MCBRLab
Cesar Valdez

MCBRLab
Cesar Valdez

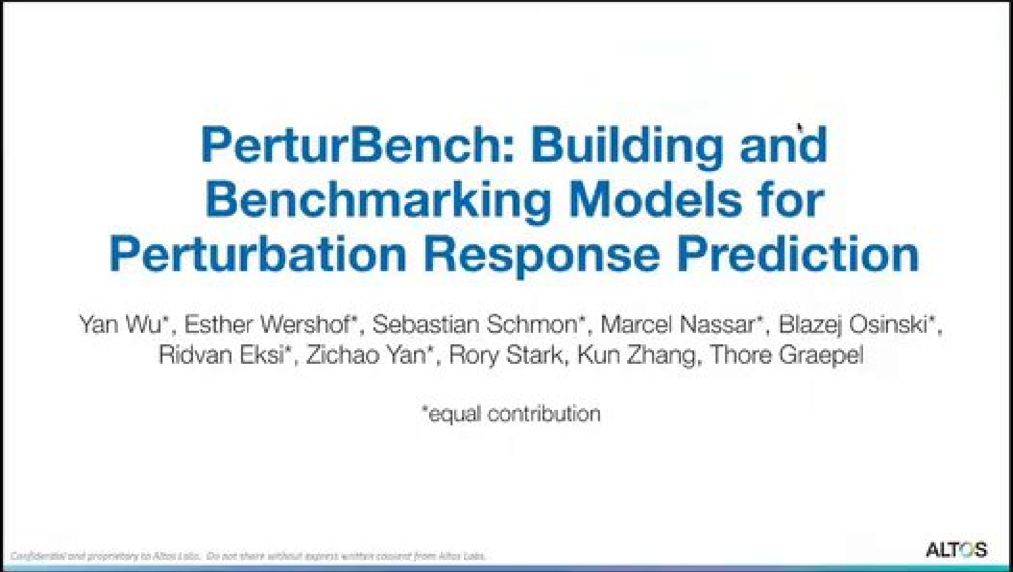
MCBRLab bilibili
Cesar Valdez

MCBRLab
Cesar Valdez

Cesar Valdez

MCBRLab bilibili
Yan Wu

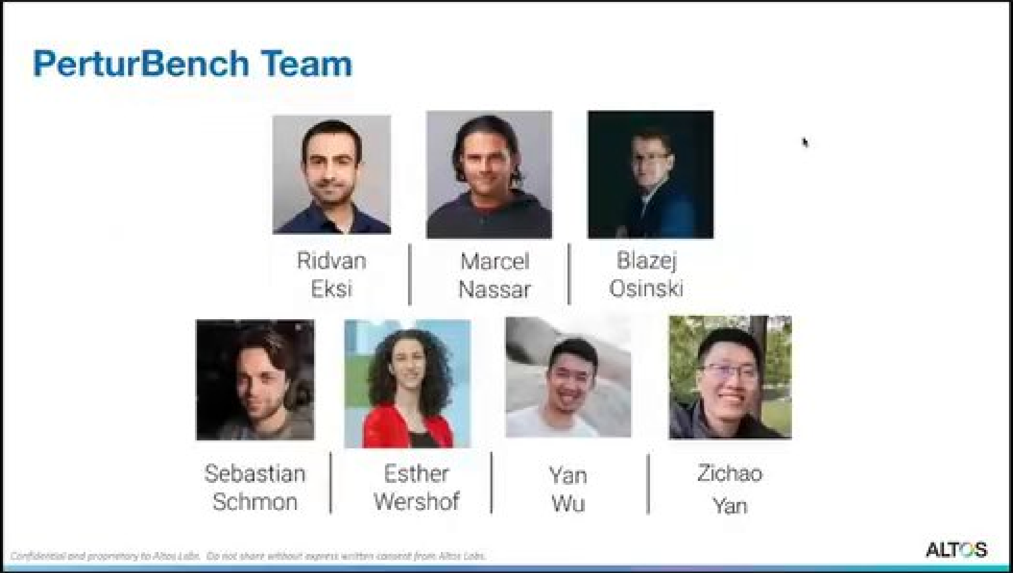
Loading...



PerturBench: Building and Benchmarking Models for Perturbation Response Prediction

Yan Wu*, Esther Wershof*, Sebastian Schmon*, Marcel Nassar*, Blazej Osinski*, Ridvan Eksi*, Zichao Yan*, Rory Stark, Kun Zhang, Thore Graepel

*equal contribution

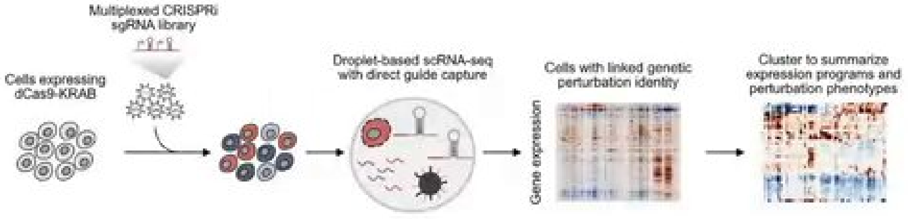


PerturBench Team

Ridvan Eksi | Marcel Nassar | Blazej Osinski

Sebastian Schmon | Esther Wershof | Yan Wu | Zichao Yan

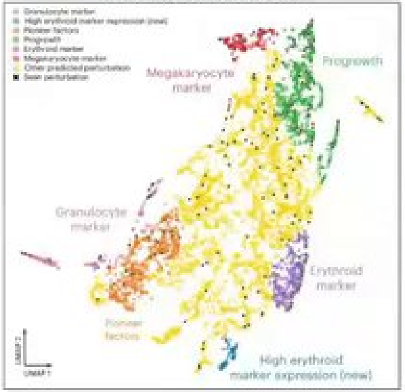
Advances in genomics enabled high-throughput screening of cell state response to perturbations


