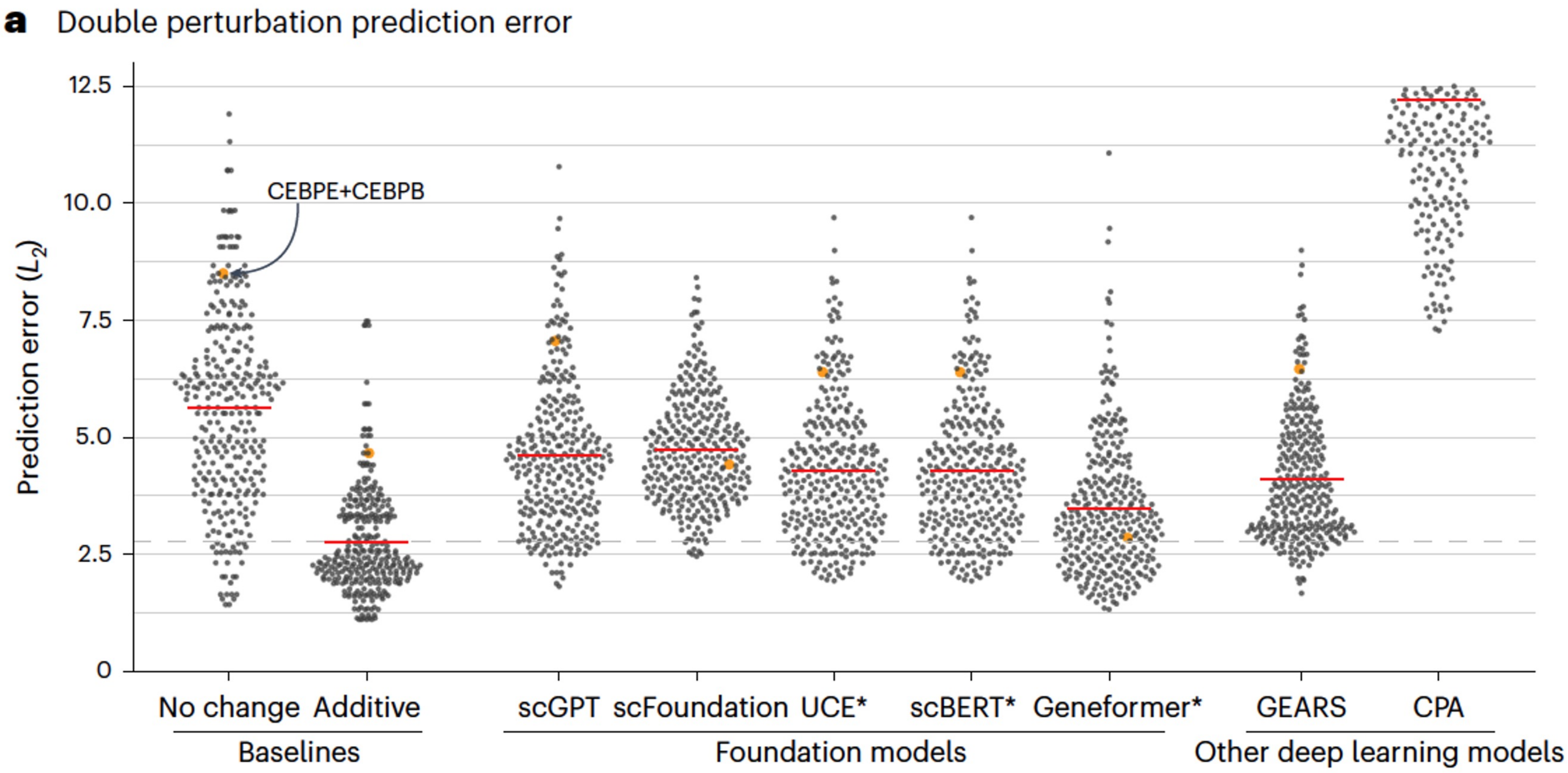
Comparison of prediction errors



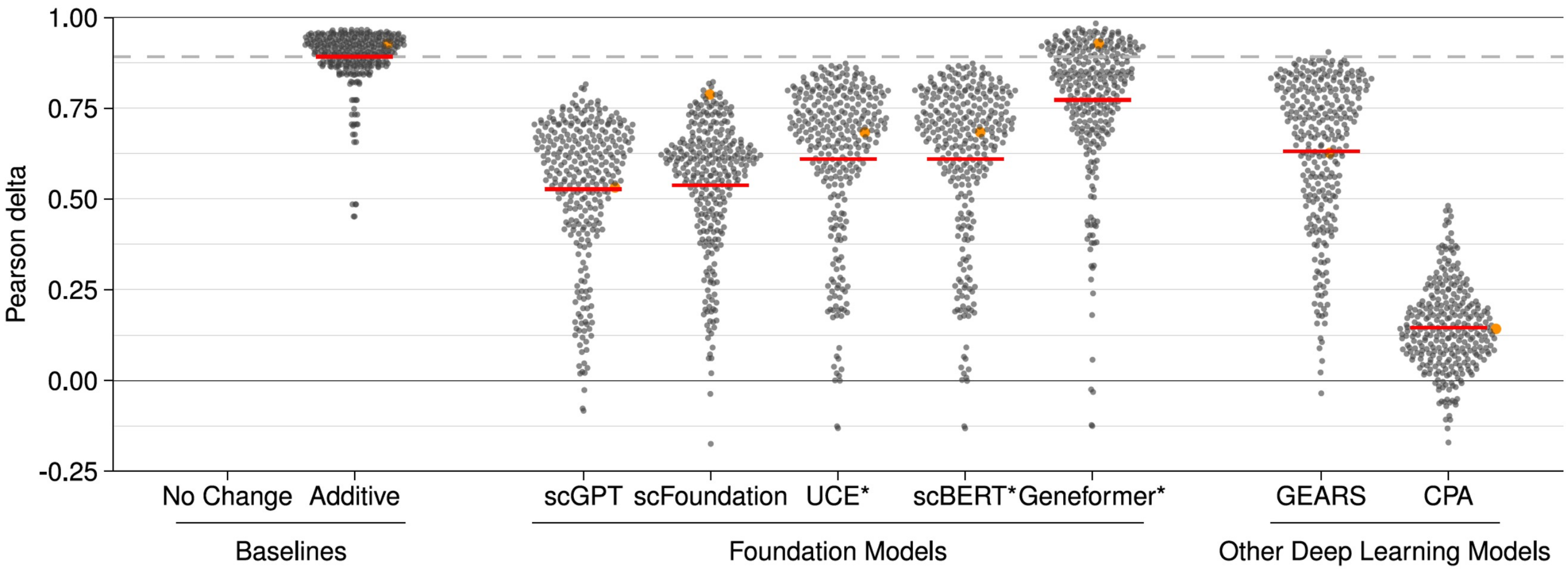




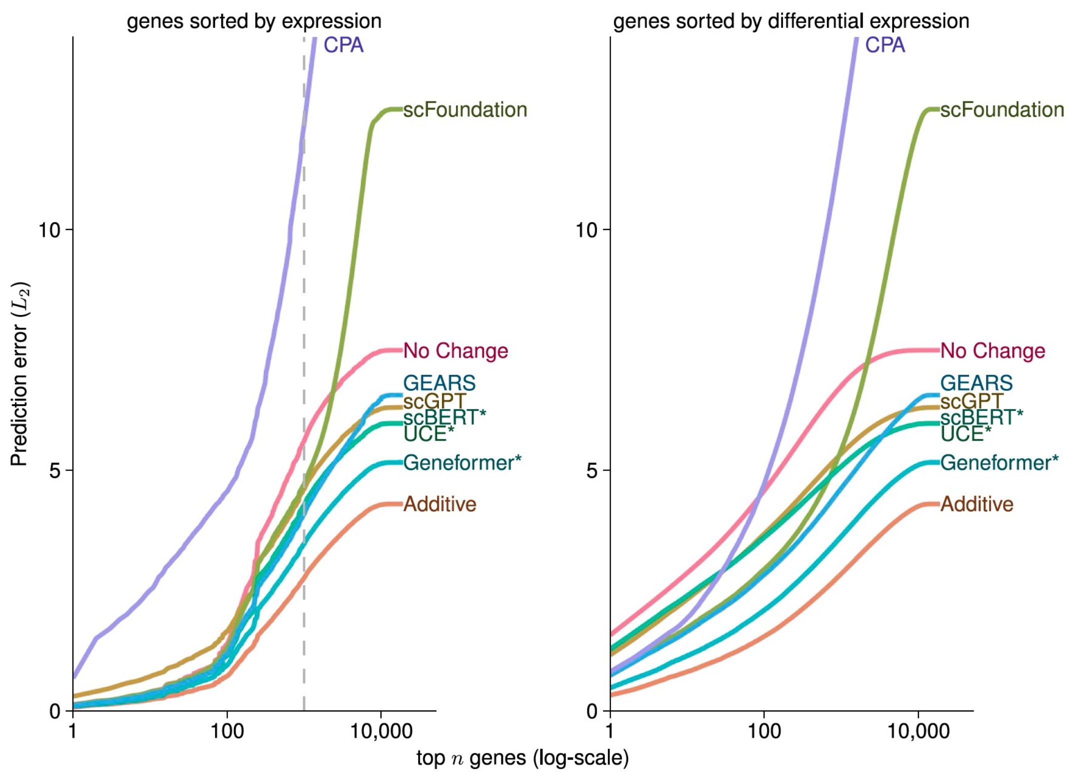
