Minimal Libraries and Synthetic Augmentation of CRISPR-Cas9 Screens for Drug Target Discovery

SMPGD 2026

Missing Values and Data Generation

29-30th January 2026

TÉCNICO LISBOA

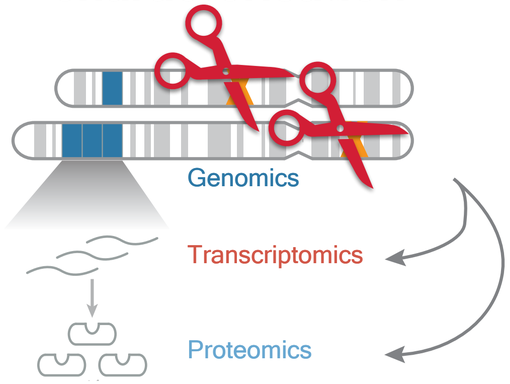
Emanuel Gonçalves

Assistant Professor

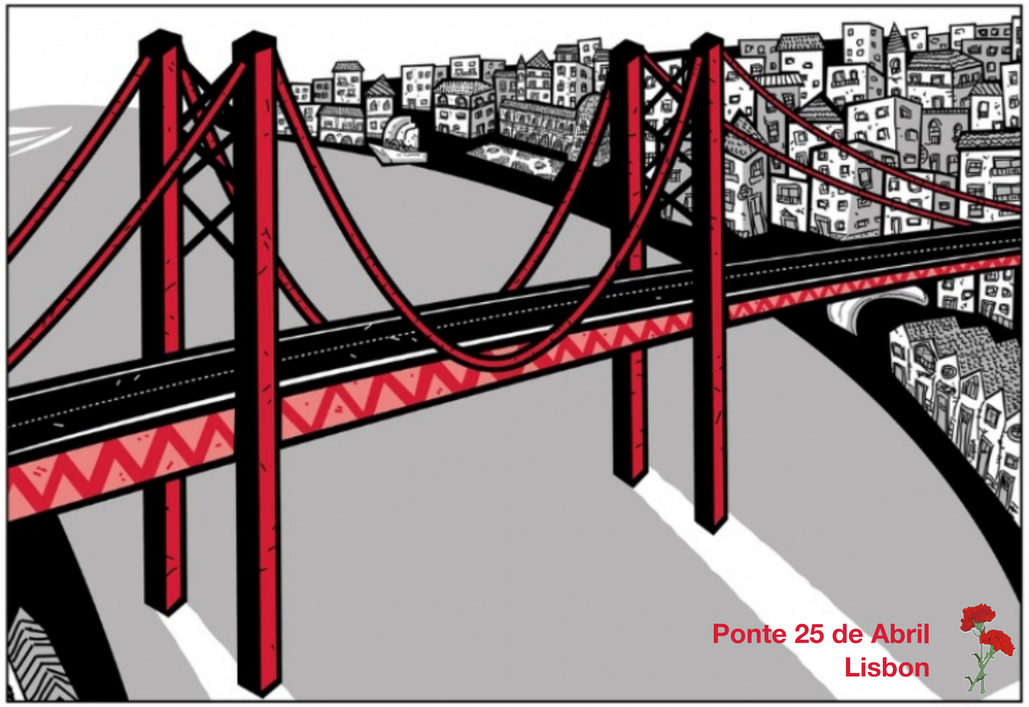
inesc id lisboa

Accelerating precision medicine with data-driven models for higher predictive validity

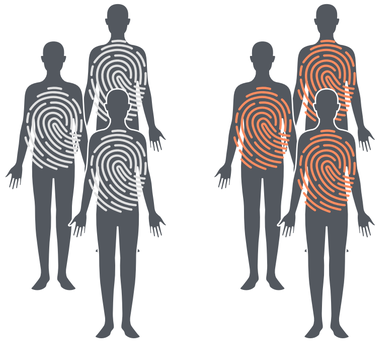
Molecular and Phenotypic characterisation



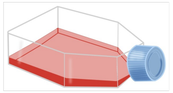
Drug target and Biomarker Discovery (e.g. synthetic lethality)



Stratify

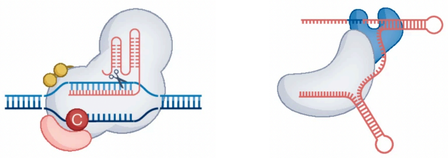


Precision Medicines

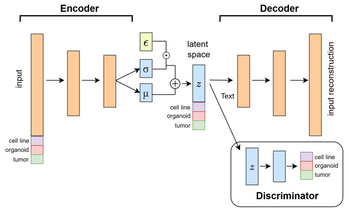


Preclinical

High-throughput Genetic Engineering (e.g. CRISPR-based)

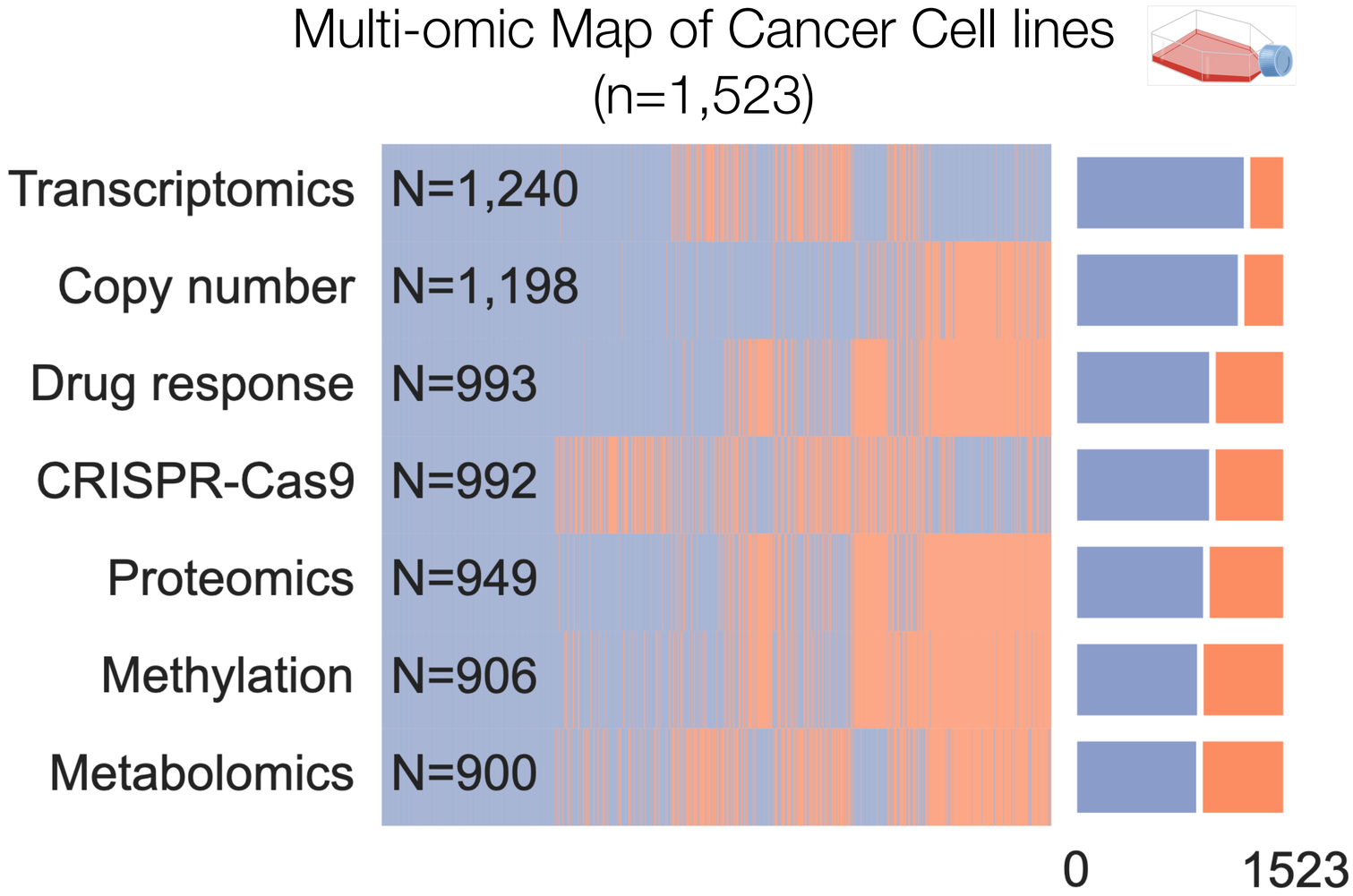


Deep Learning (e.g. Multi-modal)



Clinical

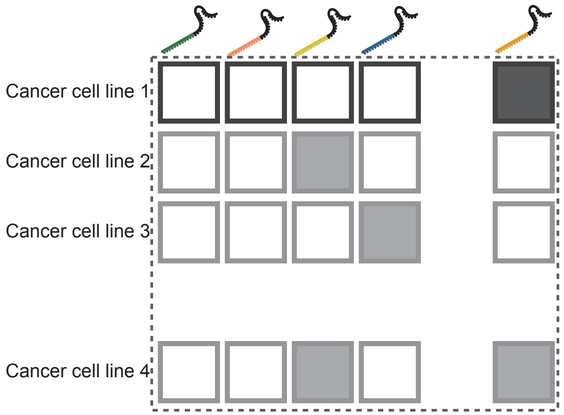
Genomics of Drug Sensitivity in Cancer | Cancer Dependency Map



Garnett et al. *Nature*, 2012.
Iorio et al. *Cell*, 2016.
Behan et al. *Nature*, 2019.
van der Meer et al., *Nucleic Acids Res*, 2019.
Gonçalves et al., *Cancer Cell*, 2022.
Pacini et al, *Cancer Cell*, 2024

Finding genetic biomarkers of essential genes in cancer

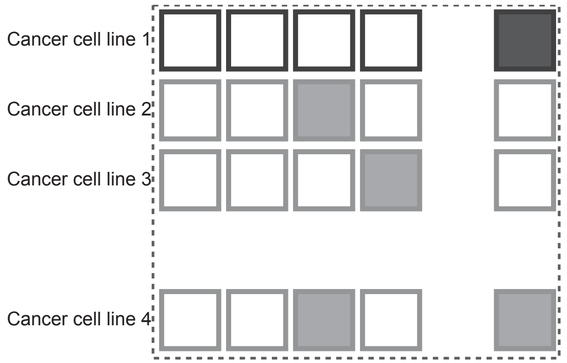
Gene Essentiality
(CRISPR-Cas9)



Continuous
(log2 fold-change)

$$Y = f(X)$$

Genetic Alterations
(e.g. mutations)



Binary
(1=present; 0=absent)

=>

```
y = f(x) Effect size;P-value;FDR;…
y = f(x) Effect size;P-value;FDR;…
y = f(x) Effect size;P-value;FDR;…
y = f(x) Effect size;P-value;FDR;…
y = f(x) Effect size;P-value;FDR;…
y = f(x) Effect size;P-value;FDR;…
…
```

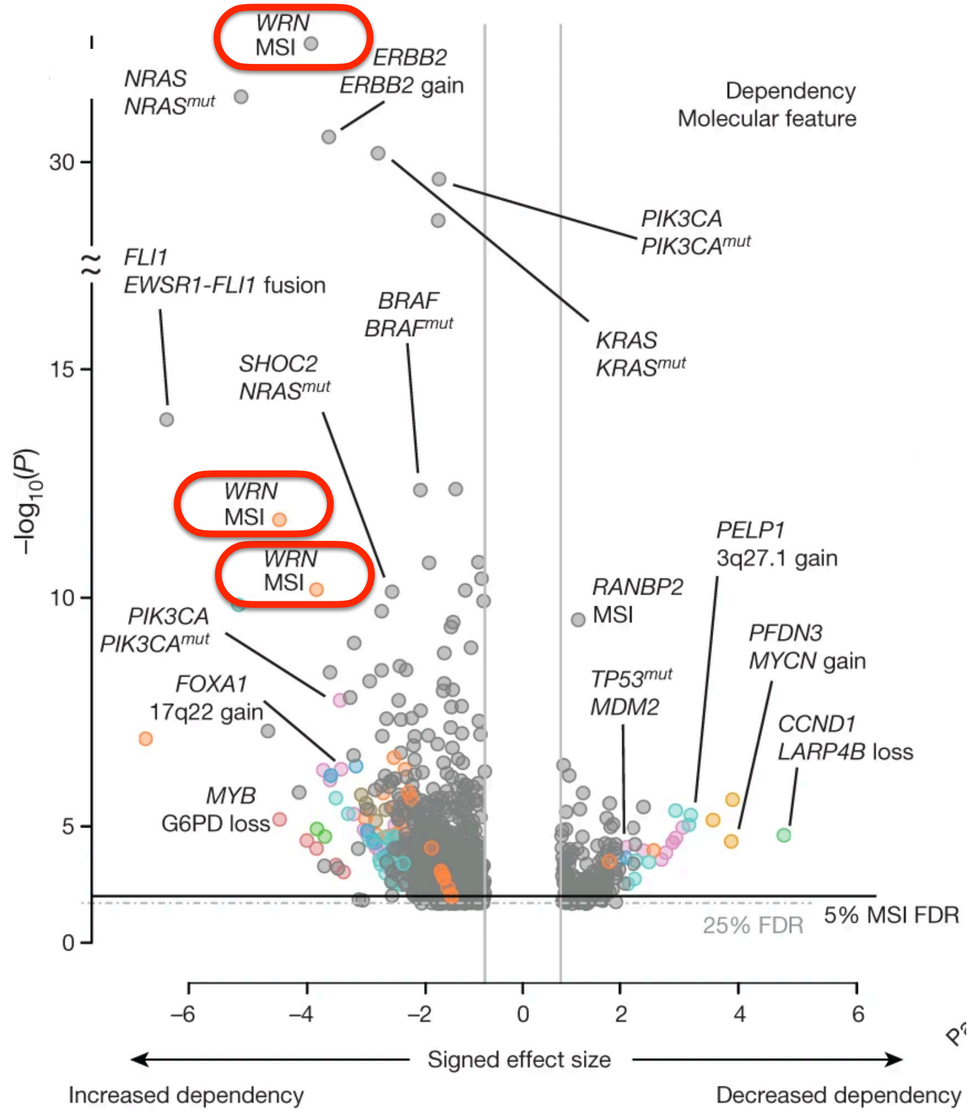
Rank

CRISPR–Cas9 screens in 324 human cancer cell lines from 30 cancer types

Systematic association between gene essentiality and genetic alterations

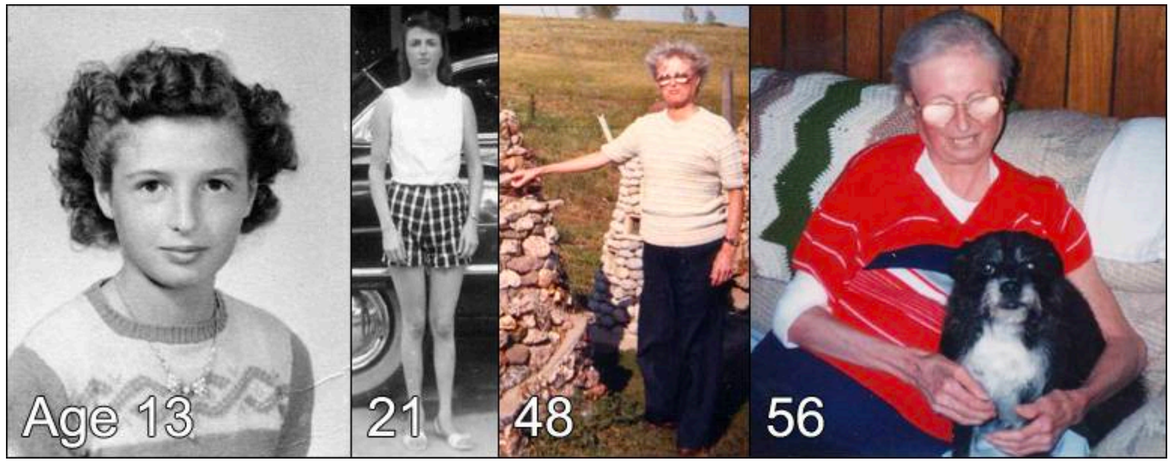
Behan et al. *Nature*, 2019.