Multiplexed CRISPRi sgRNA library
Cells expressing dCas9-KRAB
Droplet-based scRNA-seq with direct guide capture
Cells with linked genetic perturbation identity
Gene expression
Cluster to summarize expression programs and perturbation phenotypes

ML predictions enable comprehensive mapping of perturbation response space

- Exhaustively measuring perturbation effects across cell types/states cost prohibitive
  - Impossible when dealing with combinations, multiple cell states
  - Cell states key for modeling diseases
- Virtual cell models to predict perturbation effects and map out response space
- Example: GEARS predicts ~5k pairs of perturbations when trained on ~120 observed pairs
- Focus wet lab experiments on most interesting perturbations (i.e. disease modifying)

Postperturbation gene expression:
5,460 pairwise combinations (predictions)



Roohani et al, Nat Biotech. 2023

Related work and remaining gaps

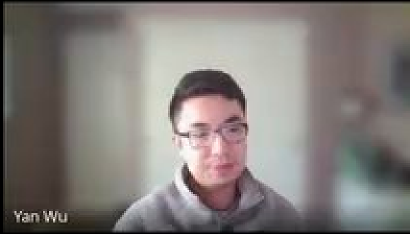
Szalata et al., Advances in Neural Information Processing Systems (NeurIPS), 2024
Ahlmann-Eltze et al., bioRxiv, 2024
Wenteler et al., bioRxiv, 2024
Csendes et al., BMC Genomics, 2025
Wong et al., bioRxiv, 2025
Kernfeld et al., bioRxiv, 2023
Li et al., bioRxiv, 2024
Li et al., bioRxiv, 2024
Roohani et al., Cell, 2025

Related benchmarks

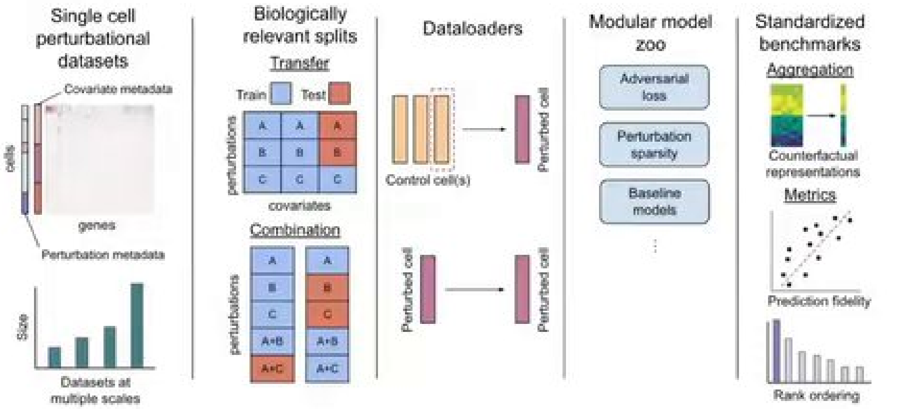
- 2023 NeurIPS perturbation prediction competition [1]: novel drug perturbation dataset in primary blood cells
- Ahlmann-Eltze et al [2], Wentler et al [3], Csendes et al [4], Wong et al [5]: fine-tuned single cell foundation models for perturbation response prediction
- Kernfield et al [6]: unseen perturbation prediction using regulatory networks
- Li et al [7] and Li et al [8]: comprehensive benchmarks across diverse datasets
- 2025 ARC Virtual Cell Competition [9]: novel genetic perturbation dataset in stem cells

Remaining gaps

- Biologically relevant metrics - ranking and distributions
- Model and data ablation experiments enabled by unified software framework