Average prediction error for tested models was larger than for the additive baseline



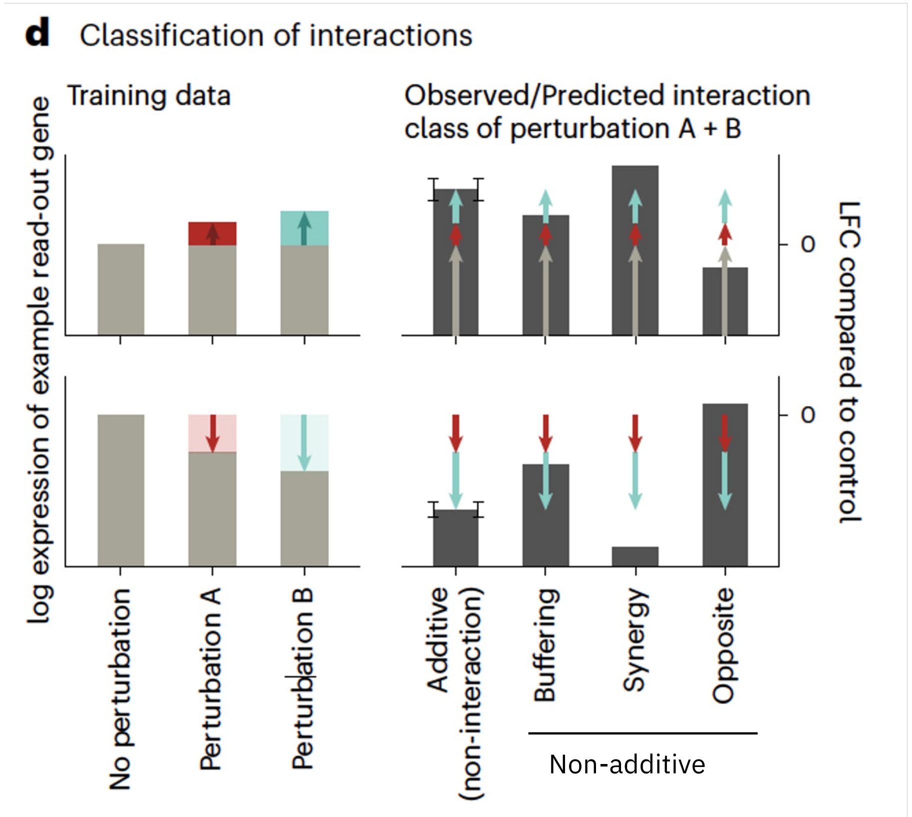
Average correlation for tested models was lower than for the additive baseline