Werner Helicase and Werner syndrome



Several cancers with microsatellite instability (MSI) are sensitive Werner Syndrome RecQ Like Helicase (WRN) knockout.

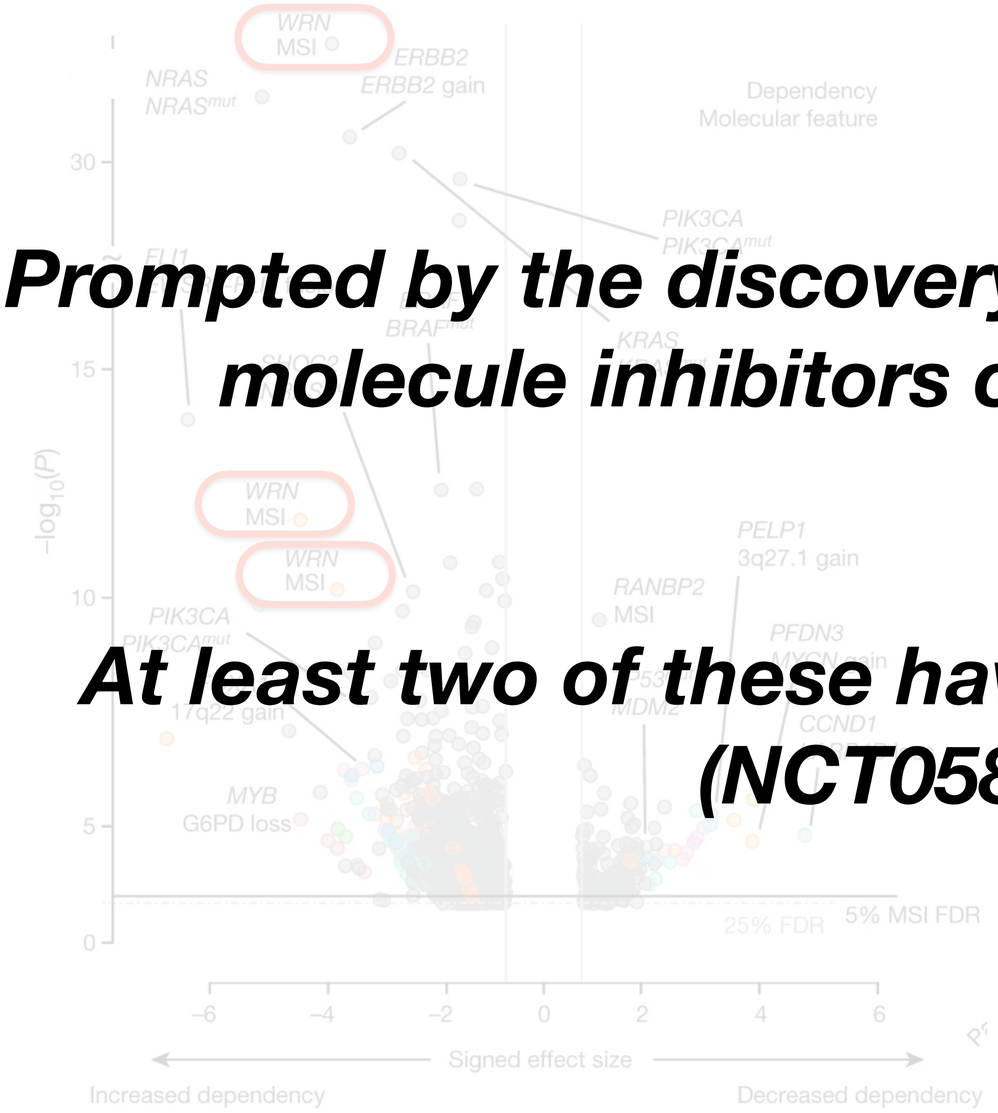
WRN is involved in DNA repair and Werner Syndrome - premature ageing and increased risk of developing cancer



Behan et al. *Nature,* 2019.
Chan et al. *Nature*, 2019.
Lieb et al. *Elife*, 2019.
Kategaya et al. *iScience*, 2019.

https://dlmp.uw.edu/research-center/werner/

Werner Helicase and Werner syndrome

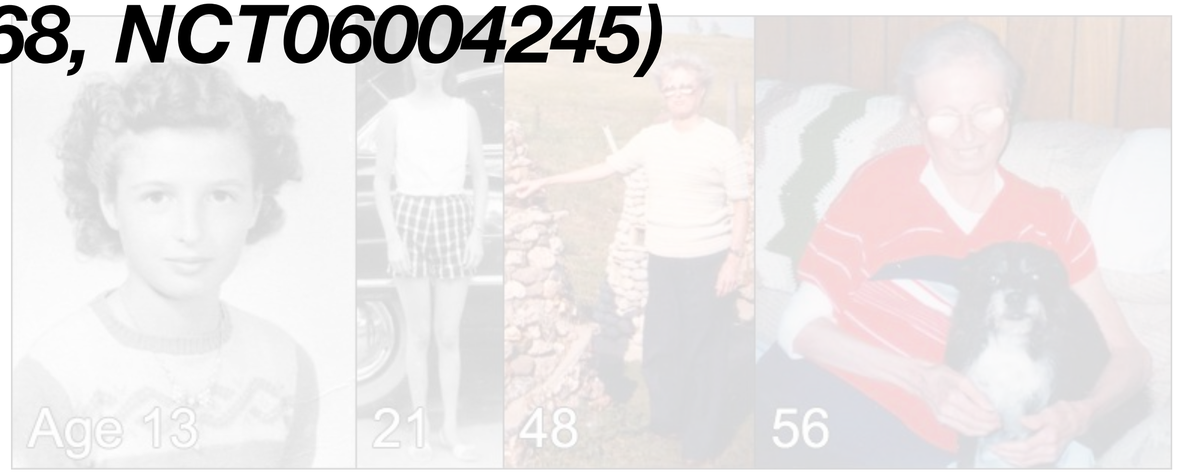


Several cancers with microsatellite instability (MSI) are sensitive Werner Syndrome RecQ Like Helicase (WRN) knockout.

Prompted by the discovery of this synthetic lethality, several small molecule inhibitors of WRN have now been developed*

WRN is involved in DNA repair and Werner Syndrome - premature ageing and increased risk of developing cancer

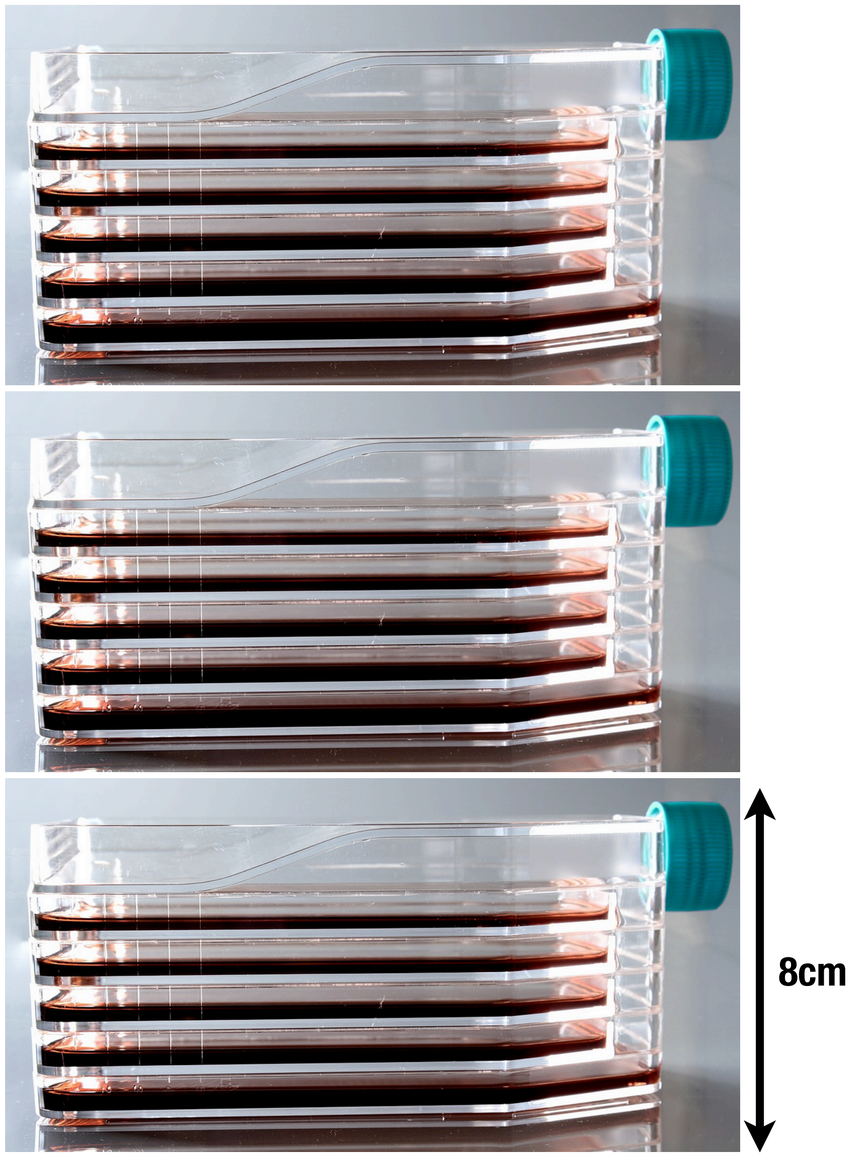
At least two of these have progressed to phase 1 clinical trials (NCT05838768, NCT06004245)



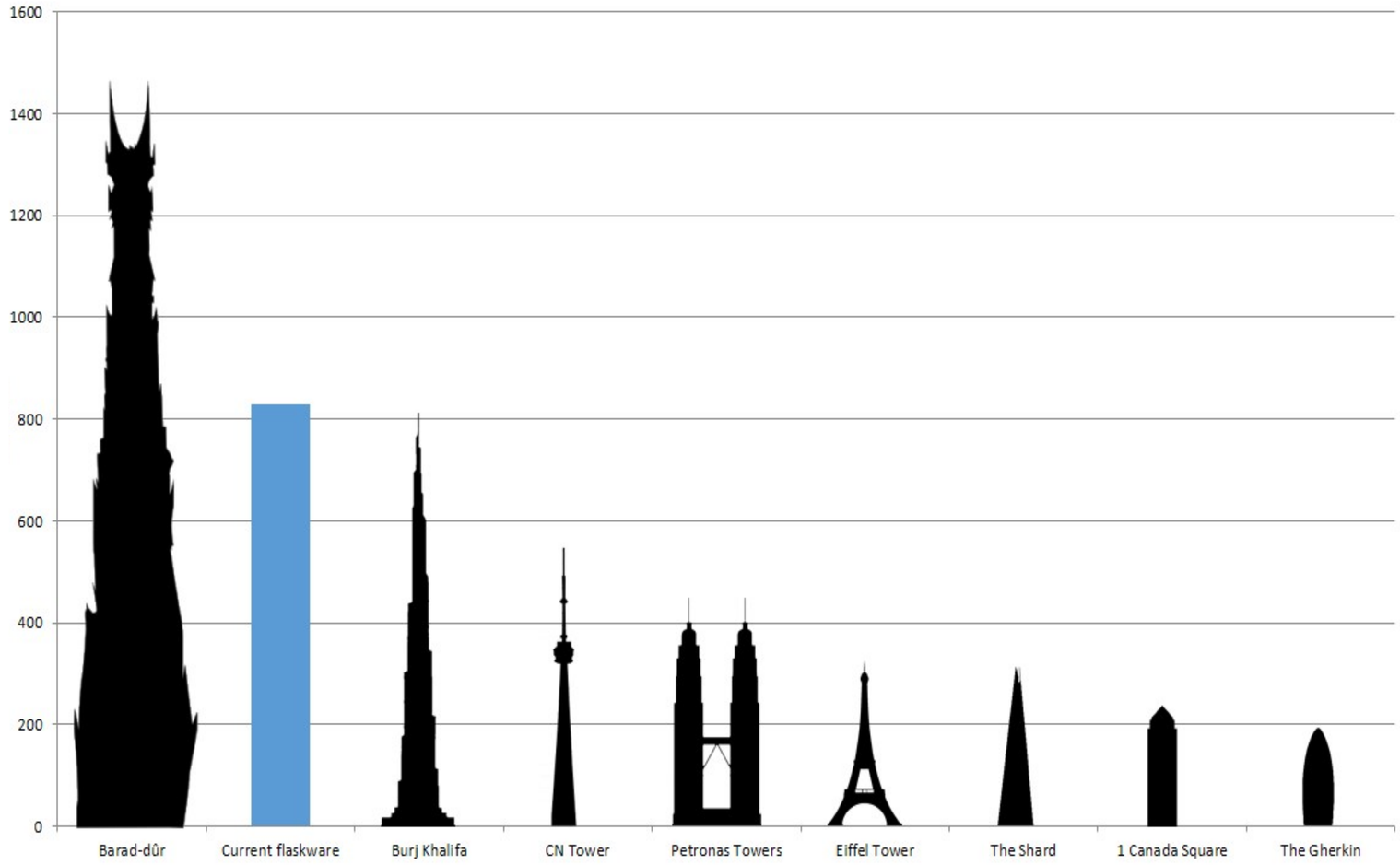
Behan et al. *Nature*, 2019.
Chan et al. *Nature*, 2019.
Lieb et al. *Elife*, 2019.
Kategaya et al. *iScience*, 2019.

*Ferretti et al. *Nature*, 2024. Kikuchi et al. *Cancer Res*, 2024. Picco et al. *Cancer Discov*, 2024.

Gargantuan effort of CRISPR-Cas9 screening at scale



~10-15 five-layer flasks per cell line



Slide courtesy of CGaP and Fiona Behan

Gargantuan effort of CRISPR-Cas9 screening at scale

We need a smaller library!

~10-15 five-layer flasks per cell line

Gargantuan effort of CRISPR-Cas9 screening at scale

We need a smaller library!