PerturBench Overview



Single cell perturbational datasets

Biologically relevant splits

Dataloaders

Modular model zoo

Adversarial loss

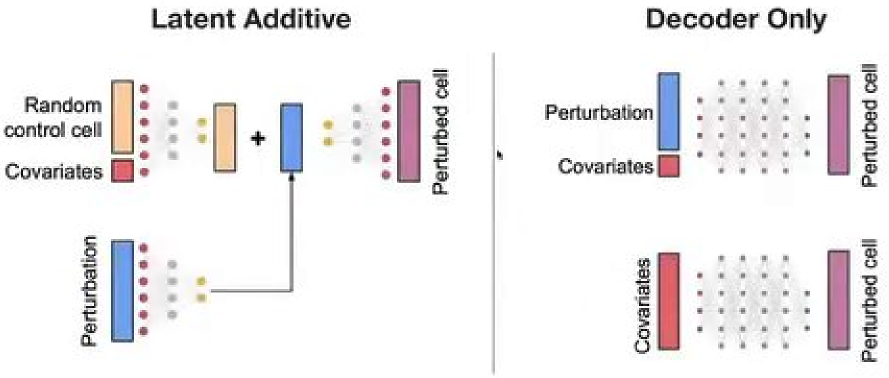
Perturbation sparsity

Baseline models

Standardized benchmarks

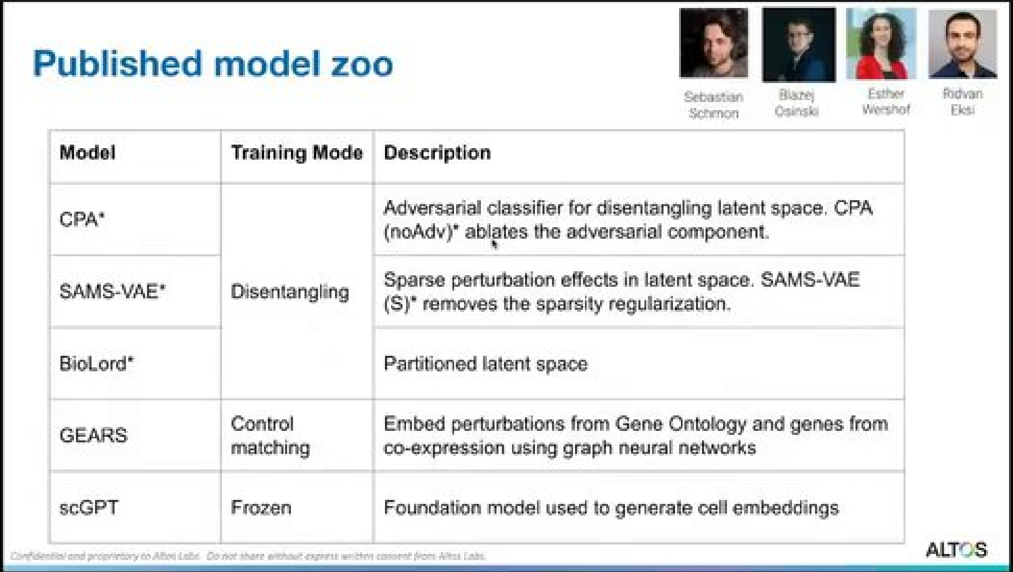
Confidential and proprietary to Altos Labs. Do not share without express written consent from Altos Labs.

Baseline models



Latent Additive

Decoder Only



Published model zoo

Sebastian Schmon

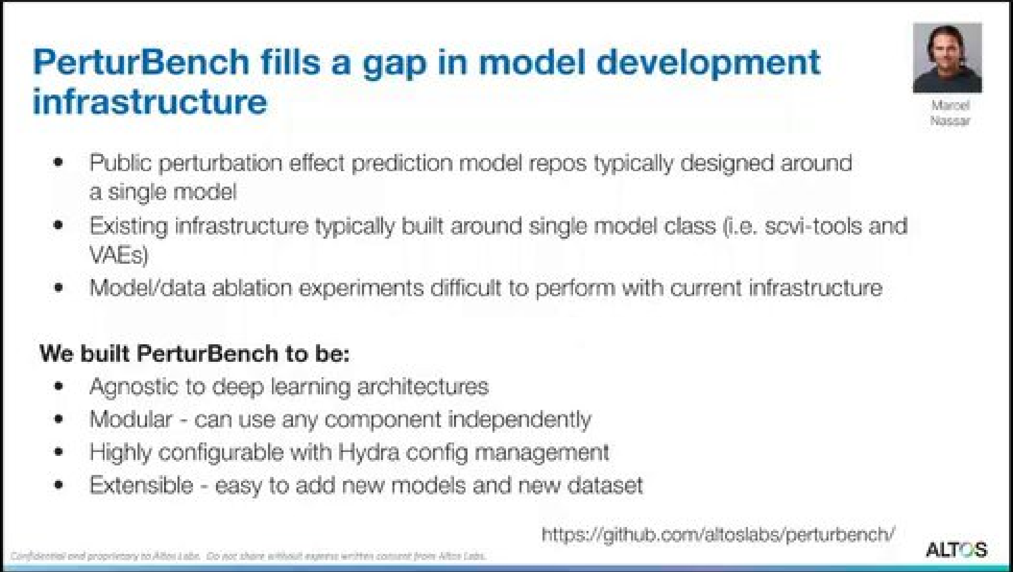
Blazej Osinski

Esther Wershof

Ridvan Eksi

Model	Training Mode	Description
CPA*	Disentangling	Adversarial classifier for disentangling latent space. CPA (noAdv)* ablates the adversarial component.
SAMS-VAE*		Sparse perturbation effects in latent space. SAMS-VAE (S)* removes the sparsity regularization.
BioLord*		Partitioned latent space
GEARS	Control matching	Embed perturbations from Gene Ontology and genes from co-expression using graph neural networks
scGPT	Frozen	Foundation model used to generate cell embeddings