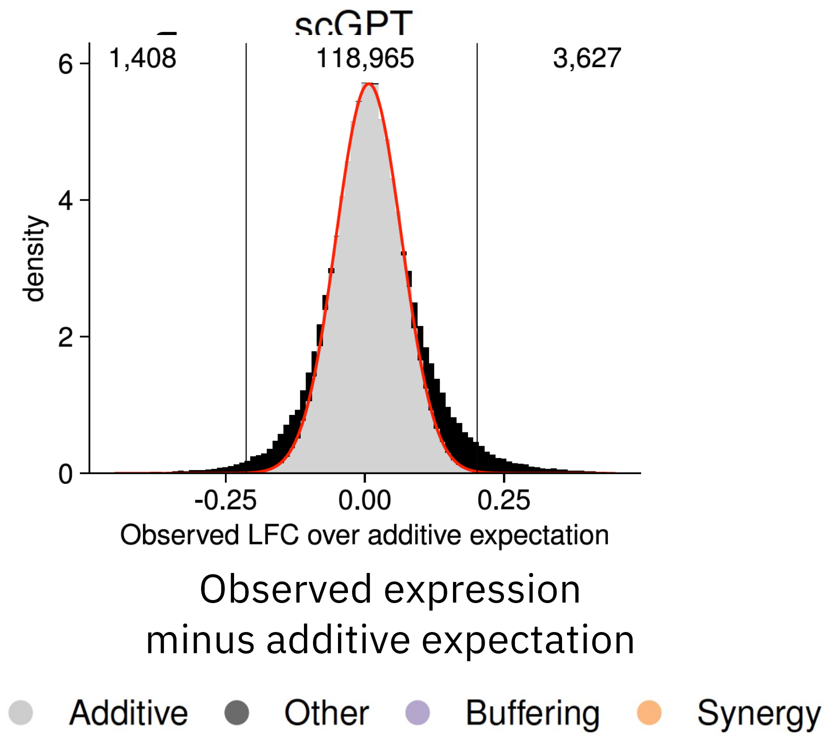


And the effect is robust across different read-out gene sets

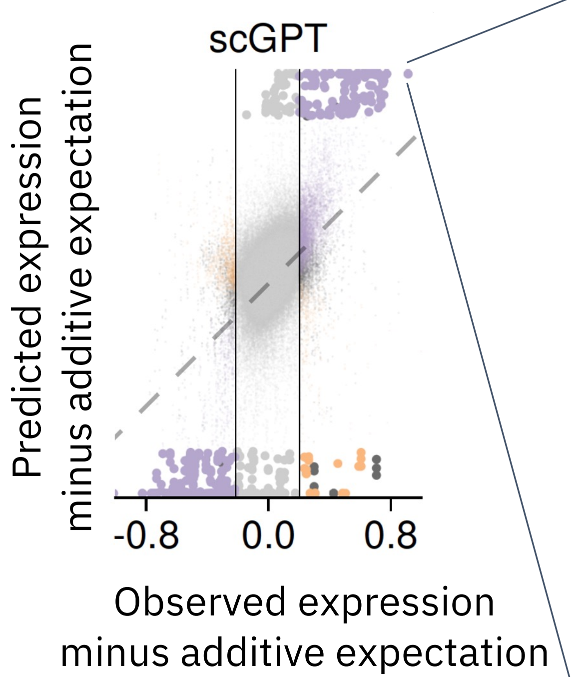


But what about prediction of non-additive effects ($\Delta AB \neq \Delta A + \Delta B$)

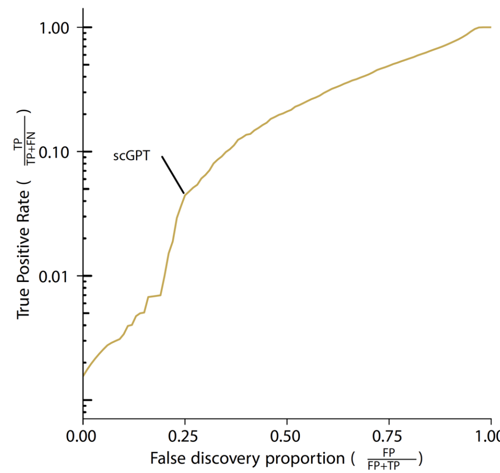




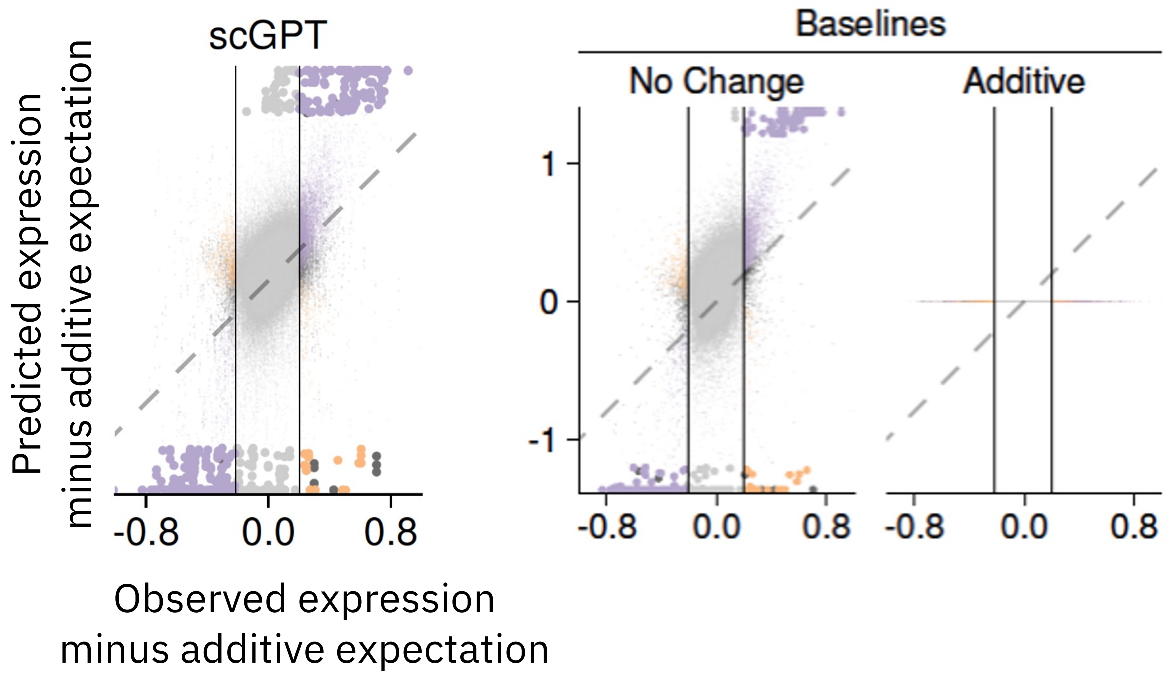
We can count how many of the most non-additive predictions are actually non-additive