Easier said than done…

~10-15 five-layer flasks per cell

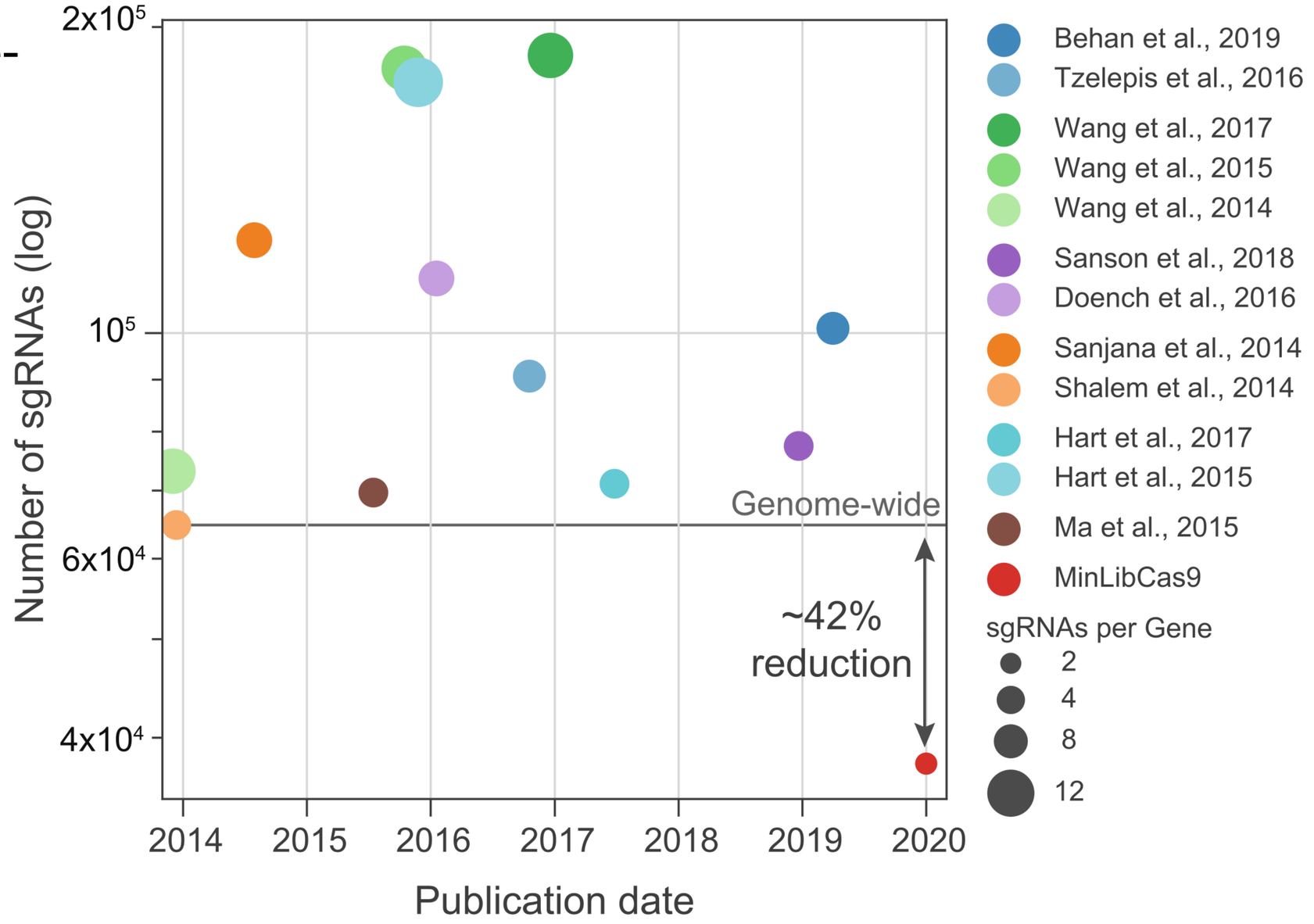
Slide courtesy of CGaP and Fiona Behan

Minimal Genome-wide CRISPR-Cas9 library

Construct a minimal genome-wide CRISPR-Cas9 library

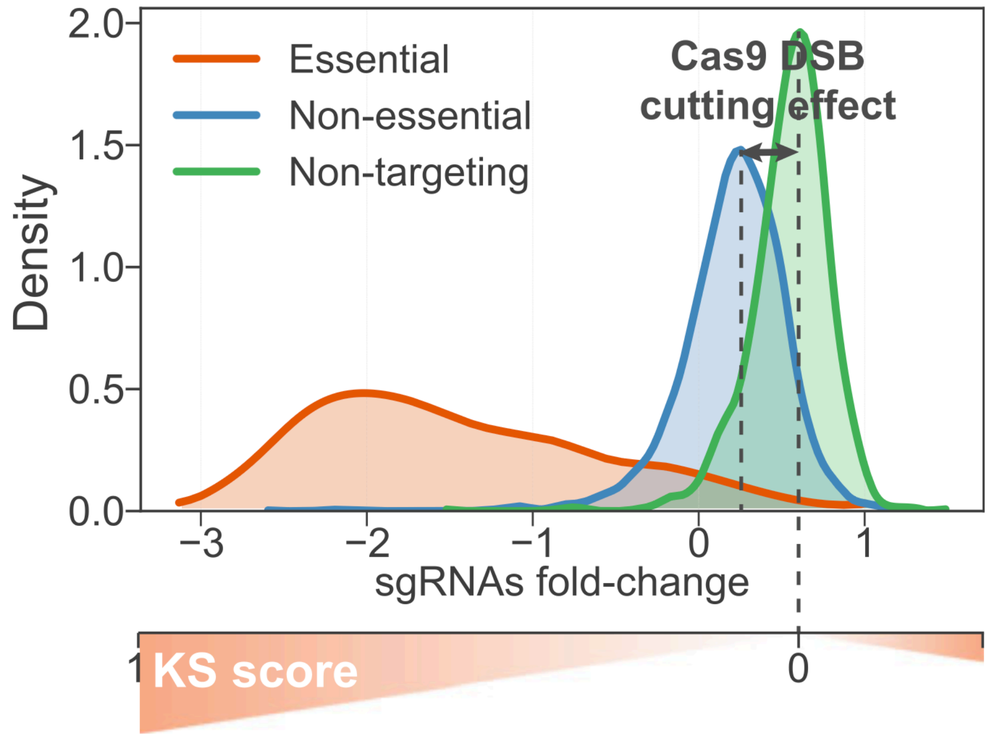
2 sgRNAs / Gene

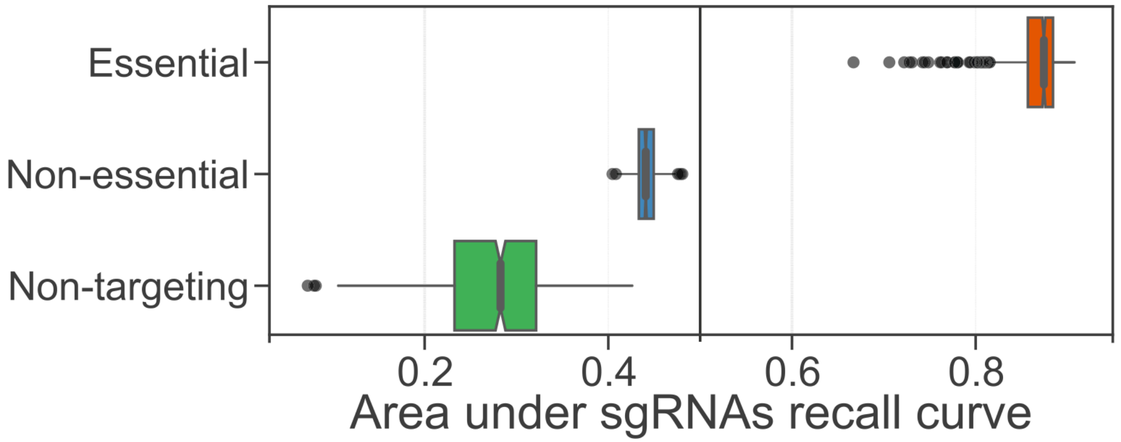
37,522 sgRNAs + 200 non-targeting sgRNAs - 62.7% decrease over the original library



Gonçalves et al. *Genome Biol*. 2021.

Prioritisation of guides with stronger “on-target” effects





CRISPR-Cas9-mediated DNA double-strand breaks induces a weak loss-of-fitness effect.

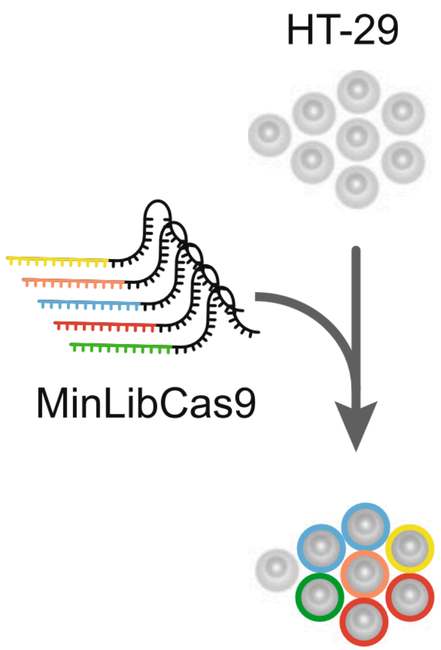
Non-targeting sgRNAs show relative growth advantage, due to the absence of a DNA double-stranded breaks.

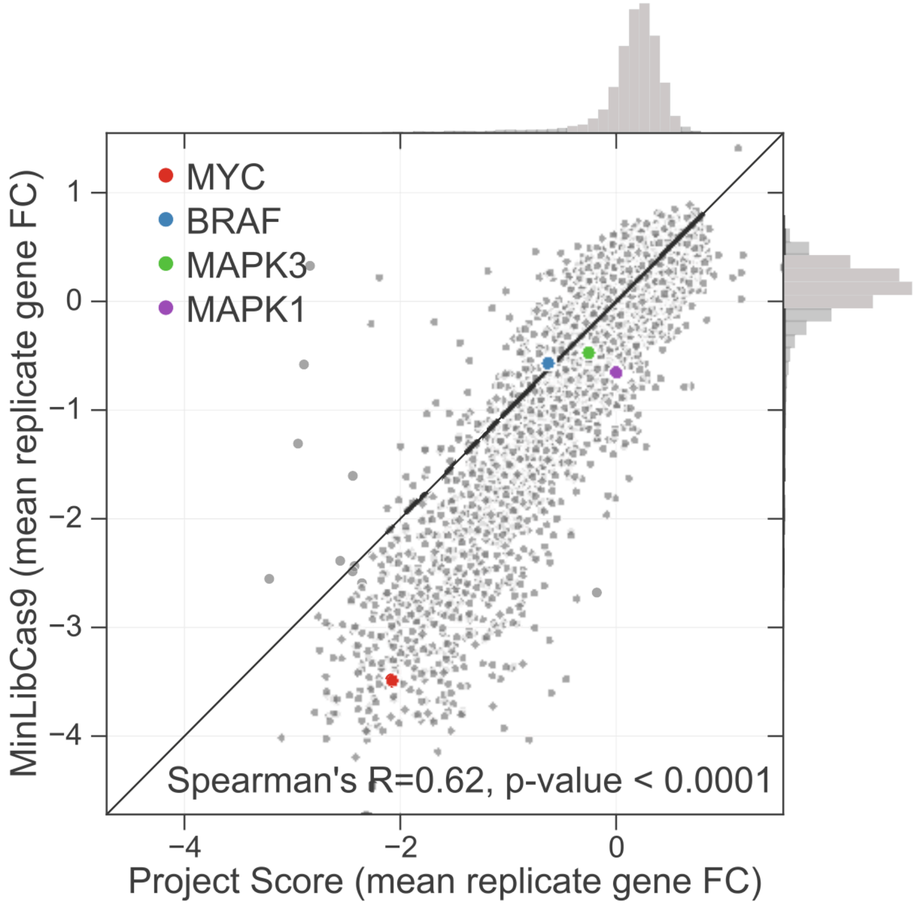
Empirically find “optimal on-target sgRNAs”, comparing sgRNA fold-change distribution to non-targeting guides.

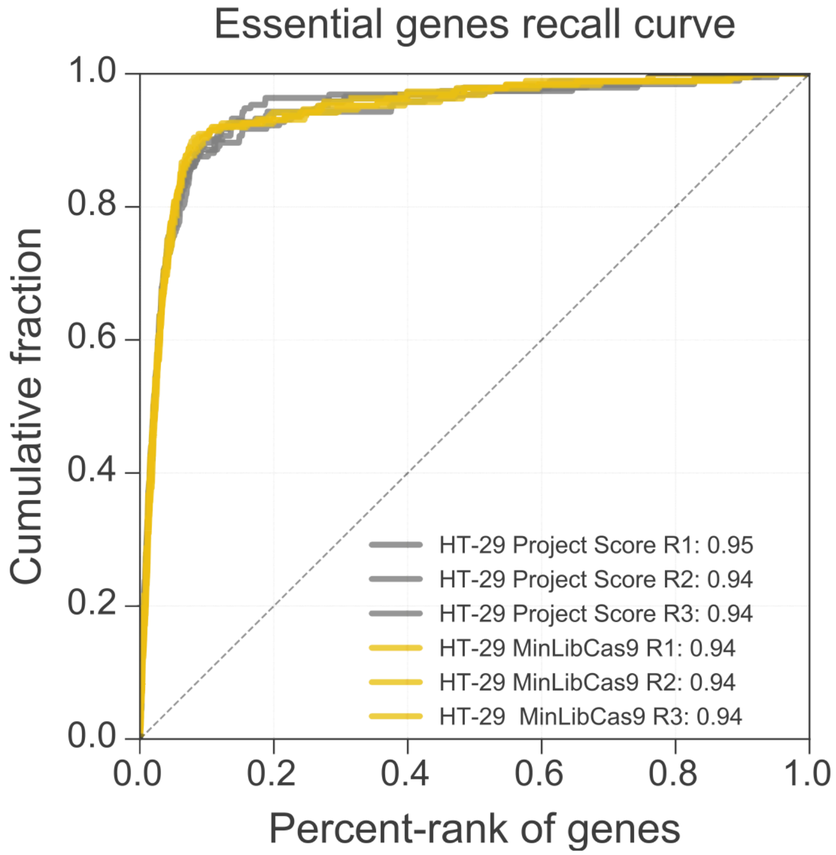
Stringent off-target and on-target filtering:

- Off-target summaries using WGE
- Multiple guide efficacy metrics (JACKS, Rule Set 2, FORECasT)
- CRISPOR scores, e.g. MIT specificity and CrisprScan

MinLibCas9 recapitulates large genome-wide libraries and increases gene fold-change range







Synthesised MinLibCas9 library and re-screened HT-29 colorectal cancer cell line.

MinLibCas9 showed an higher fold-change range, improving the identification of cancer dependencies.