ALTOS



PerturBench fills a gap in model development infrastructure

Marcel Nassar

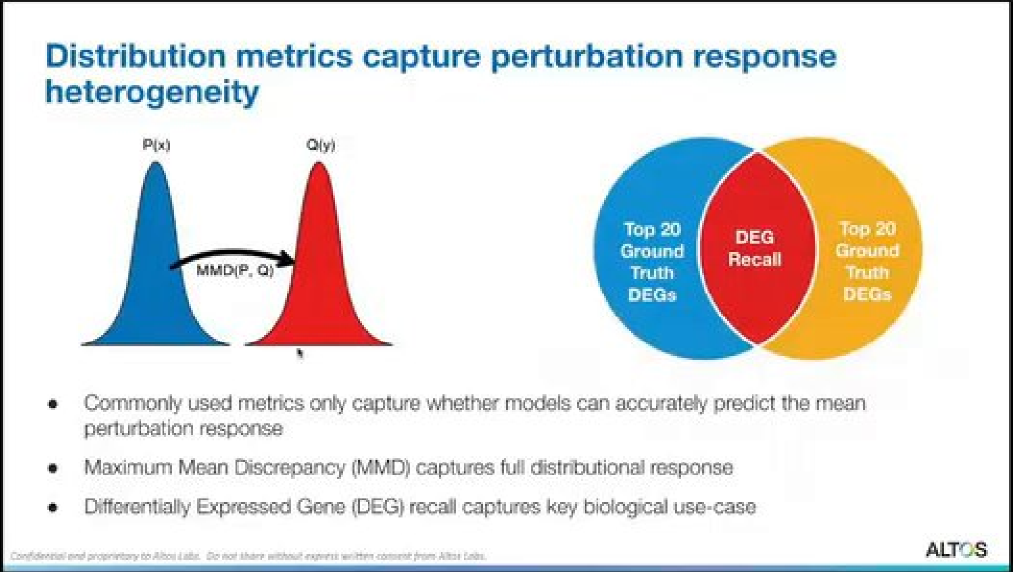
- Public perturbation effect prediction model repos typically designed around a single model
- Existing infrastructure typically built around single model class (i.e. scvi-tools and VAEs)
- Model/data ablation experiments difficult to perform with current infrastructure

We built PerturBench to be:

- Agnostic to deep learning architectures
- Modular - can use any component independently
- Highly configurable with Hydra config management
- Extensible - easy to add new models and new dataset

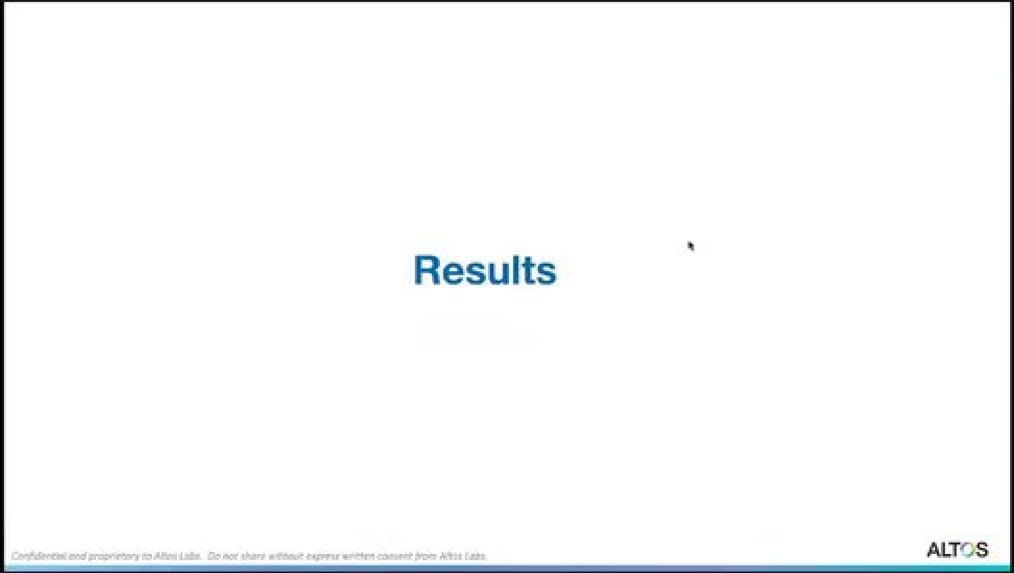
https://github.com/altoslabs/perturbench/

ALTOS

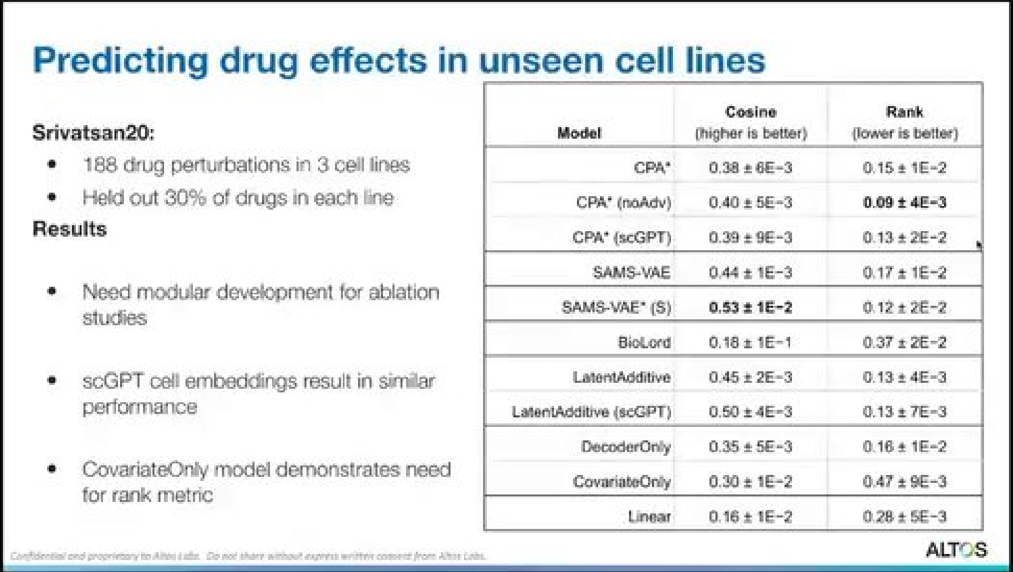


Distribution metrics capture perturbation response heterogeneity

- Commonly used metrics only capture whether models can accurately predict the mean perturbation response
- Maximum Mean Discrepancy (MMD) captures full distributional response
- Differentially Expressed Gene (DEG) recall captures key biological use-case