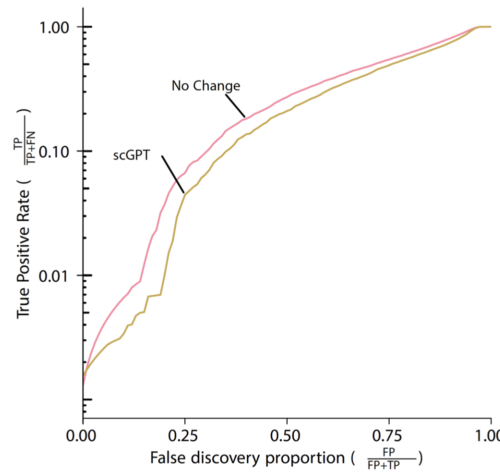


	perturbation	gene_name	additive_model	observed	prediction	tp	fp	`fp/(fp+tp)`
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>
1	CEBPB+MAPK1	HBG2	-0.160	0.537	2.27	1	0	0
2	SNAI1+DLX2	HBZ	0.166	0.752	2.56	2	0	0
3	CEBPB+CEBPA	HBG2	0.00883	0.782	2.14	3	0	0
4	ZBTB10+SNAI1	HBZ	0.384	0.791	2.50	4	0	0
5	CEBPE+CEBPB	HBG2	0.139	0.860	2.17	5	0	0
6	SNAI1+DLX2	HBG2	0.317	0.784	2.29	6	0	0
7	CEBPE+RUNX1T1	HBZ	0.670	1.03	2.64	7	0	0
8	ZC3HAV1+CEBPA	HBZ	0.767	0.970	2.64	8	0	0
9	FOXF1+FOXL2	HBZ	0.740	0.919	2.60	8	1	0.111
10	CEBPB+CEBPA	GYPB	-0.0275	0.459	1.72	9	1	0.1
11	CEBPB+MAPK1	HBZ	0.835	0.950	2.56	9	2	0.182
12	CEBPB+CEBPA	AVP	0.0262	0.101	1.74	9	3	0.25
13	CEBPE+RUNX1T1	HBG2	0.634	0.934	2.27	10	3	0.231
14	FOXL2+MEIS1	HBZ	0.975	1.16	2.59	10	4	0.286
15	CEBPE+CEBPB	GYPB	0.0971	0.595	1.69	11	4	0.267
16	ZC3HAV1+CEBPA	HBG2	0.768	0.969	2.33	12	4	0.25
17	FOSB+OSR2	HBZ	0.854	0.919	2.42	12	5	0.294
18	ZBTB10+SNAI1	HBG2	0.698	0.794	2.23	12	6	0.333
19	CEBPB+CEBPA	HBZ	1.00	0.985	2.53	12	7	0.368
20	CEBPE+CEBPB	HBZ	0.992	1.06	2.49	12	8	0.4
21	CEBPE+CEBPB	AVP	0.0503	0.0833	1.54	12	9	0.429
22	ZC3HAV1+CEBPA	GYPB	0.294	0.579	1.70	13	9	0.409
23	POU3F2+FOXL2	HBZ	1.14	1.39	2.54	14	9	0.391
24	FOSB+OSR2	HBG2	0.870	0.858	2.24	14	10	0.417
25	CEBPE+SPI1	AVP	0.0608	0.0453	1.43	14	11	0.44
26	CEBPB+MAPK1	HBG1	-0.170	0.167	1.19	15	11	0.423
27	CEBPB+CEBPA	GYPA	-0.226	0.351	1.13	16	11	0.407
28	CEBPB+CEBPA	HBG1	-0.0917	0.278	1.23	17	11	0.393
29	CEBPE+RUNX1T1	GYPB	0.348	0.682	1.65	18	11	0.379
30	CEBPE+CNN1	AVP	0.0469	0.0563	1.34	18	12	0.4
31	CEBPB+PTPN12	AVP	0.0328	0.0647	1.31	18	13	0.419
32	CEBPE+CEBPB	GYPA	-0.135	0.424	1.15	19	13	0.406
33	ZBTB10+SNAI1	AVP	0.0166	0.0285	1.29	19	14	0.424
34	CEBPB+MAPK1	GYPB	0.357	0.775	1.62	20	14	0.412