MinLibCas9 V2.0

Motivation

•New library sgRNA metrics available (e.g. RuleSet3, VBS, CRISPRspec)

•50-80 MinLibCas9 screens available

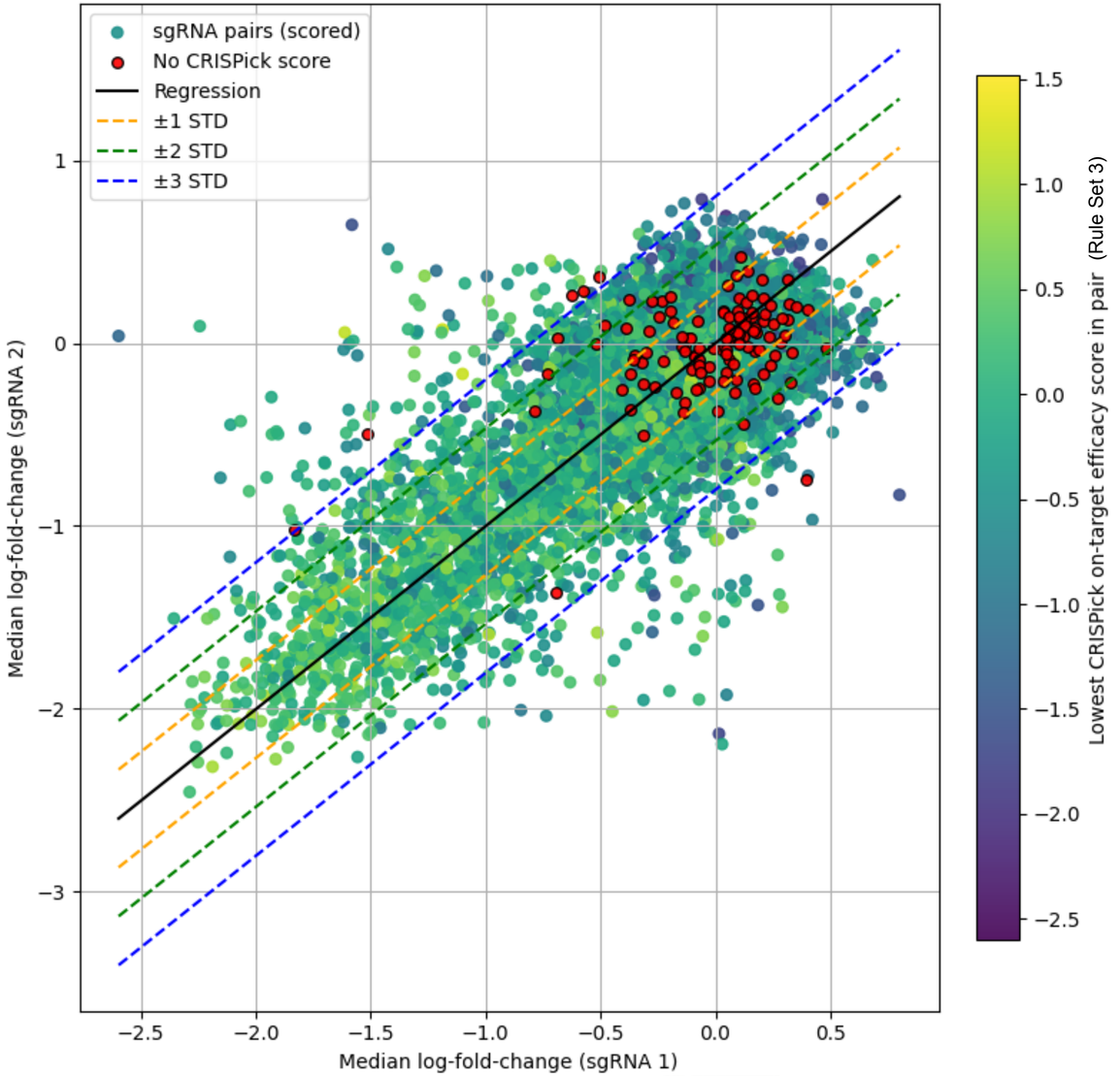
Contribution

•MinLibV2.0, a refined and improved library with additional gene coverage

Challenges

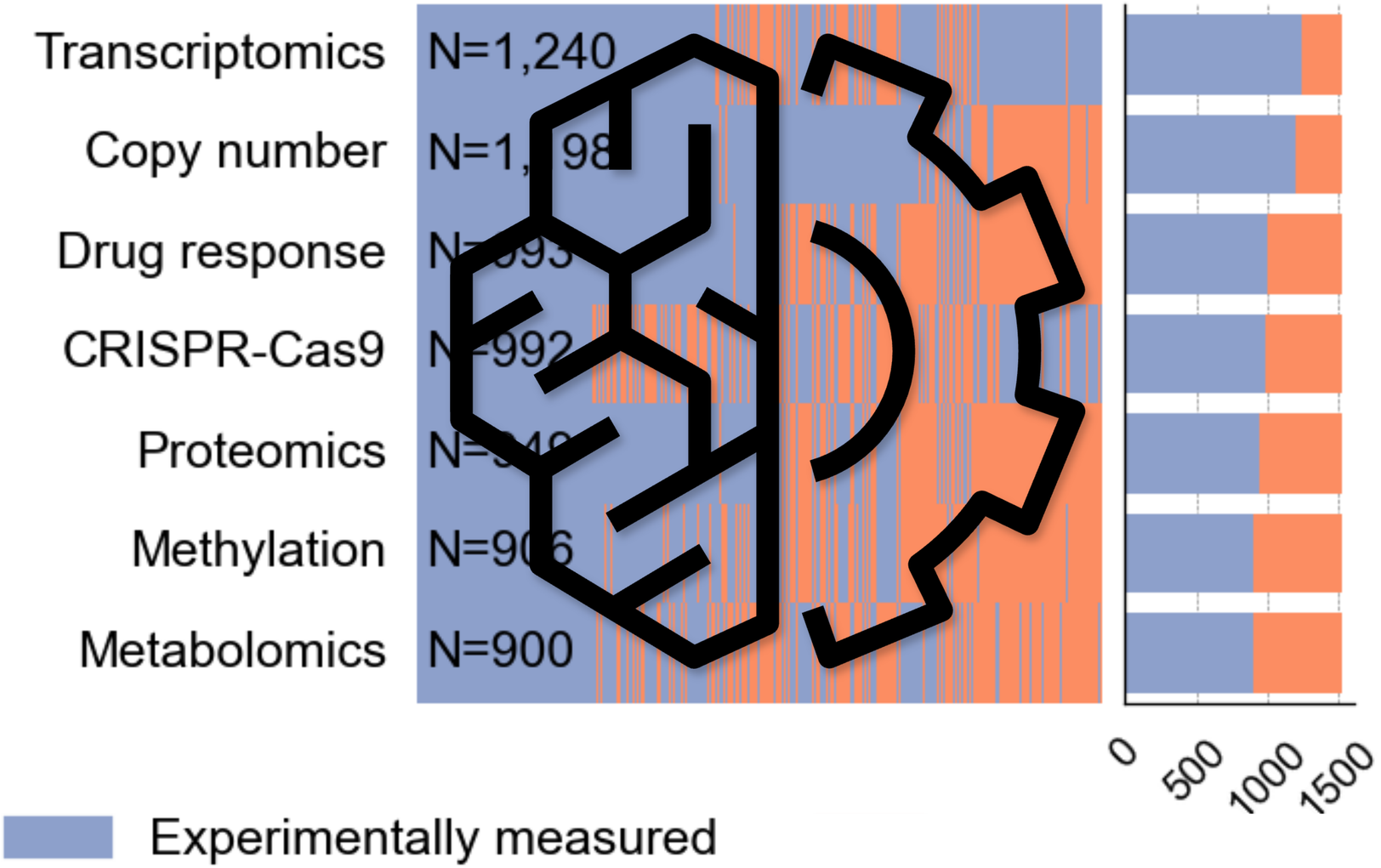
•Many disagreeing sgRNAs have little evidence from existing metrics

•Several sgRNAs show disagreement only in specific cancer cell moldes



In silico CRISPR-Cas9 screening

Multi-omics integration using deep generative models

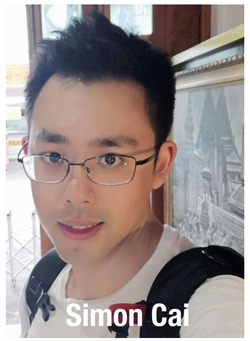


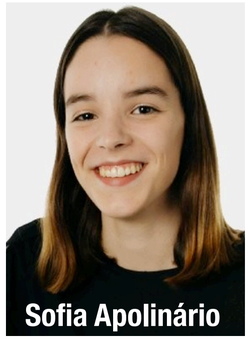
Multi-Omic Synthetic Augmentation

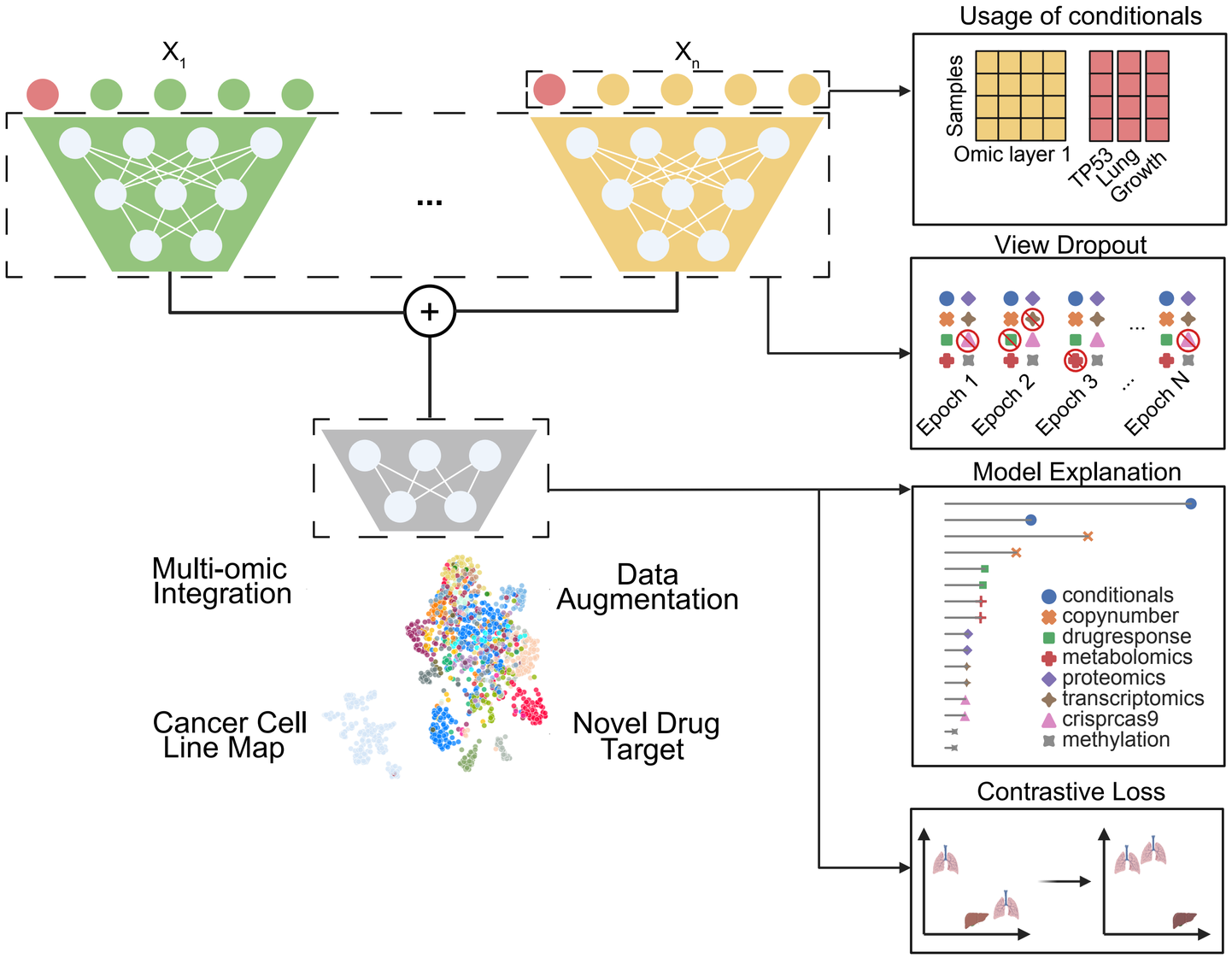
Unsupervised deep learning model — Multi-view masked variational autoencoder

Natively developed for CancerDepMap data, integrating 7 different omics

Generative model with the capacity of augmenting current cancer datasets







Cai, Apolinário, …, Gonçalves. *Nature Comms*. 2024.

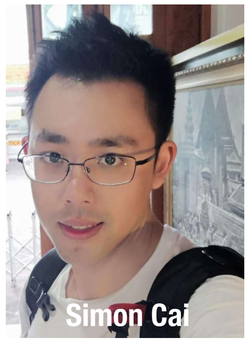
Omics | Mutations + Tissue + Other | Omics (removed features) | Latent dimensions

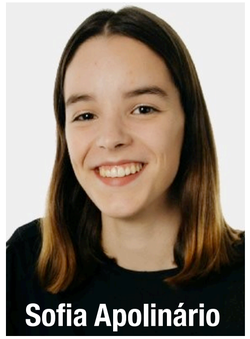
Multi-Omic Synthetic Augmentation

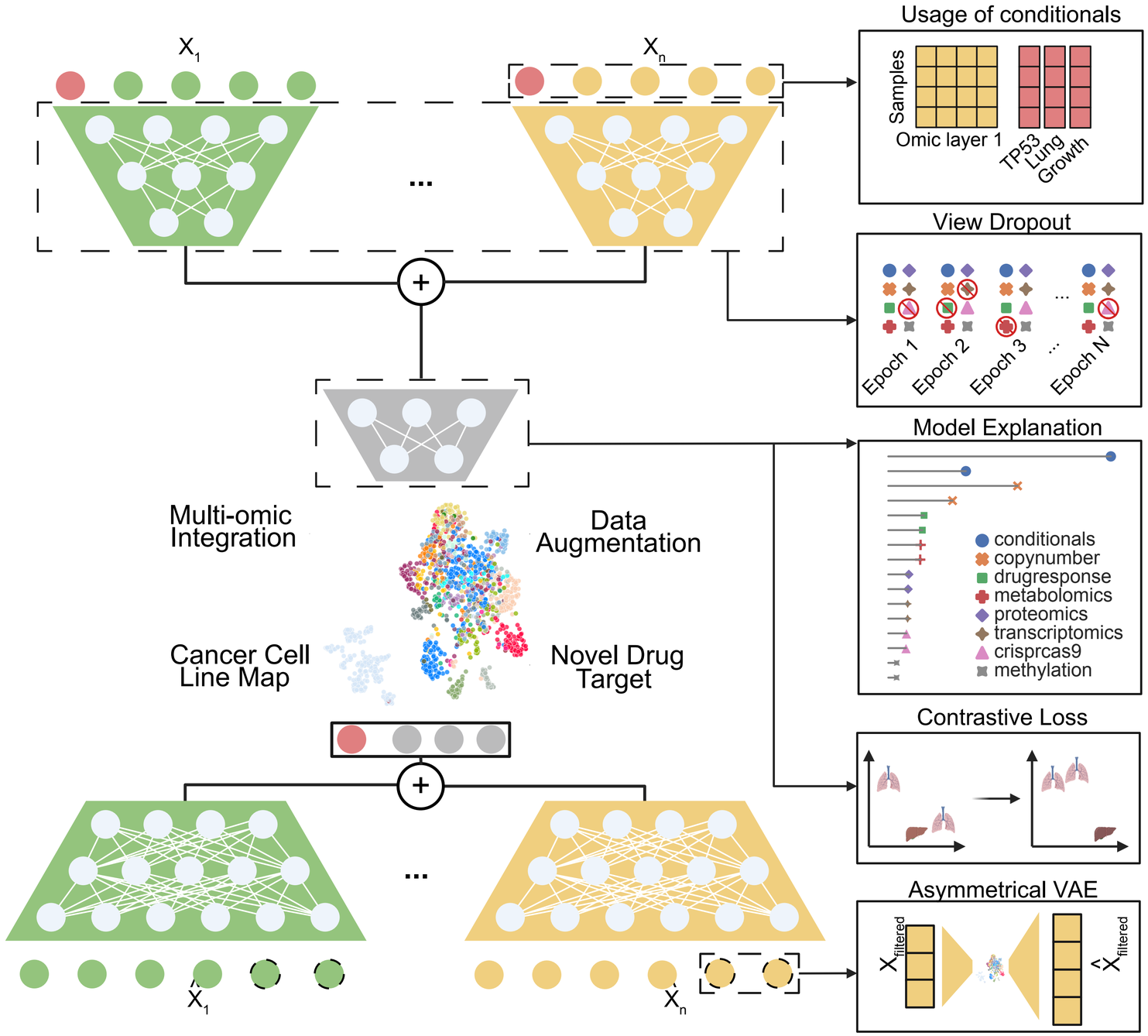
Unsupervised deep learning model — Multi-view masked variational autoencoder