Results



Predicting drug effects in unseen cell lines

Srivatsan20:

- 188 drug perturbations in 3 cell lines
- Held out 30% of drugs in each line

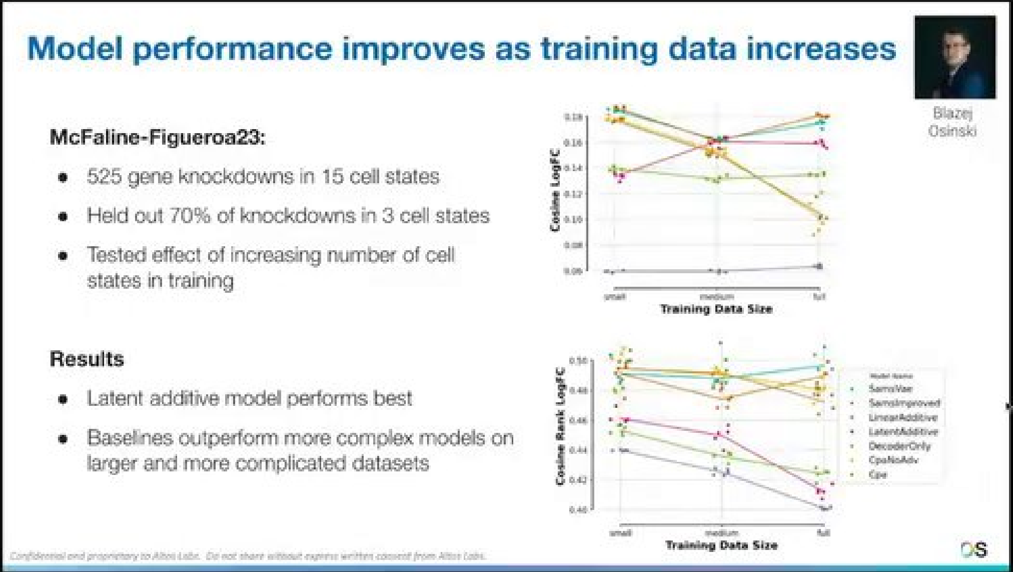
Results

- Need modular development for ablation studies
- scGPT cell embeddings result in similar performance
- CovariateOnly model demonstrates need for rank metric

Model	Cosine (higher is better)	Rank (lower is better)
CPA*	0.38 ± 6E-3	0.15 ± 1E-2
CPA* (noAdv)	0.40 ± 5E-3	**0.09 ± 4E-3**
CPA* (scGPT)	0.39 ± 9E-3	0.13 ± 2E-2
SAMS-VAE	0.44 ± 1E-3	0.17 ± 1E-2
SAMS-VAE* (S)	**0.53 ± 1E-2**	0.12 ± 2E-2
BioLord	0.18 ± 1E-1	0.37 ± 2E-2
LatentAdditive	0.45 ± 2E-3	0.13 ± 4E-3
LatentAdditive (scGPT)	0.50 ± 4E-3	0.13 ± 7E-3
DecoderOnly	0.35 ± 5E-3	0.16 ± 1E-2
CovariateOnly	0.30 ± 1E-2	0.47 ± 9E-3
Linear	0.16 ± 1E-2	0.28 ± 5E-3

Confidential and proprietary to Altos Labs. Do not share without express written consent from Altos Labs.

ALTOS



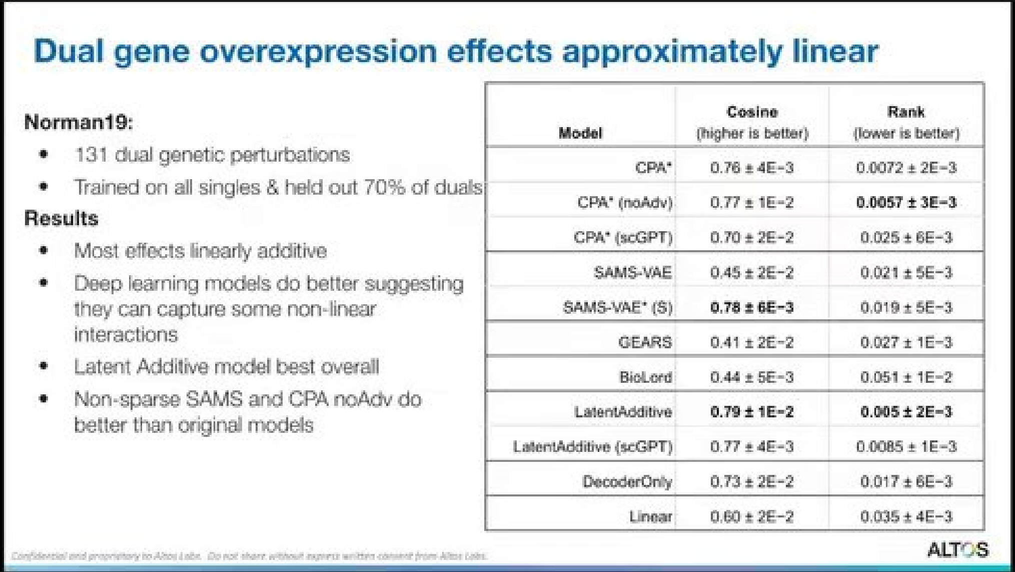
Model performance improves as training data increases