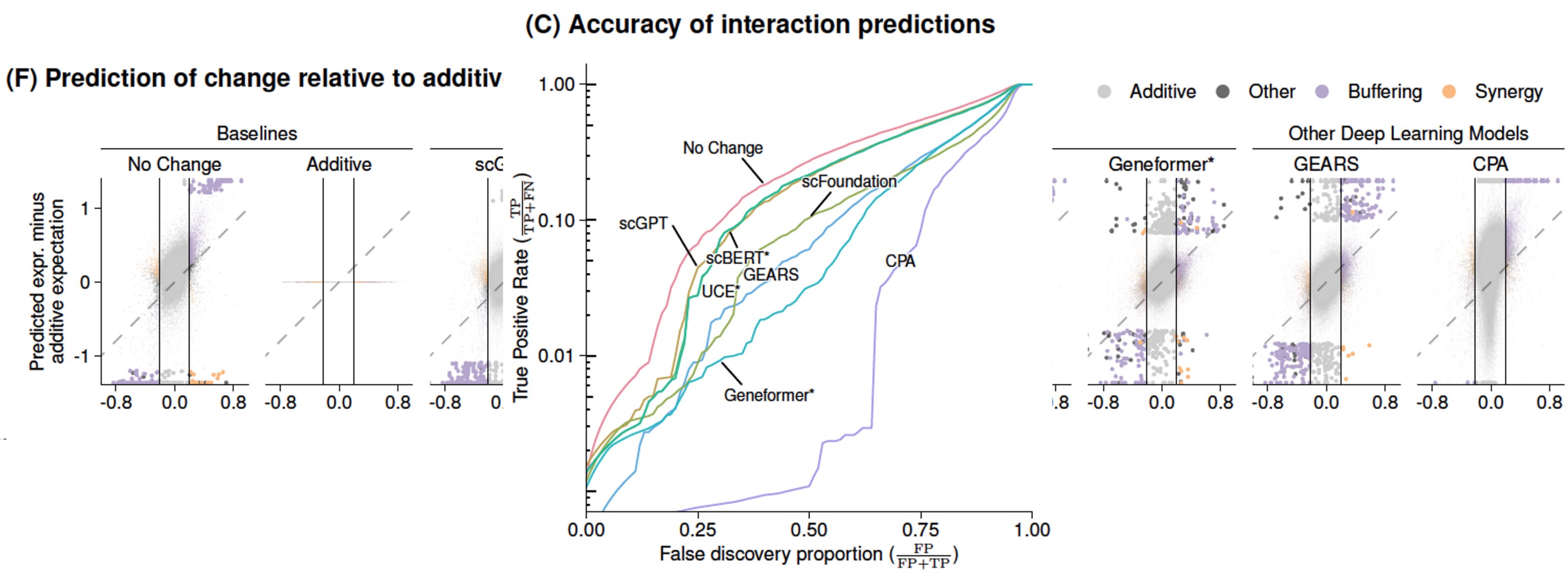


scGPT finds fewer non-additive expression changes than the no-change baseline

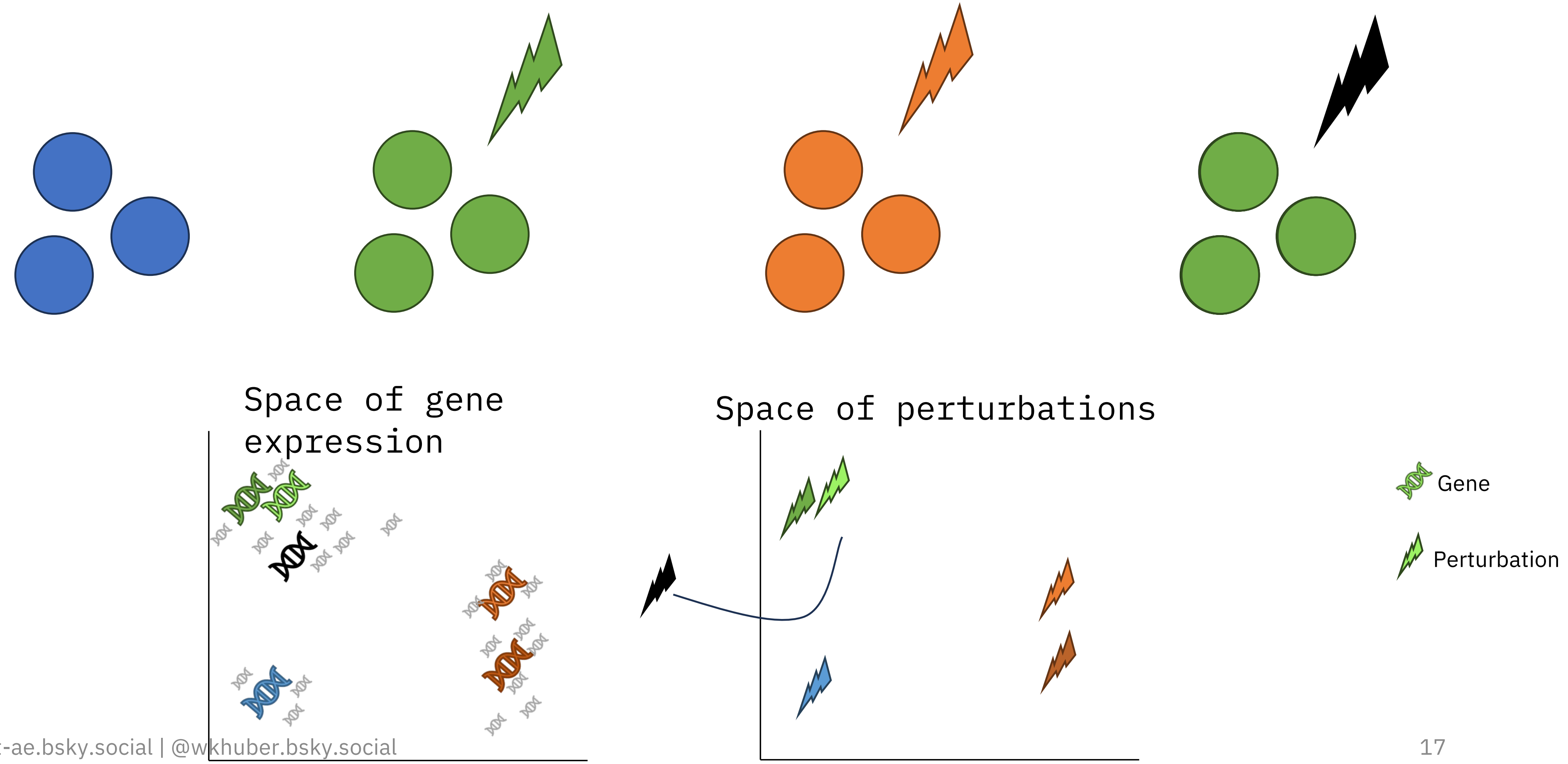




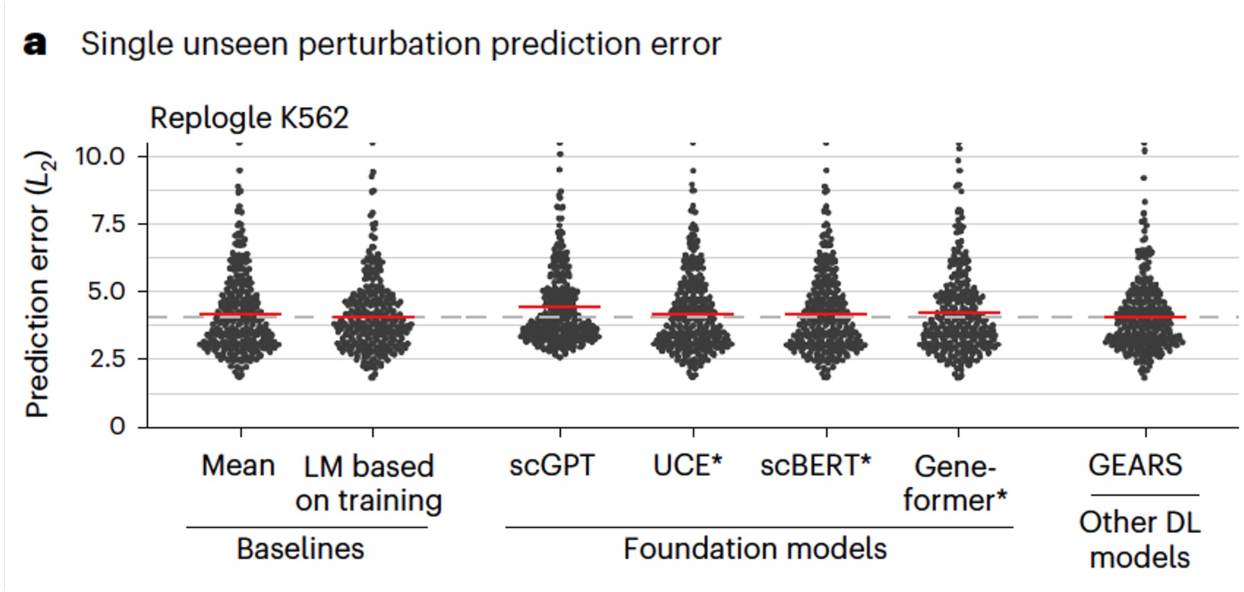
All methods perform worse at identifying non-additive interactions than the no-change baseline



How to predict the effect of unseen gene perturbations?