Natively developed for CancerDepMap data, integrating 7 different omics

Generative model with the capacity of augmenting current cancer datasets

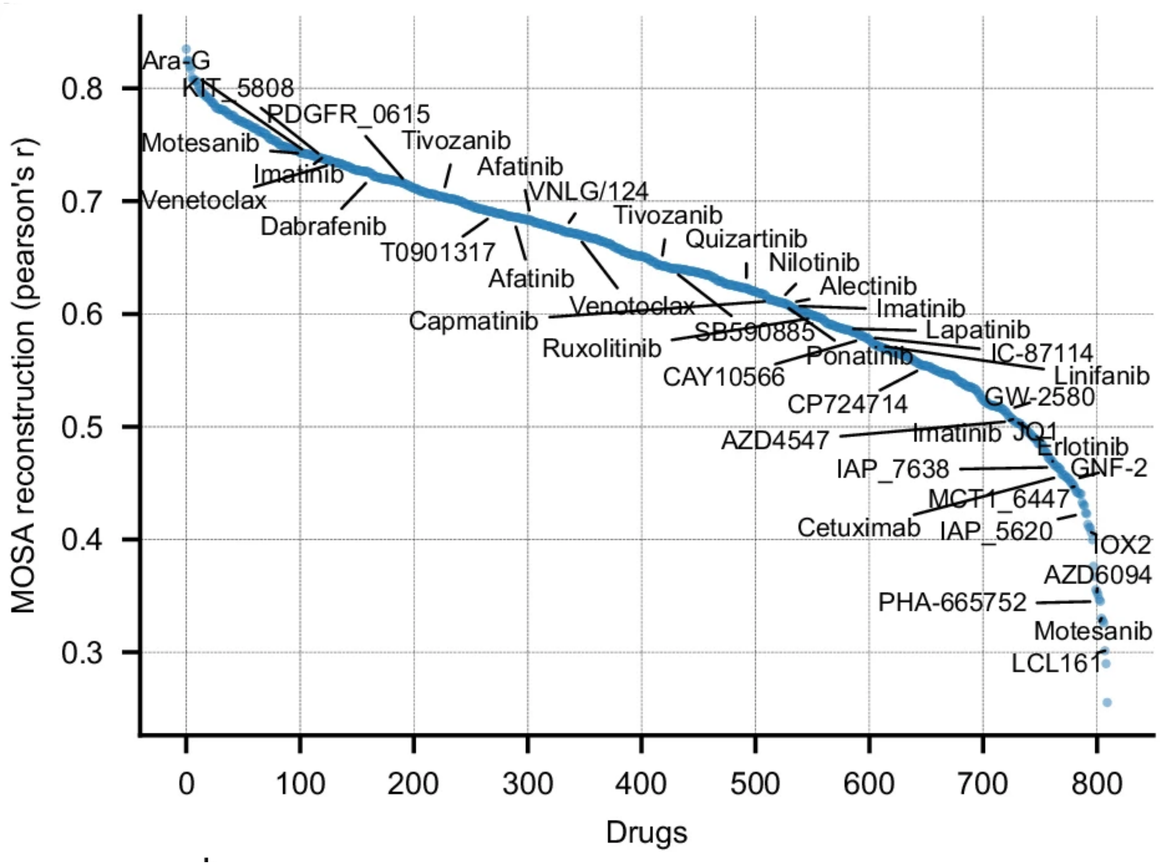


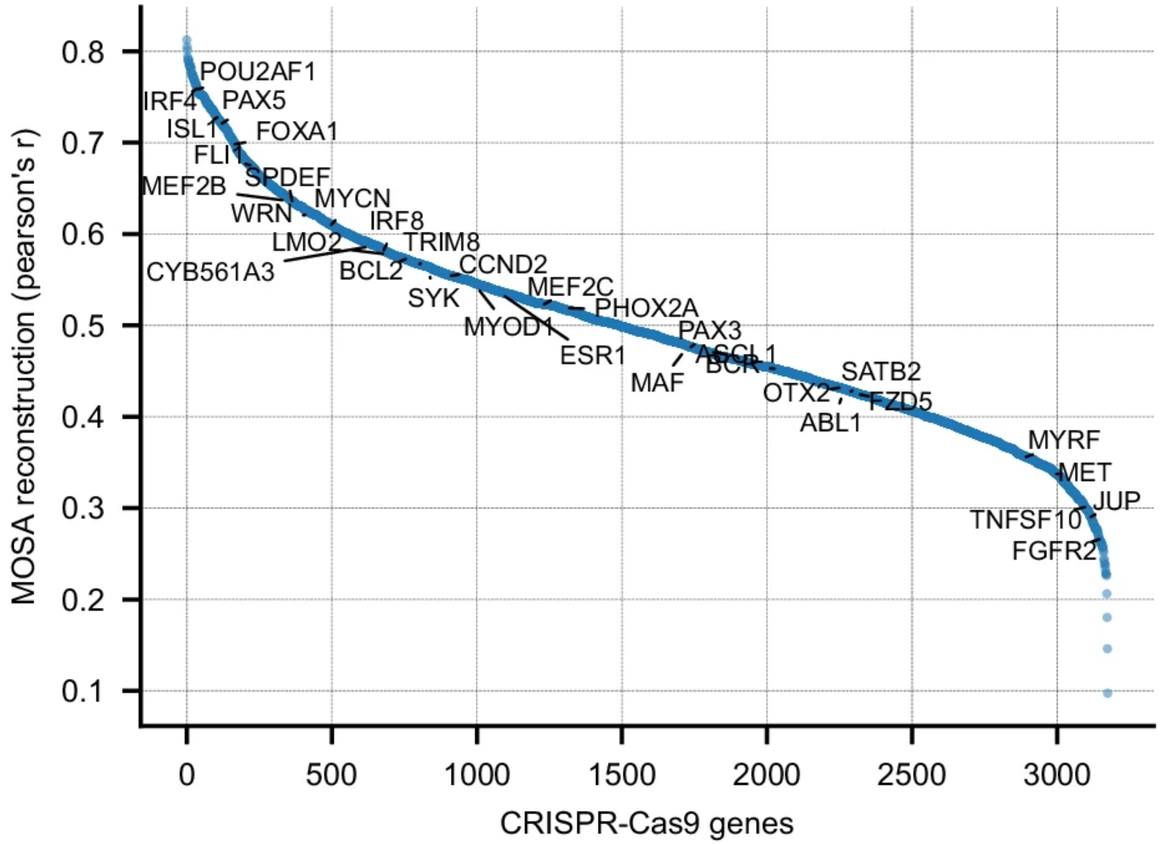




Cai, Apolinário, …, Gonçalves. *Nature Comms*. 2024.

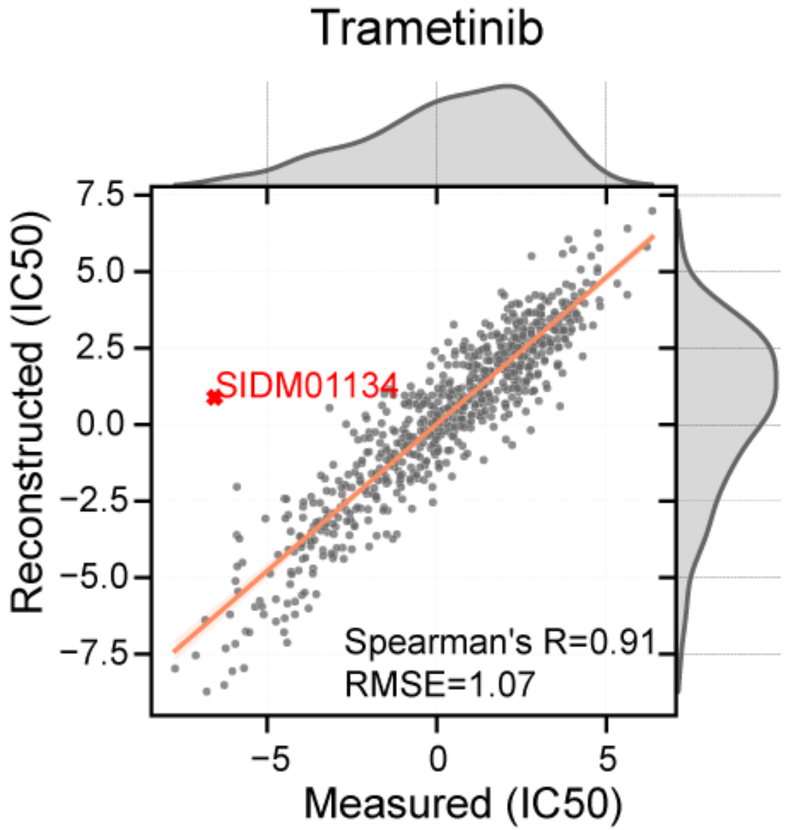
Overall reconstruction of drug response and CRISPR-Cas9





MOSA reconstruction evaluation using a 10-fold cross-validation - all test folds concatenated and the quality score is calculated as the Pearson's r between the reconstructed and actual measured values.

Inconsistencies between synthetic and original measurements



Inconsistencies revealed i) likely incorrect experimental measurements and ii) drugs (e.g. venetoclax) or classes of drugs (e.g. antiapoptotic inhibitors) without effective molecular biomarkers

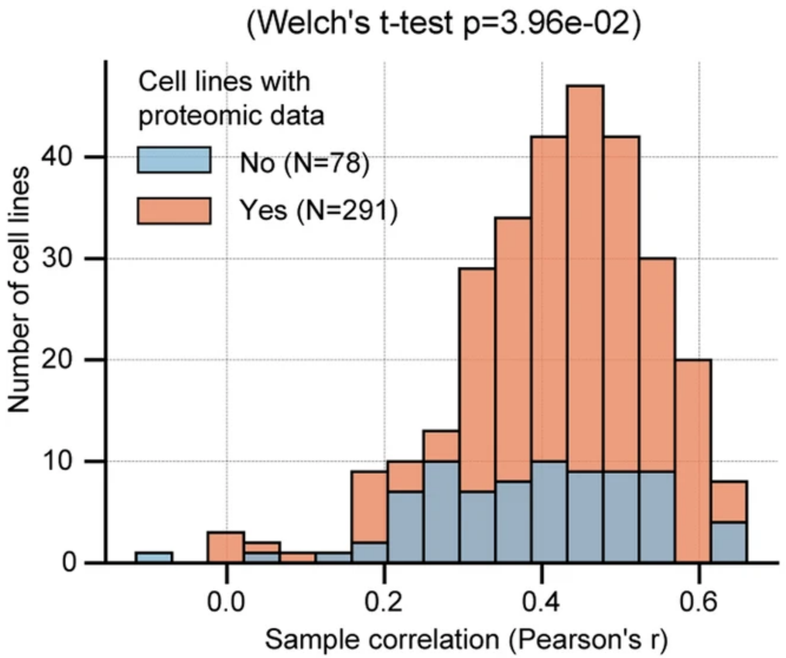
View/omic dropout improves model generalisation

Distribution of proteomics cancer cell lines correlation with an independent dataset (CCLE*).

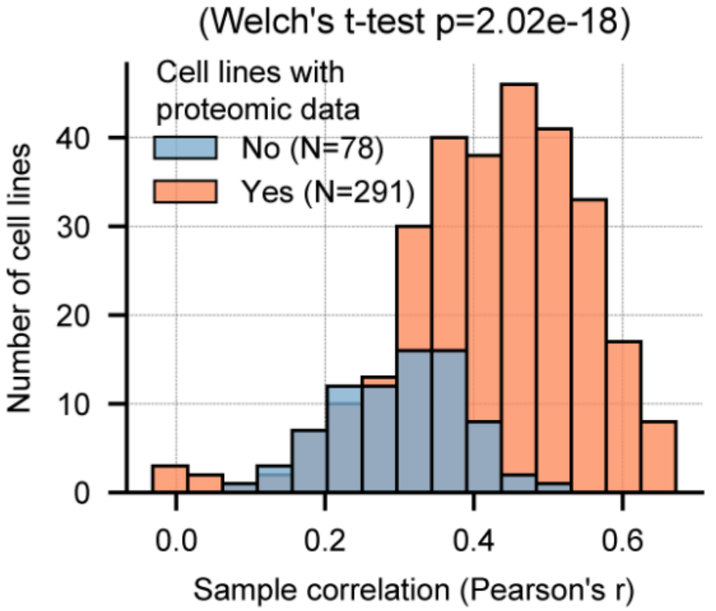
Grouping cancer cell lines that had proteomic data for the model training (orange, n = 291) versus cell lines without any proteomics prior (light blue, n = 78).

View dropout schema improves synthetic proteomics profiles.