McFaline-Figueroa23:

- 525 gene knockdowns in 15 cell states
- Held out 70% of knockdowns in 3 cell states
- Tested effect of increasing number of cell states in training

Results

- Latent additive model performs best
- Baselines outperform more complex models on larger and more complicated datasets



Dual gene overexpression effects approximately linear

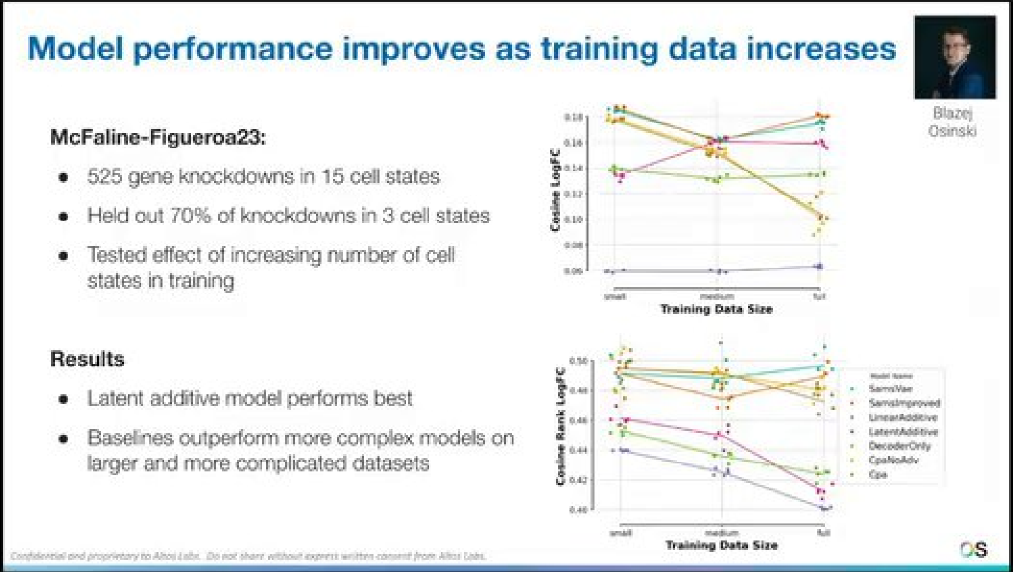
Norman19:

- 131 dual genetic perturbations
- Trained on all singles & held out 70% of duals

Results

- Most effects linearly additive
- Deep learning models do better suggesting they can capture some non-linear interactions
- Latent Additive model best overall
- Non-sparse SAMS and CPA noAdv do better than original models

Model	Cosine (higher is better)	Rank (lower is better)
CPA*	0.76 ± 4E-3	0.0072 ± 2E-3
CPA* (noAdv)	0.77 ± 1E-2	**0.0057 ± 3E-3**
CPA* (scGPT)	0.70 ± 2E-2	0.025 ± 6E-3
SAMS-VAE	0.45 ± 2E-2	0.021 ± 5E-3
SAMS-VAE* (S)	**0.78 ± 6E-3**	0.019 ± 5E-3
GEARS	0.41 ± 2E-2	0.027 ± 1E-3
BioLord	0.44 ± 5E-3	0.051 ± 1E-2
LatentAdditive	**0.79 ± 1E-2**	**0.005 ± 2E-3**
LatentAdditive (scGPT)	0.77 ± 4E-3	0.0085 ± 1E-3
DecoderOnly	0.73 ± 2E-2	0.017 ± 6E-3
Linear	0.60 ± 2E-2	0.035 ± 4E-3



Model performance improves as training data increases

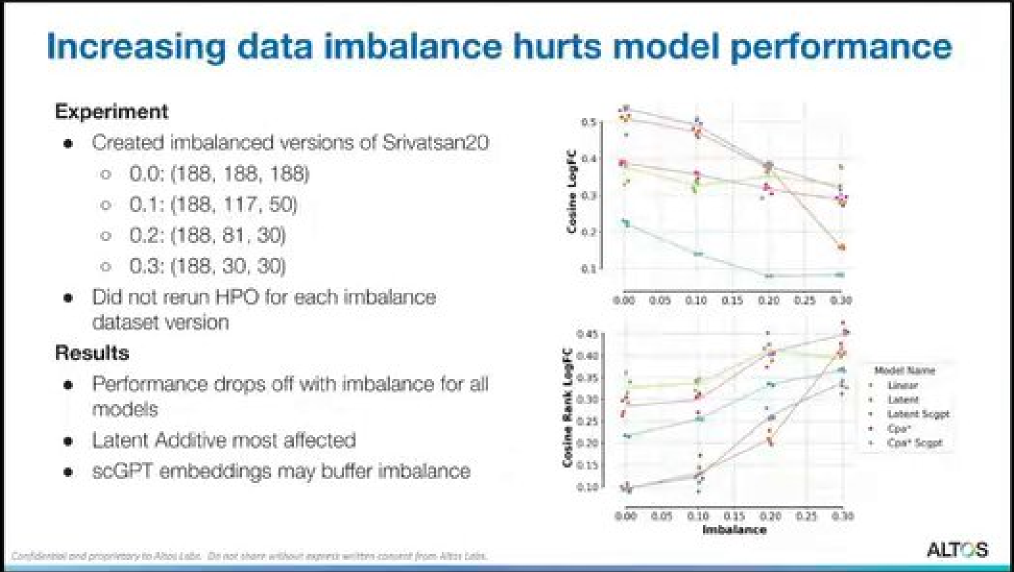
Blazej Osinski

McFaline-Figueroa23:

- 525 gene knockdowns in 15 cell states
- Held out 70% of knockdowns in 3 cell states
- Tested effect of increasing number of cell states in training

Results

- Latent additive model performs best
- Baselines outperform more complex models on larger and more complicated datasets



Increasing data imbalance hurts model performance

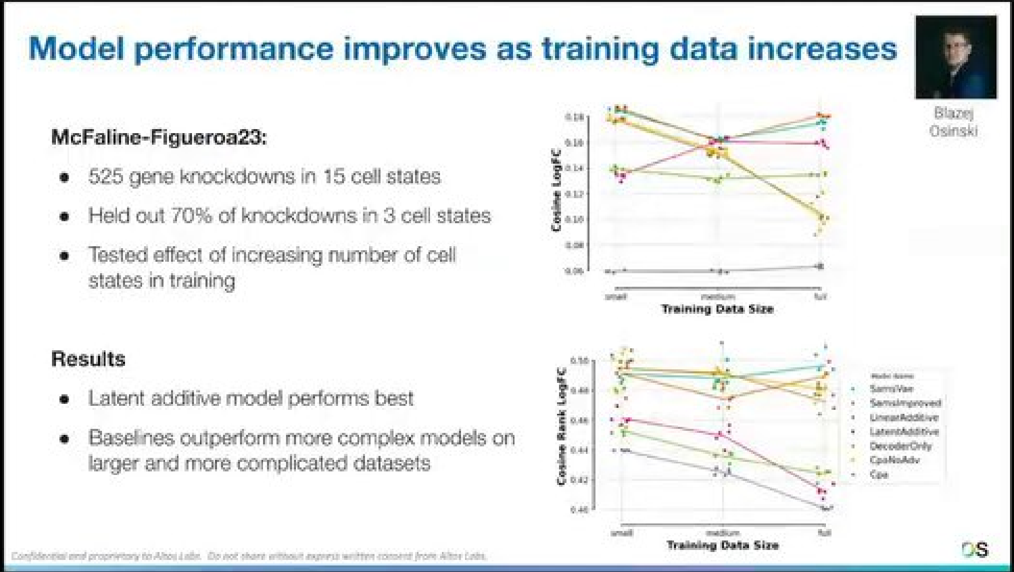
Experiment

- Created imbalanced versions of Srivatsan20
  - 0.0: (188, 188, 188)
  - 0.1: (188, 117, 50)
  - 0.2: (188, 81, 30)
  - 0.3: (188, 30, 30)
- Did not rerun HPO for each imbalance dataset version

Results

- Performance drops off with imbalance for all models
- Latent Additive most affected
- scGPT embeddings may buffer imbalance

ALTOS



Model performance improves as training data increases

McFaline-Figueroa23:

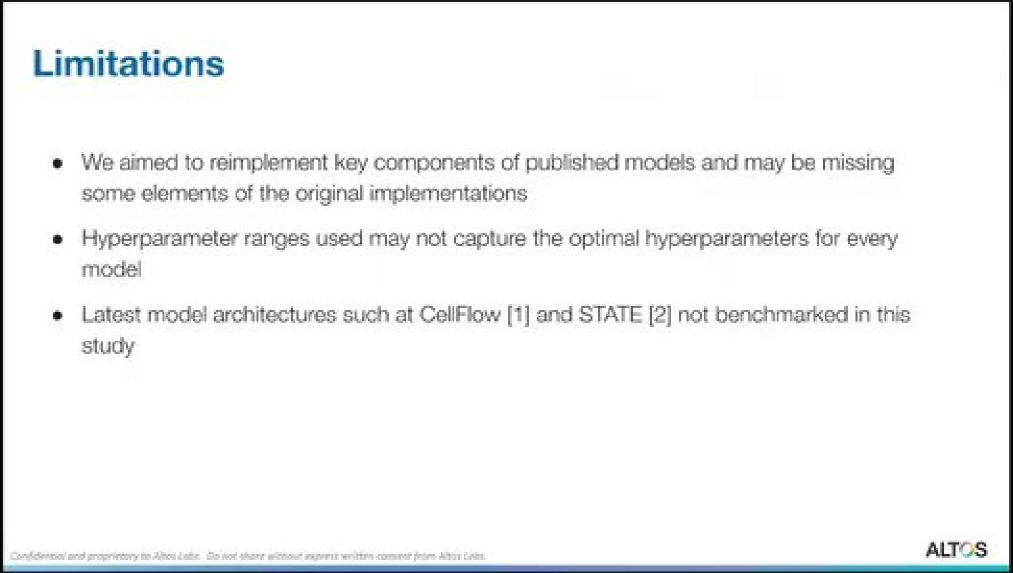
- 525 gene knockdowns in 15 cell states
- Held out 70% of knockdowns in 3 cell states
- Tested effect of increasing number of cell states in training