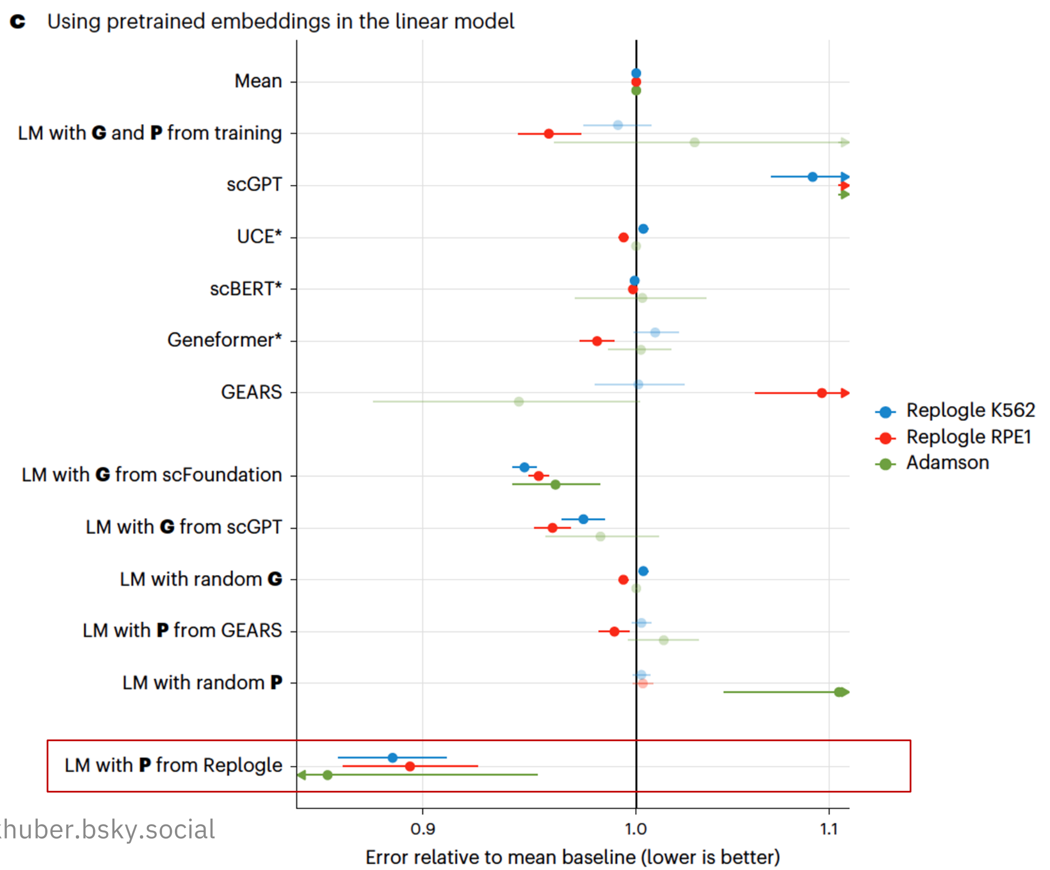


A linear model performs as good or better than the deep learning models for single perturbation prediction



Same trends for Replogle RPE1 and Adamson

Pre-training on another perturbation dataset increases performance



Evaluating the Utilities of Foundation Models in Single-cell Data Analysis

Tianyu Liu[1,2], Kexing Li[1,2], Yuge Wang[2], Hongyu Li[2], Hongyu Zhao[1,2*]

[1]Interdepartme...
Bioinfo...
[2]Department ...

A Systematic Comparison of Single-Cell Perturbation Response Prediction Models

Lanxiang Li[1,2†], Yue You[1*†], Wenyu Liao[3†], Xueying Fan[4,5,6,7], Shihong Lu[1], Ye Cao[1], Bo Li[1], Wenle Ren[1], Yunlin Fu[1], Jiaming Kong[8], Shuangjia Zheng[9], Jizheng Chen[1,10], Xiaodong Liu[4,5,6,7], Luyi Tian[1,2*]

A systematic comparison of computational methods for expression forecasting

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, Patric...

doi: https://doi.org/10.1101/2023.07.28.551039

Abstract

Expression forecasting methods use machine learning models to predict how ... transcriptome upon perturbation. Such methods are enticing because they pr... pressing questions in fields ranging from developmental genetics to cell fate e... because they are a fast, cheap, and accessible complement to the correspon... However, the absolute and relative accuracy of these methods is poorly chara... their informed use, their improvement, and the interpretation of their predictio... these issues, we created a benchmarking platform that combines a panel of 1... perturbation datasets with an expression forecasting software engine that enc... interfaces to a wide variety of methods. We used our platform to systematical... methods, parameters, and sources of auxiliary data, finding that performance ... on the choice of metric, and especially for simple metrics like mean squared e... uncommon for expression forecasting methods to out-perform simple baselines. Our platform

Benchmarking Transcriptomics Foundation Models for Perturbation Analysis : one PCA still rules them all

Ihab Bendidi, ...ole Normale Supérieure, Paris, France
Shawn Whitfield, Valence Labs, Montreal, Canada
Kian Kenyon-Dean, Recursion, Toronto, Canada
Ben Yedder, ...nce Labs, ...al, Canada
Yassir El Mesbahi, Valence Labs, Montreal, Canada
Emmanuel Noutahi, Valence Labs, Montreal, Canada
Alisandra K. Denton, Valence Labs, Montreal, Canada

Abstract

...nderstanding the relationships among genes, compounds, and their interact... living organisms remains limited due to technological constraints and the c... plexity of biological data. Deep learning has shown promise in exploring t... relationships using various data types. However, transcriptomics, which prov... detailed insights into cellular states, is still underused due to its high noise le...

Machine Learning for Genomics Explorations workshop at ICLR 2024

ENHANCING GENERATIVE PERTURBATION MODELS WITH LLM-INFORMED GENE EMBEDDINGS

Kaspar Märtens, Rory Donovan-Maiye & Jesper Ferkinghoff-Borg
Digital Science & Innovation, Novo Nordisk
{KQTM,RZDM,JFGB}@novonordisk.com

ABSTRACT

Genetic perturbations are key to understanding how genes regulate cell behavior, yet the ability to predict responses to these perturbations remains a significant ...

Benchmarking AI Models for *In Silico* Gene Perturbation of Cells

Chen Li[1,2#], Haoxiang Gao[2#], Yuli She[2#], Haiyang Bian[1,2], Qing Chen[2], Kai Liu[2*], ... and Xuegong Zhang[1,3*]
...ioinformatics Division of BNRIST, Department of ... 4, China
...5, Singapore

PERTEVAL-SCFM: BENCHMARKING SI... FOUNDATION MODELS FOR PERTURBAT... PREDICTION

A. Wenteler[1*], M. Occhetta[1], N. Branson[1], M. Huebner[1], V. Curean[2] W. T. Dee[1], W. T. Connell, A. Hawkins-Hooker[4], S. P. Chung[1], Y. E... A. Gallagher-Syed[1], C. M. V. Córdova[6,7]
[1]Queen Mary University of London, [2]University of Medicine and Pharmac... [3]STAR-UBB Institute Cluj, [4]University College London, [5]Harvard Univers...

SHORT REPORT — Open Access

Zero-shot evaluation reveals limitations of single-cell foundation models

Kasia Z. Kedzierska[1], Lorin Crawford[2], Ava P. Amini[2] and Alex X. Lu[2*]

*Correspondence: lualex@microsoft.com
[1] University of Oxford, Oxford, UK
[2] Microsoft Research, Cambridge, MA, USA

Abstract

Foundation models such as scGPT and Geneformer have not been rigorously evaluated in a setting where they are used without any further training (i.e., zero-shot). Understanding the performance of models in zero-shot settings is critical to applications that exclude the ability to fine-tune, such as discovery settings where labels are unknown. Our evaluation of the zero-shot performance of Geneformer and scGPT suggests that, in some cases, these models may face reliability challenges and could be outperformed by simpler methods. Our findings underscore the importance of zero-shot evaluations in development and deployment of foundation models in single-cell research.