View Dropout = **True**



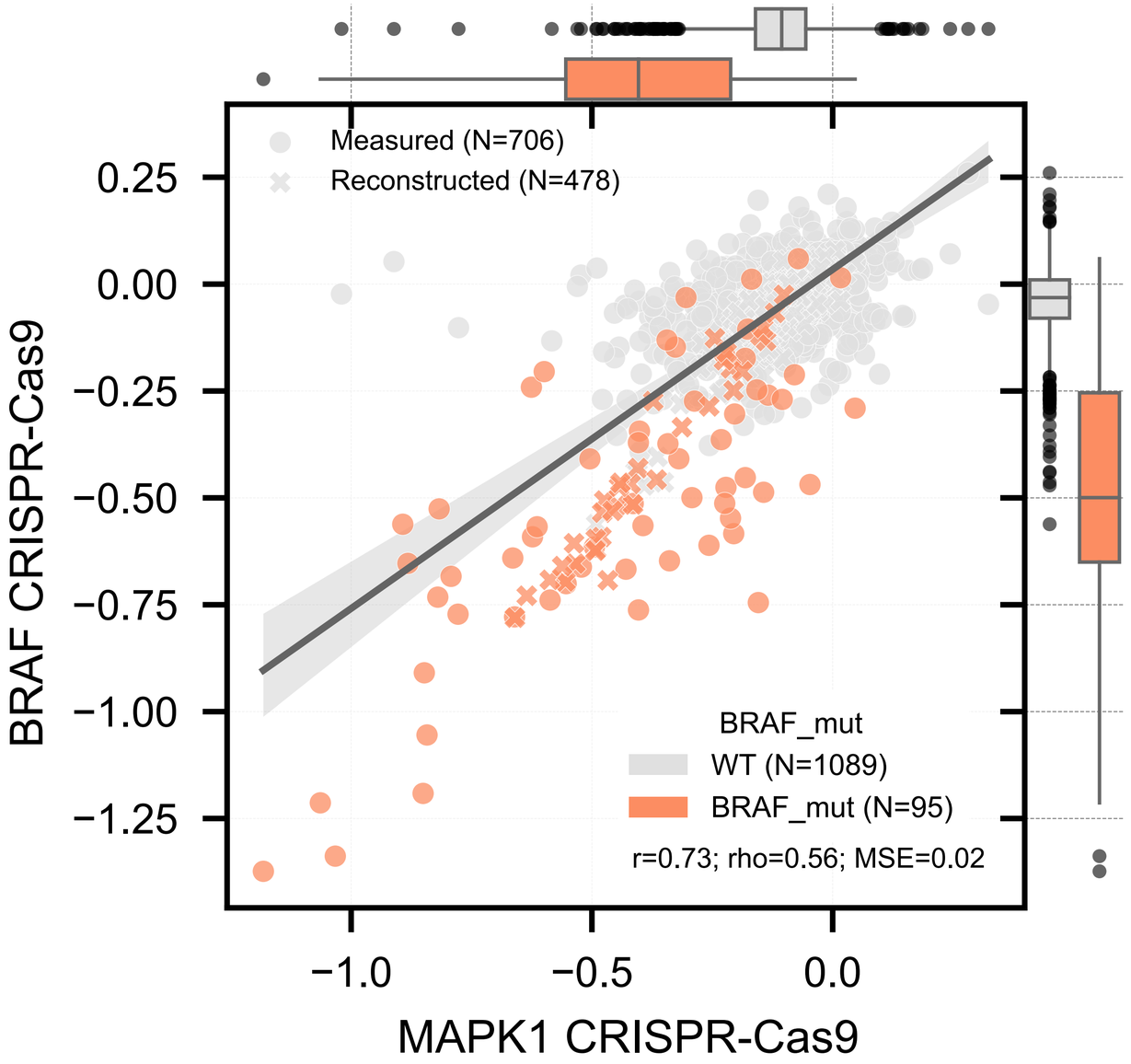
View Dropout = **False**

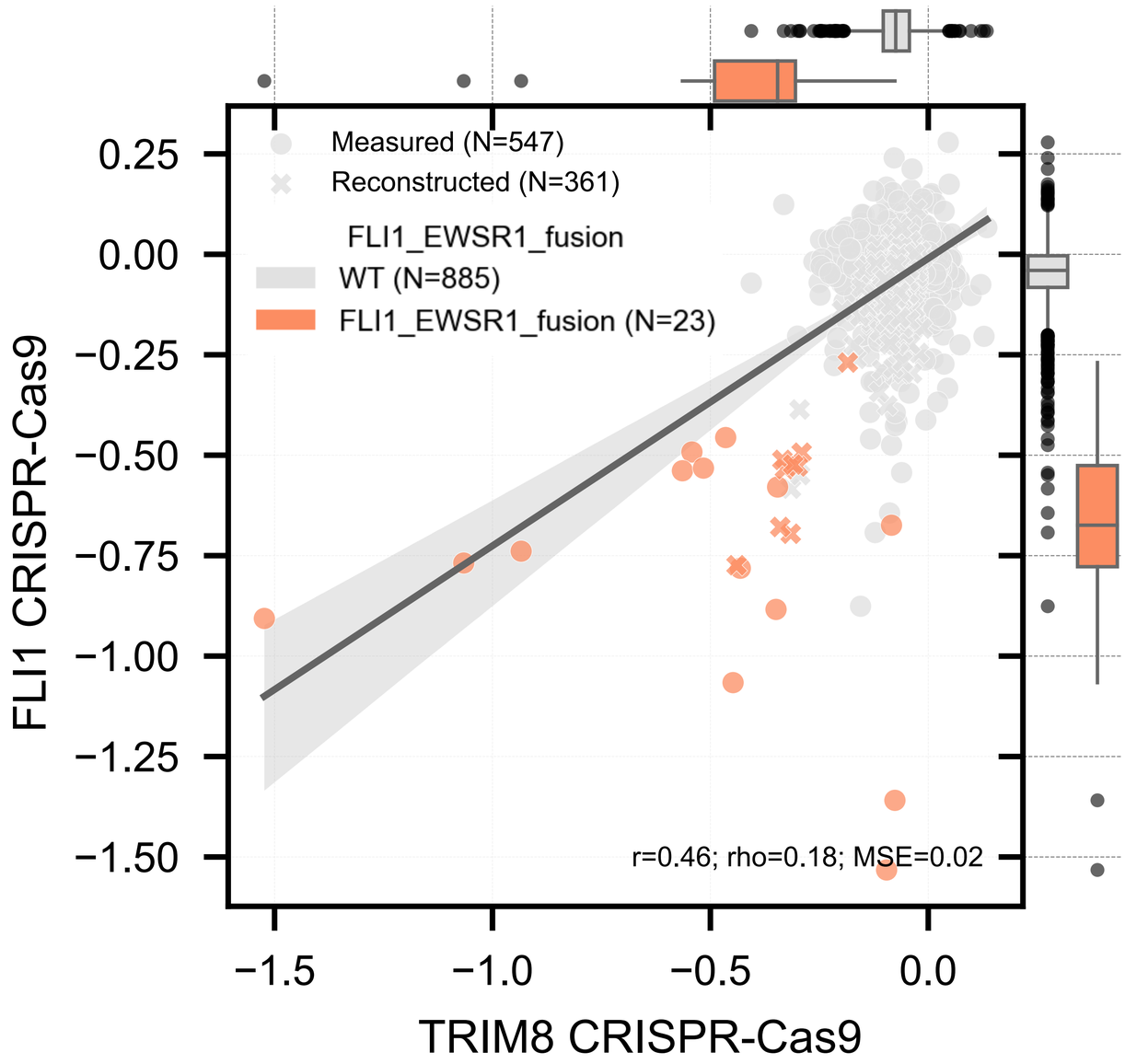


* Nusinow, et al. *Cell*. 2020.

MOSA reconstructs realistic Transcriptomics and CRISPR-Cas9 screens

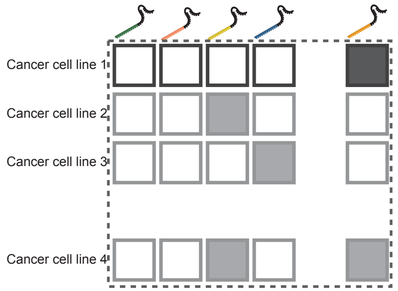
MOSA synthetic generation of CRISPR-Cas9 screens





Finding genetic biomarkers of essential genes with synthetically augmented data

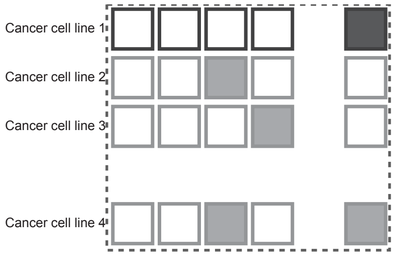
Gene Essentiality
(CRISPR-Cas9)



Continuous
(log2 fold-change)

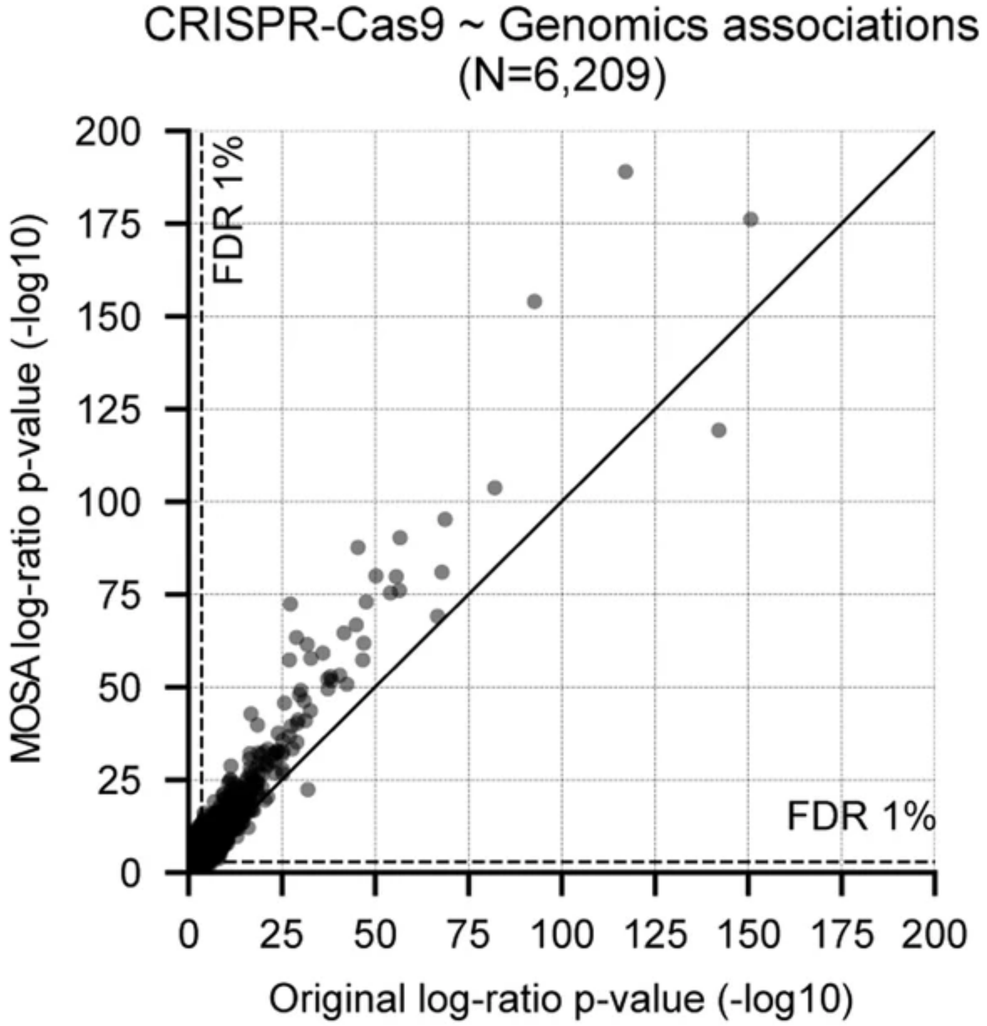
`Y = f(X)`

Genetic Alterations
(e.g. mutations)



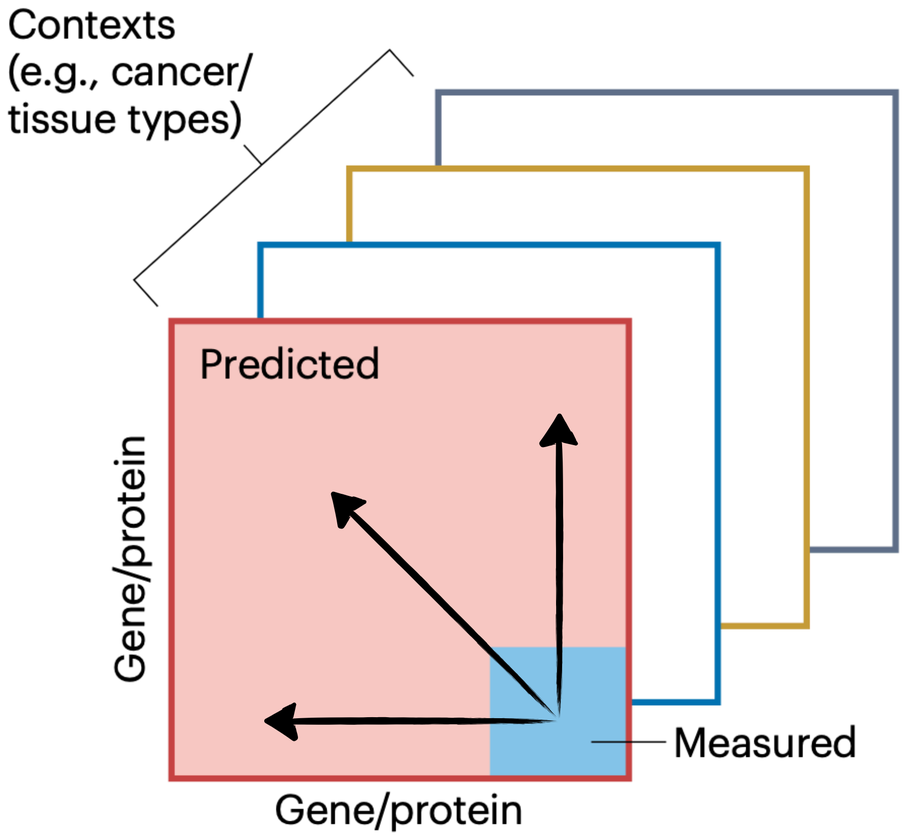
Binary
(1=present; 0=absent)

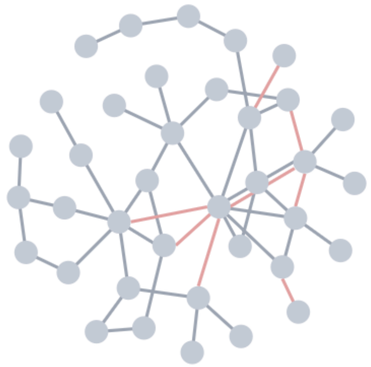
=>



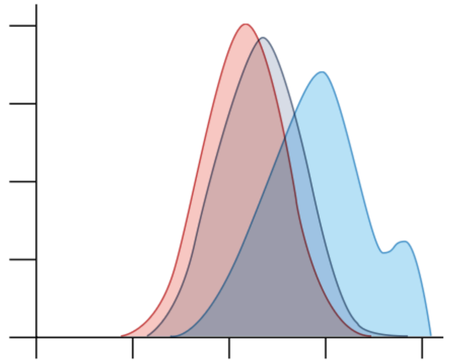
One-sided log-ratio test p-value of genetic associations with CRISPR-Cas9 gene essentiality with the original dataset (x-axis) and the augmented MOSA dataset (y-axis).