Results

- Latent additive model performs best
- Baselines outperform more complex models on larger and more complicated datasets

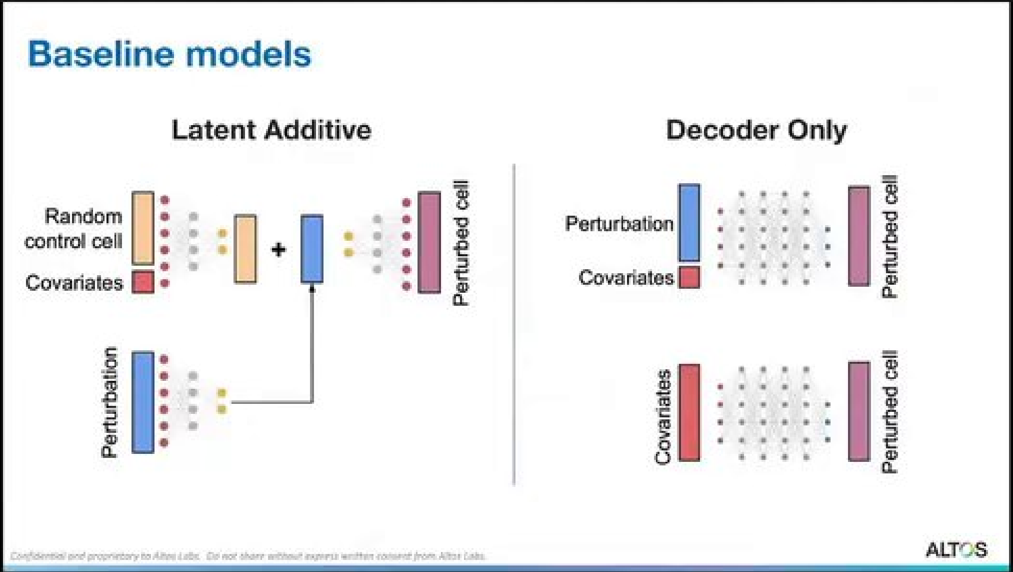
Blazej Osinski

Confidential and proprietary to Altos Labs. Do not share without express written consent from Altos Labs.



Limitations

- We aimed to reimplement key components of published models and may be missing some elements of the original implementations
- Hyperparameter ranges used may not capture the optimal hyperparameters for every model
- Latest model architectures such at CellFlow [1] and STATE [2] not benchmarked in this study



Baseline models

Latent Additive

Random control cell

Covariates

Perturbation

+

Perturbed cell

Decoder Only

Perturbation

Covariates

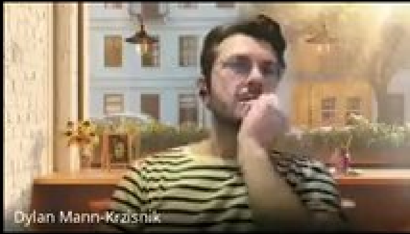
Perturbed cell

Covariates

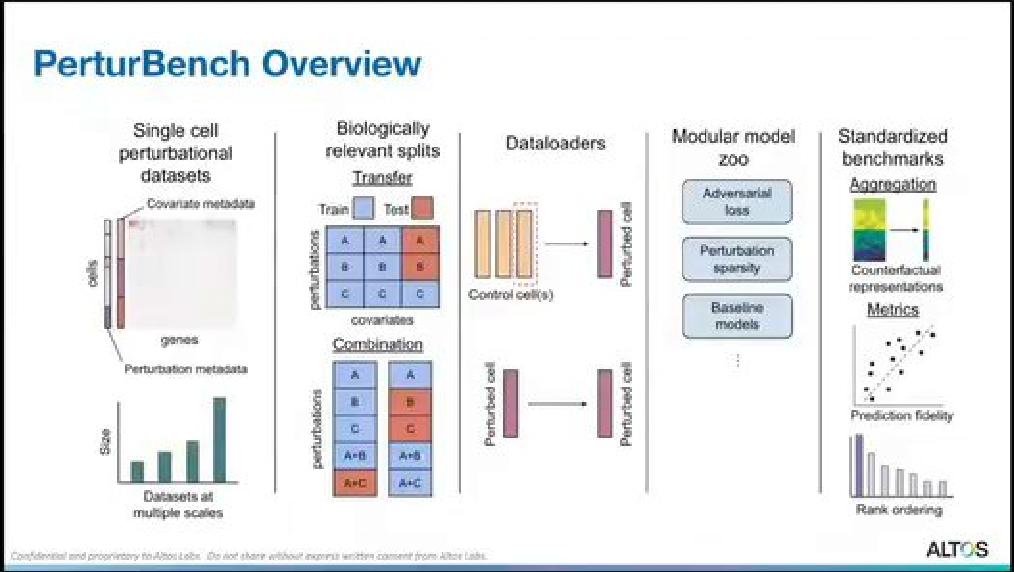
Perturbed cell

Confidential and proprietary to Altos Labs. Do not share without express written consent from Altos Labs.

ALTOS



Dylan Mann-Krzisnik



PerturBench Overview

Single cell perturbational datasets

Biologically relevant splits

Transfer

A	A	A
B	B	B
C	C	C

Combination

A	A
B	B
C	C
A+B	A+B
A+C	A+C

Dataloaders

Modular model zoo

- Adversarial loss
- Perturbation sparsity
- Baseline models

Standardized benchmarks

Aggregation

Metrics