Benchmarking a foundational ... for post-perturbation RNAseq prediction

Gerold Csendes[1], Kristóf Z. Szalay[1], Bence Szalai[1,*]
[1] Turbine Ltd., Budapest, Hungary
* correspondence: bence.szalai@turbine.ai

Abstract

Accurately predicting cellular responses to perturbations is essential for understanding cell behaviour in both healthy and diseased states. While perturbation data is ideal for building such predictive models, it is considerably sparser than baseline (non-perturbed) cellular data. To address this limitation, several foundational cell models have been developed using large-scale single-cell gene expression data. These models are fine-tuned after pre-training for specific tasks, such as predicting post-perturbation gene expression profiles, and are considered state-of-the-art for these problems. However, proper benchmarking of these models remains an unsolved challenge.

...arking Machine Learning ...r Perturbation Analysis

Yan Wu*, Altos Labs, San Diego, US
Esther Wershof*, Altos Labs, Cambridge, UK
Sebastian M Schmon*, Altos Labs, Cambridge, UK
Marcel Nassar*, Altos Labs, San Diego, US
Błażej Osiński*, Altos Labs, Cambridge, UK
Ridvan Eksi*, Altos Labs, San Diego, US
Kun Zhang, Altos Labs, San Diego, US
Thore Graepel, Altos Labs, Cambridge, UK

But maybe we are looking at the wrong metrics. Two alternative proposals:

nature biotechnology

View all

Explore content | About the journal | Publish with us

nature > nature biotechnology > articles > article

Article | Open access | Published: 25 August 2025

Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan & Maria Brbić

Nature Biotechnology (2025) | Cite this article

33k Accesses | **3** Citations | **53** Altmetric | Metrics

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | SUB

New Results

Follow this preprint

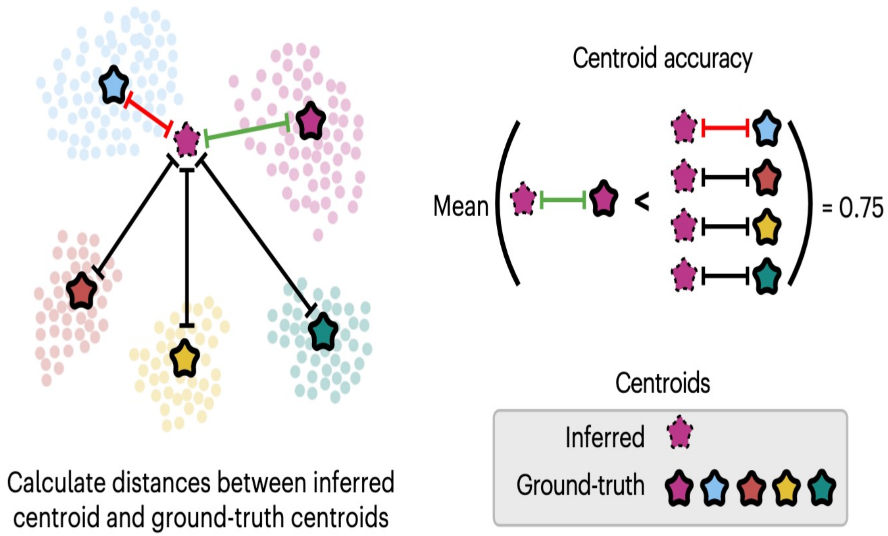
Deep Learning-Based Genetic Perturbation Models *Do* Outperform Uninformative Baselines on Well-Calibrated Metrics

Henry E. Miller, Gabriel M. Mejia, Francis J.A. Leblanc, Bo Wang, Brendan Swain, Lucas Paulo de Lima Camillo

doi: https://doi.org/10.1101/2025.10.20.683304

This article is a preprint and has not been certified by peer review [what does this mean?].

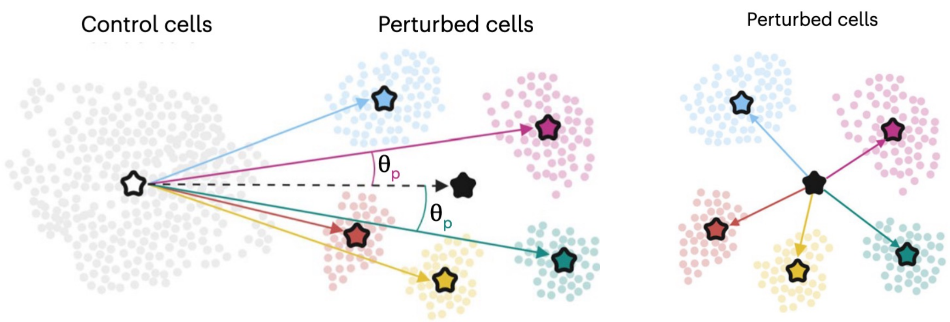
Systema: how good is the prediction relative to the other perturbations?



Viñas Torné et al., Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. Nature Biotechnology

Systematic differences between perturbations and control inflate the Pearson Delta score