Explore synthetic lethal interactions

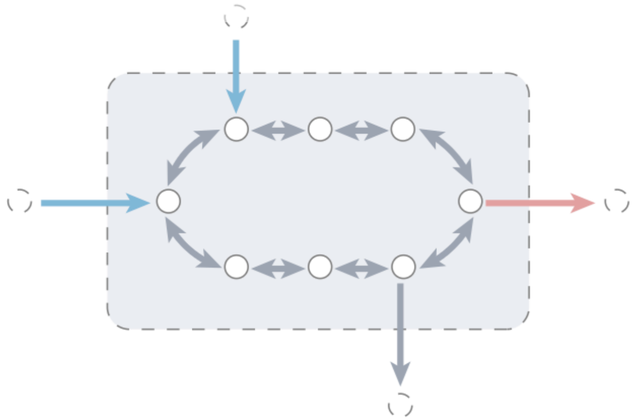


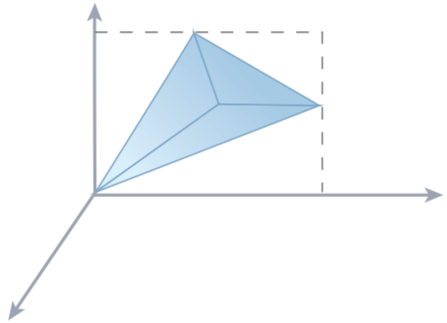


Gene/protein
network-based



Statistical/machine
learning based



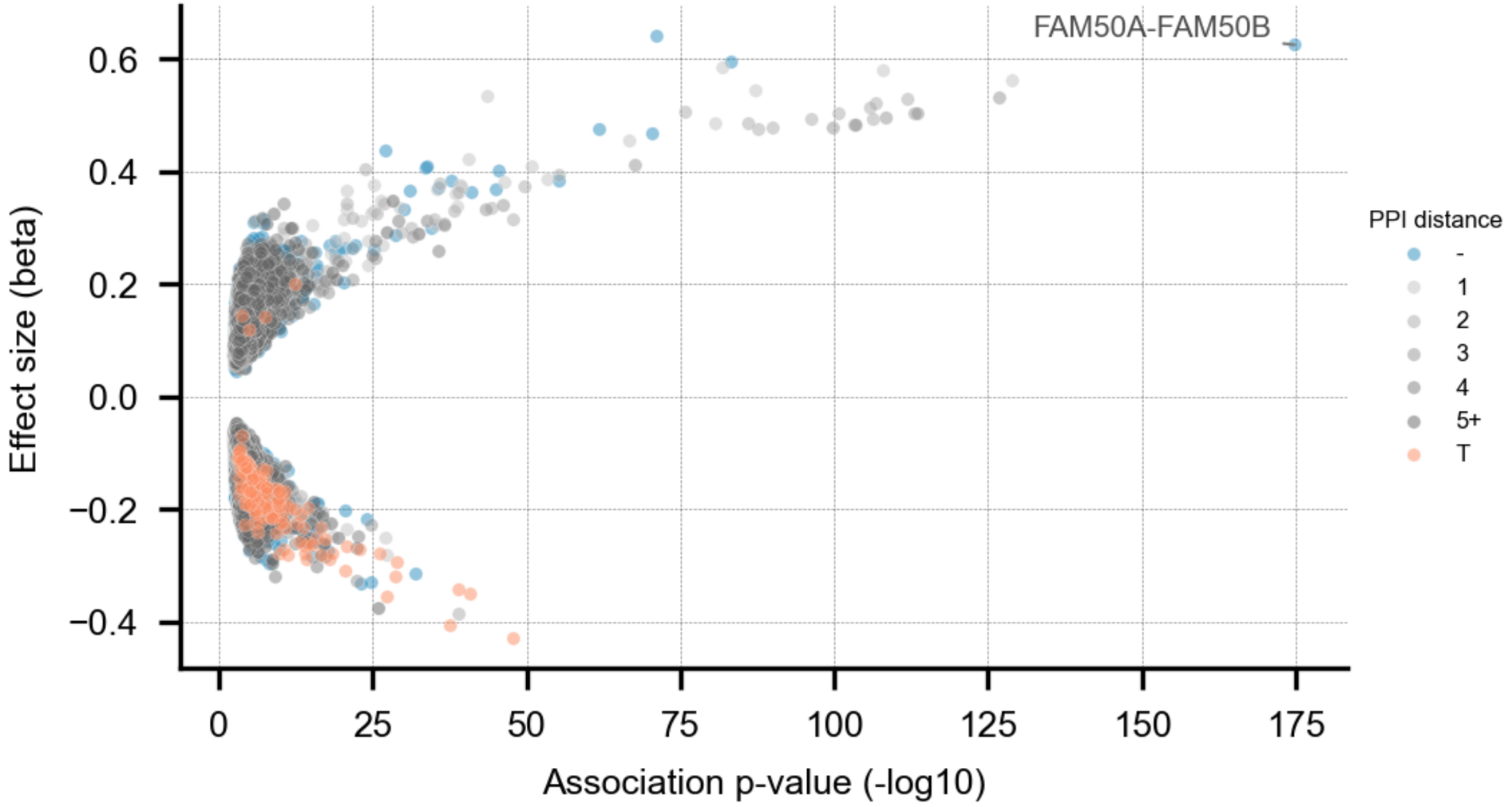


Genome-scale metabolic models

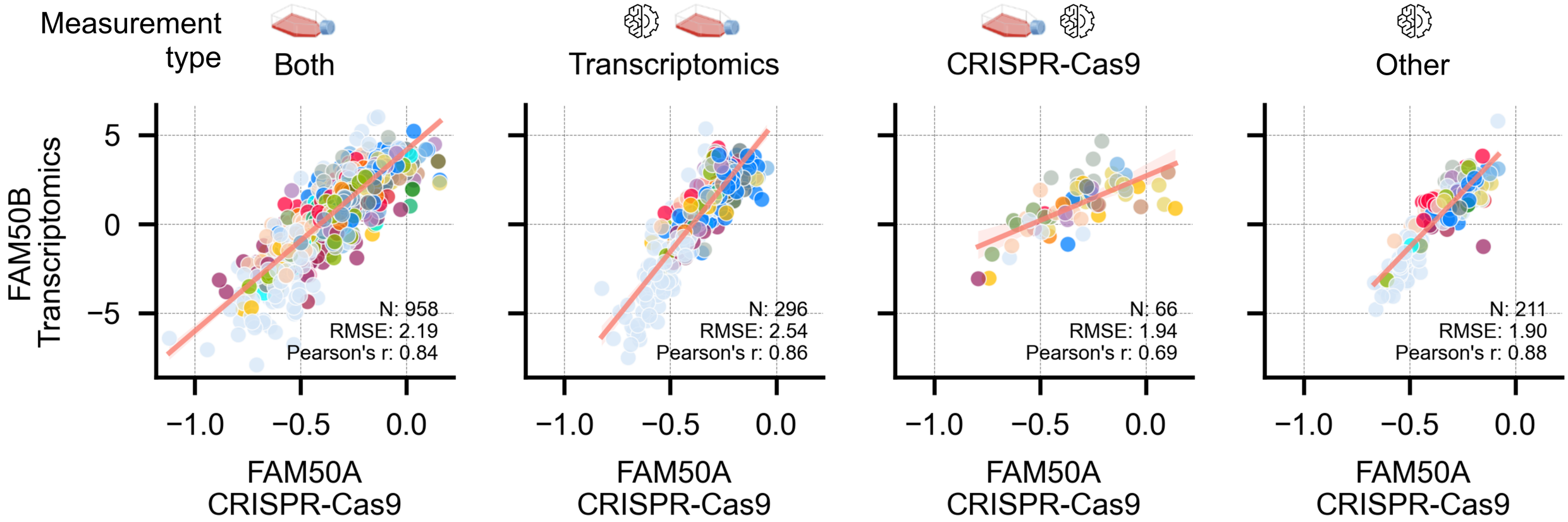
Gene/protein — Predicting SL interaction — Gene/protein

Gonçalves, Ryan, Adams. *Nature Reviews Drug Discovery*. 2025.

Finding synthetic lethal interactions using *in silico* augmented Transcriptomics and CRISPR-Cas9