Using perturbation mean as reference for Pearson Delta

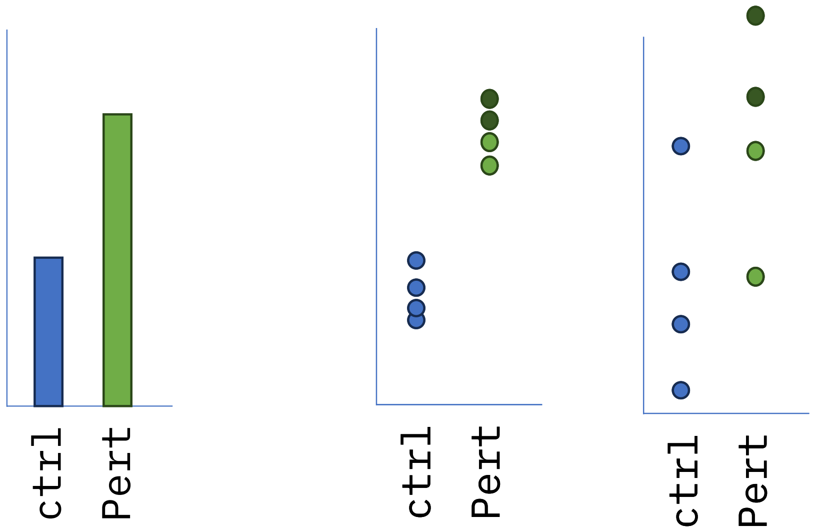


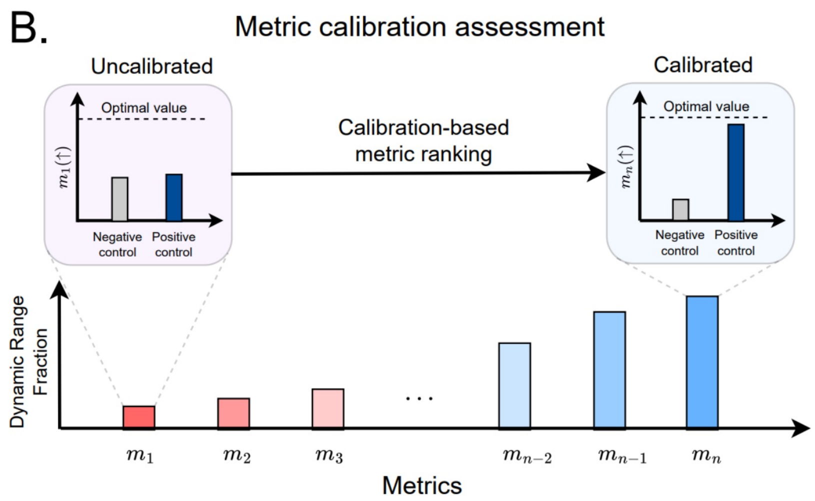

```python
X_true = pert_adata.layers['obs'][condition,:]
X_pred = pert_adata.layers['pred'][condition,:]

pert_mean = pert_adata.X.mean(axis=0)
ctrl_mean = ctrl_adata.X.mean(axis=0)

# Pearson Correlation
pearsonr(X_true, X_pred)
# Pearson Delta wrt. to control
pearsonr(X_true - ctrl_mean, X_pred - ctrl_mean)
# Pearson Delta wrt. to pertubation mean
pearsonr(X_true - pert_mean, X_pred - pert_mean)
```

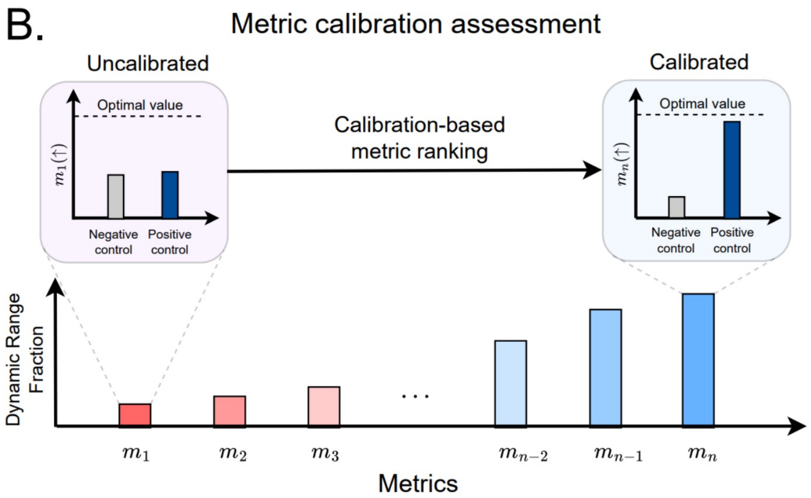
Miller et al. propose to use metrics that separate positive and negative control





Miller et al., Deep Learning-Based Genetic Perturbation Models Do Outperform Uninformative Baselines on Well-Calibrated Metrics. bioRxiv

Miller et al. propose to use metrics that separate positive and negative control



→ MSE weighted by DE is “well calibrated”

Miller et al., Deep Learning-Based Genetic Perturbation Models Do Outperform Uninformative Baselines on Well-Calibrated Metrics. bioRxiv

What to measure?

- Mean squared error
- Pearson correlation
- Pearson Delta correlation
  - Delta wrt. to control
  - Delta wrt. to pert. mean
- Centroid accuracy
- Recall of truly non-additive genes

$\times$

Error calculated across

- All genes
- Highly expressed genes
- Differentially expressed genes

Measure	Pro	Con
Mean squared error	Interpretable	Sensitive to outliers
Pearson correlation	Interpretable	Typically very close to 1
Pearson Delta wrt. to control	Interpretable	Systematic effects increase mean predictor performance
Pearson Delta wrt. to perturbation mean	Robust to systematic changes	Less interpretable. Unclear what is a good baseline
Centroid accuracy	Interpretable	Output depends on perturbation similarity
Recall of truly non-additive genes	Interpretable, relevant	Only meaningful for double perturbations

Gene subset	Pro	Con
All	Comprehensive	Noise can dominate signal
Highly Expressed	Informative	Not always relvant
Most differentially expressed	Emphasizes affected genes	Lacks all negative examples

Measure	Pro	Con
Mean squared error	Interpretable	Sensitive to outliers
Pearson correlation	Interpretable	Typically very close to 1
Pearson Delta wrt. to control	Interpretable	Systematic effects increase mean predictor performance
Pearson Delta wrt. to perturbation mean	Robust to systematic changes	**Less interpretable. Unclear what is a good baseline**
Centroid accuracy	Interpretable	Output depends on perturbation similarity
Recall of truly non-additive genes	Interpretable, relevant	Only meaningful for double perturbations

Gene subset	Pro	Con
All	Comprehensive	Noise can dominate signal
Highly Expressed	Informative	Not always relvant
Most differentially expressed	Emphasizes affected genes	Lacks all negative examples

Measure	Pro	Con
Mean squared error	Interpretable	Sensitive to outliers
Pearson correlation	Interpretable	Typically very close to 1
Pearson Delta wrt. to control	Interpretable	Systematic effects increase mean predictor performance
Pearson Delta wrt. to perturbation mean	Robust to systematic changes	Less interpretable. Unclear what is a good baseline
Centroid accuracy	Interpretable	Output depends on perturbation similarity
Recall of truly non-additive genes	Interpretable, relevant	Only meaningful for double perturbations

Gene subset	Pro	Con
All	Comprehensive	Noise can dominate signal
Highly Expressed	Informative	Not always relvant
Most differentially expressed	Emphasizes affected genes	**Lacks all negative examples**

Thinking inside vs outside the box

Inside:

Finding the best metric to train a Deep Learning model for Perturb-Seq data

- MSE on highly expressed genes

Outside:

Predict

- Proliferation
- T cell exhaustion
- Contractile strength
- Cell-cell interactions
- ...

learned from spatiotemporal data

Thank you for your attention

- **Wolfgang Huber**
- **Simon Anders**
- **Constantin Ahlmann-Eltze**

https://www.nature.com/articles/s41592-025-02772-6

EMBL

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

UCL

erc
European Research Council

BioQuant
MODEL base of LIFE