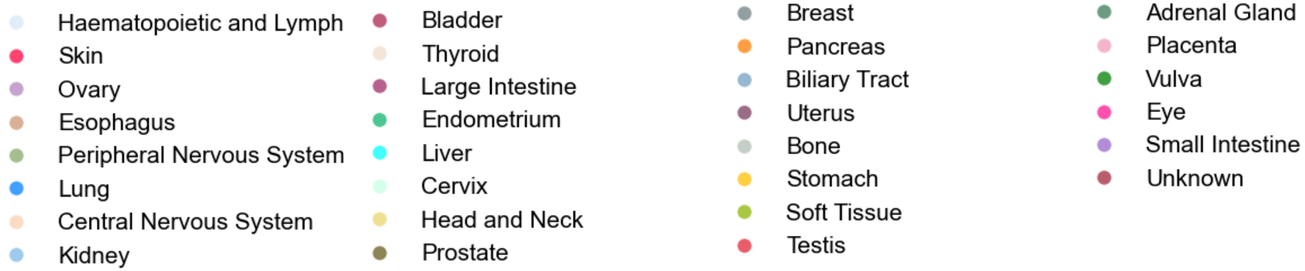


Robust synthetic reconstruction of gene expression and essentiality

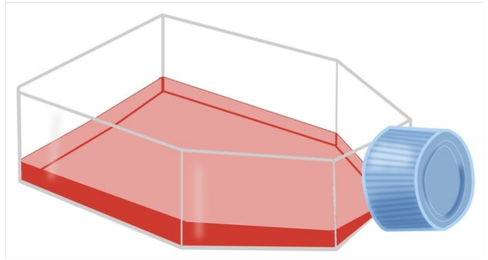


In silico prediction of cell line vulnerabilities

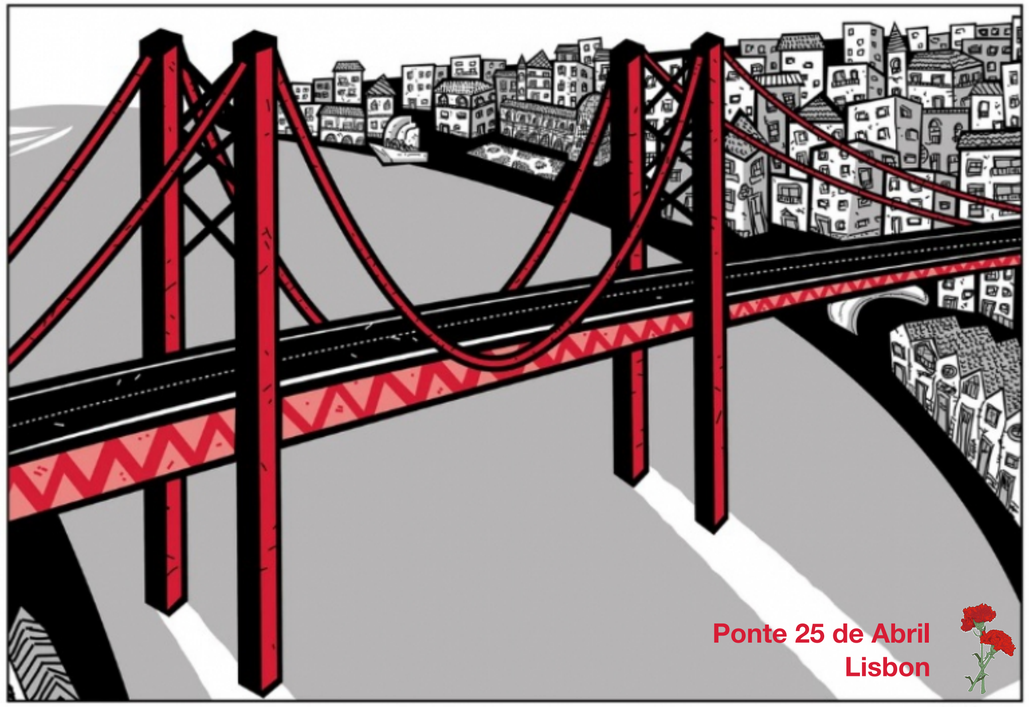
Guide further screens



Identify synthetic lethal associations that translate to the clinic

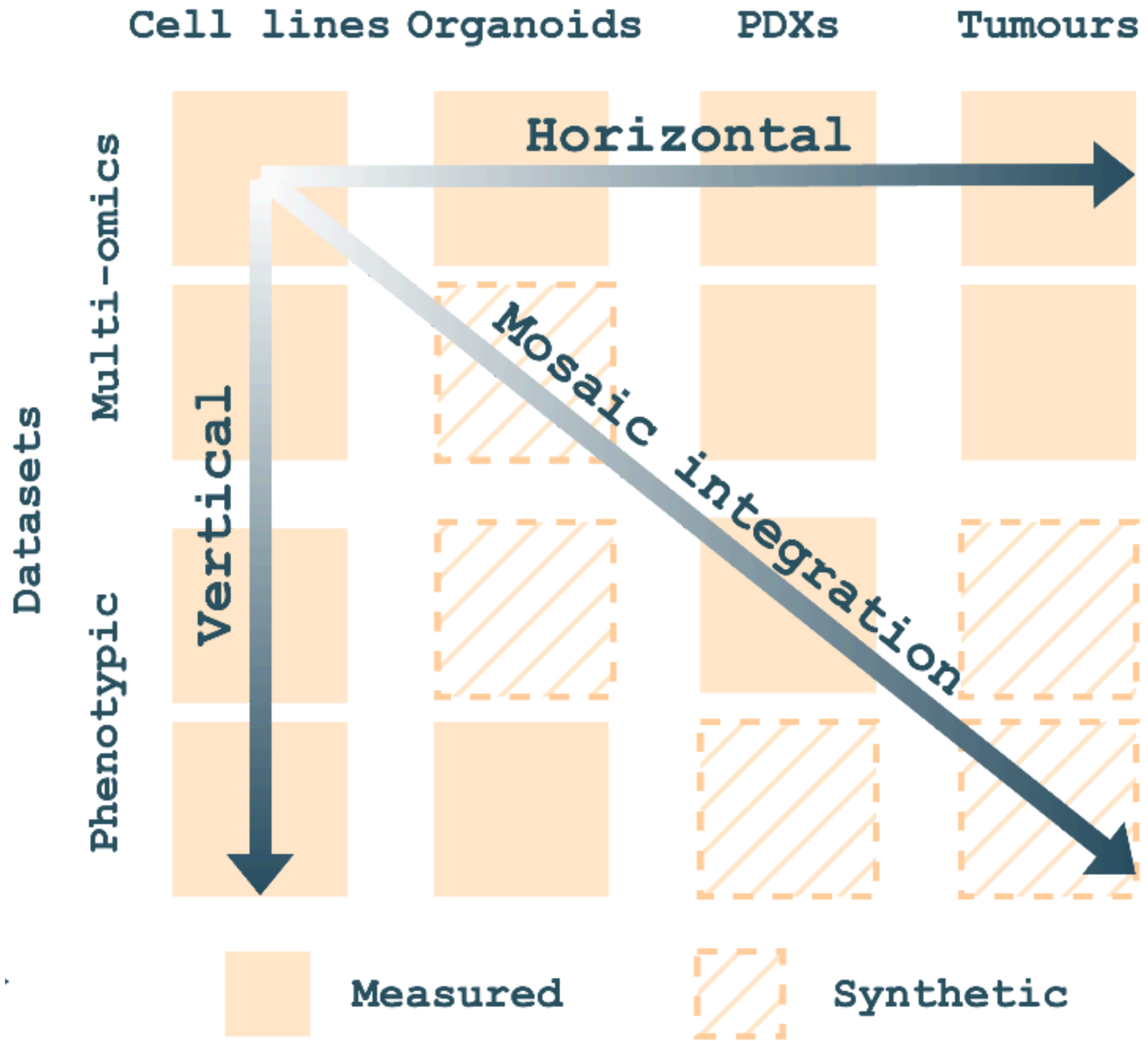


Lab

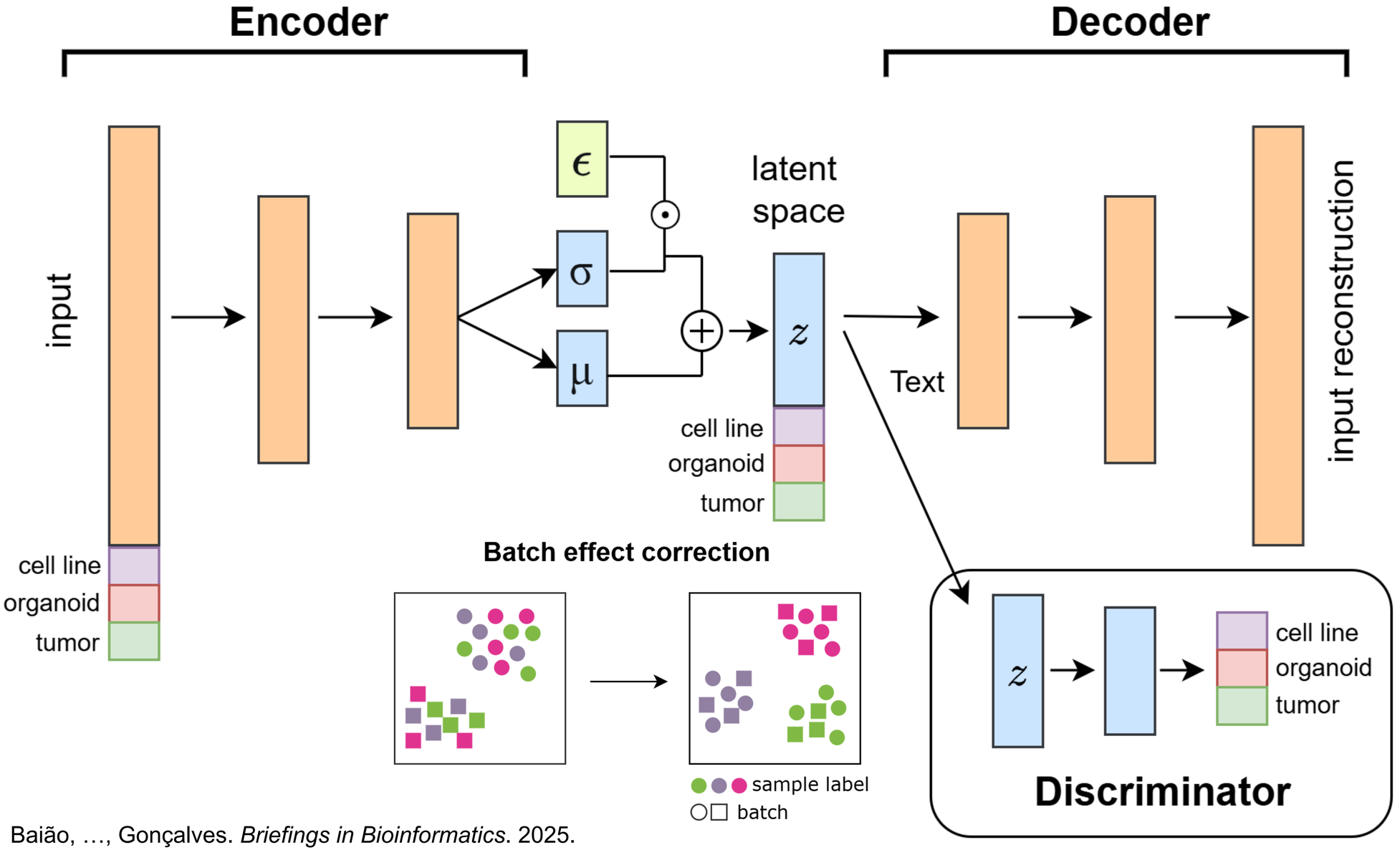


Clinic

Expanding integration across cancer cell models - Mosaic integration

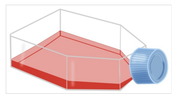


Deep learning methods for multi-omics integration across cell models



Baião, …, Gonçalves. *Briefings in Bioinformatics*. 2025.

Horizontal integration of transcriptomics across cancer cell models



Cell Lines

Organoids

Tumours

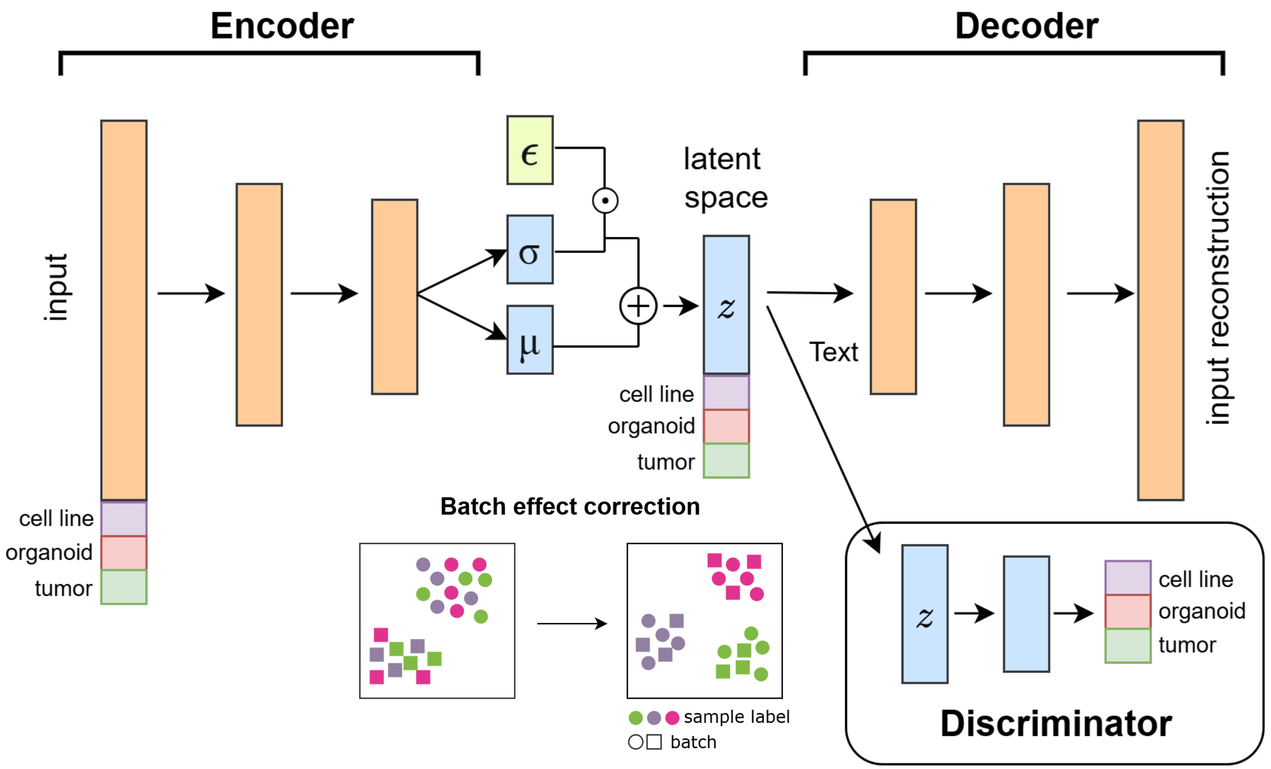
Transcriptomics
(genes=15,905)

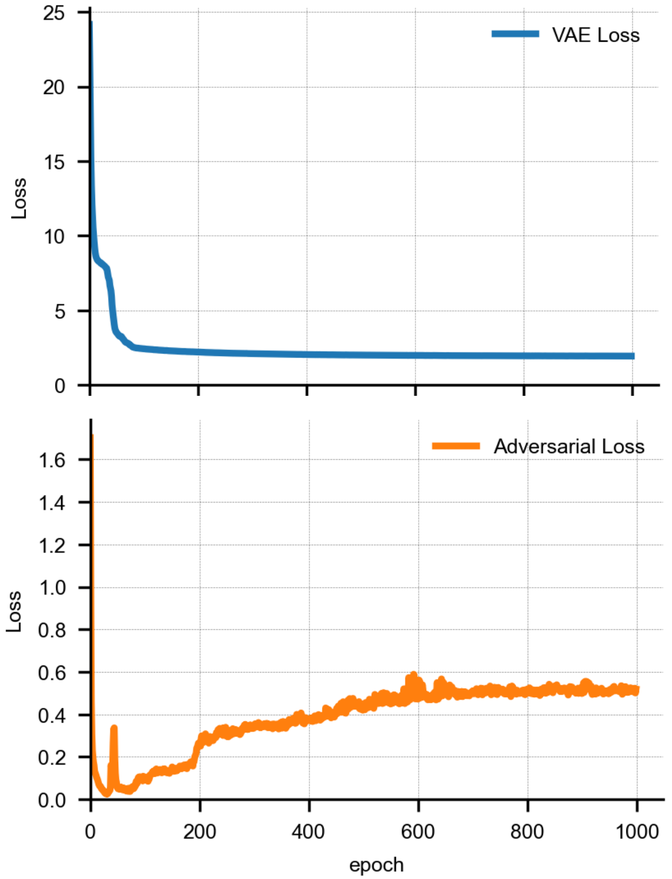
n = 1,293

n = 138

n = 3,474

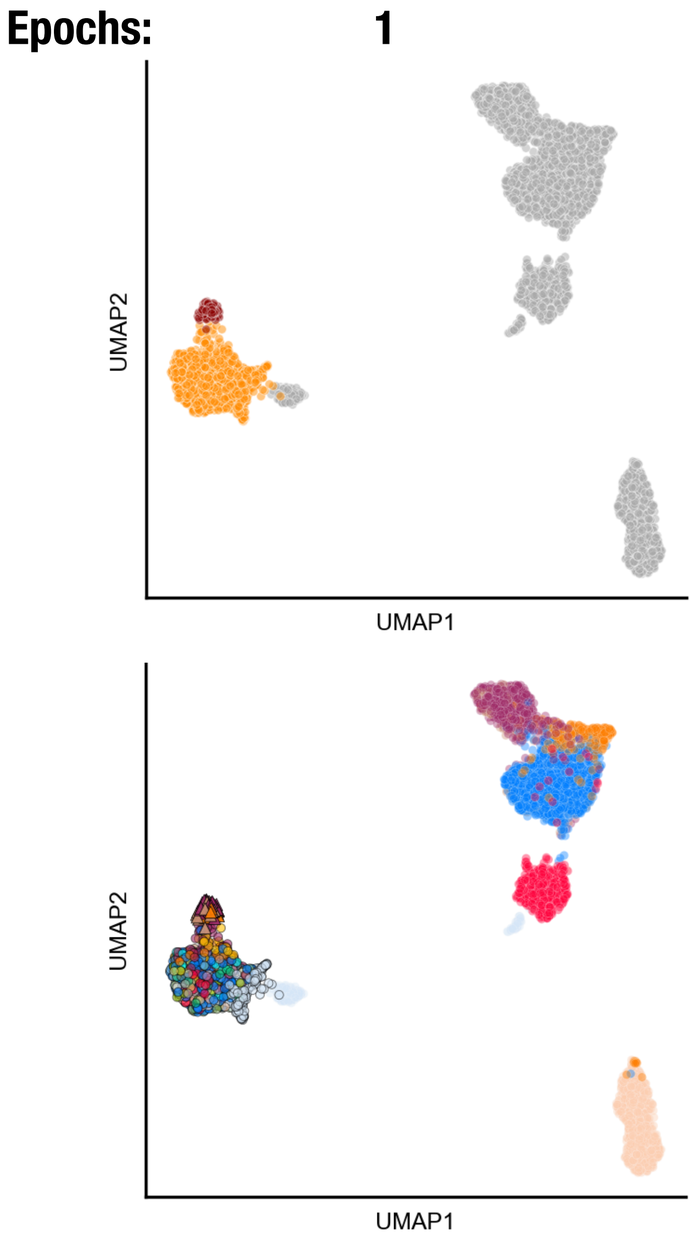
Horizontal Integration





Unpublished

Machine learning across cancer cell models is possible

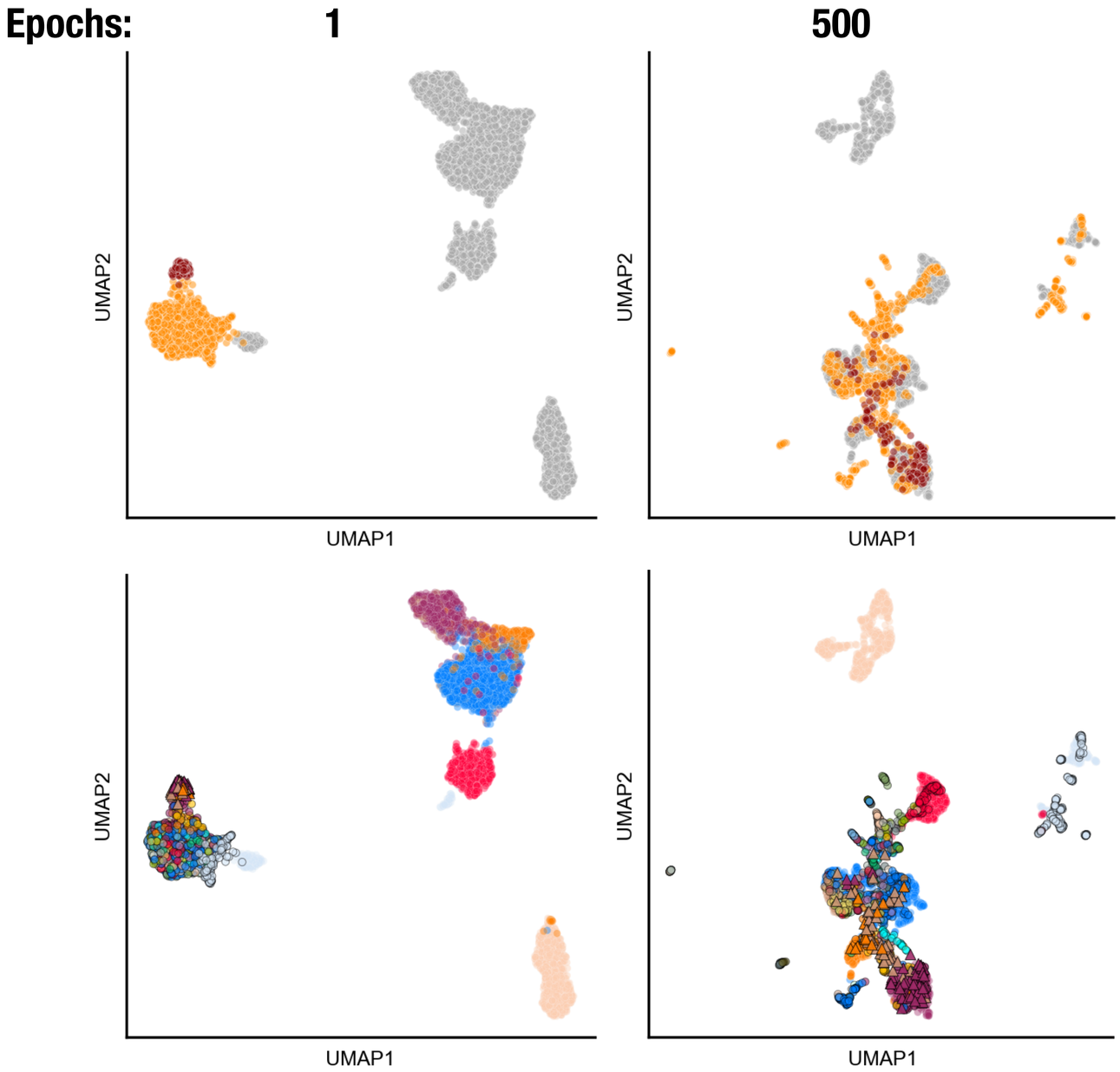


Sample Type
- Tumor
- Cell Line
- Organoid

Sample Type
- Tumor
- Cell Line
- Organoid

- Haematopoietic and Lymphoid
- Peripheral Nervous System
- Skin
- Breast
- Ovary
- Large Intestine
- Esophagus
- Lung
- Head and Neck
- Central Nervous System
- Kidney
- Bladder
- Stomach
- Pancreas
- Bone
- Thyroid
- Liver
- Prostate
- Endometrium
- Biliary Tract
- Uterus
- Cervix
- Testis
- Soft Tissue
- Small Intestine
- Adrenal Gland
- Vulva
- Placenta
- Unknown

Machine learning across cancer cell models is possible

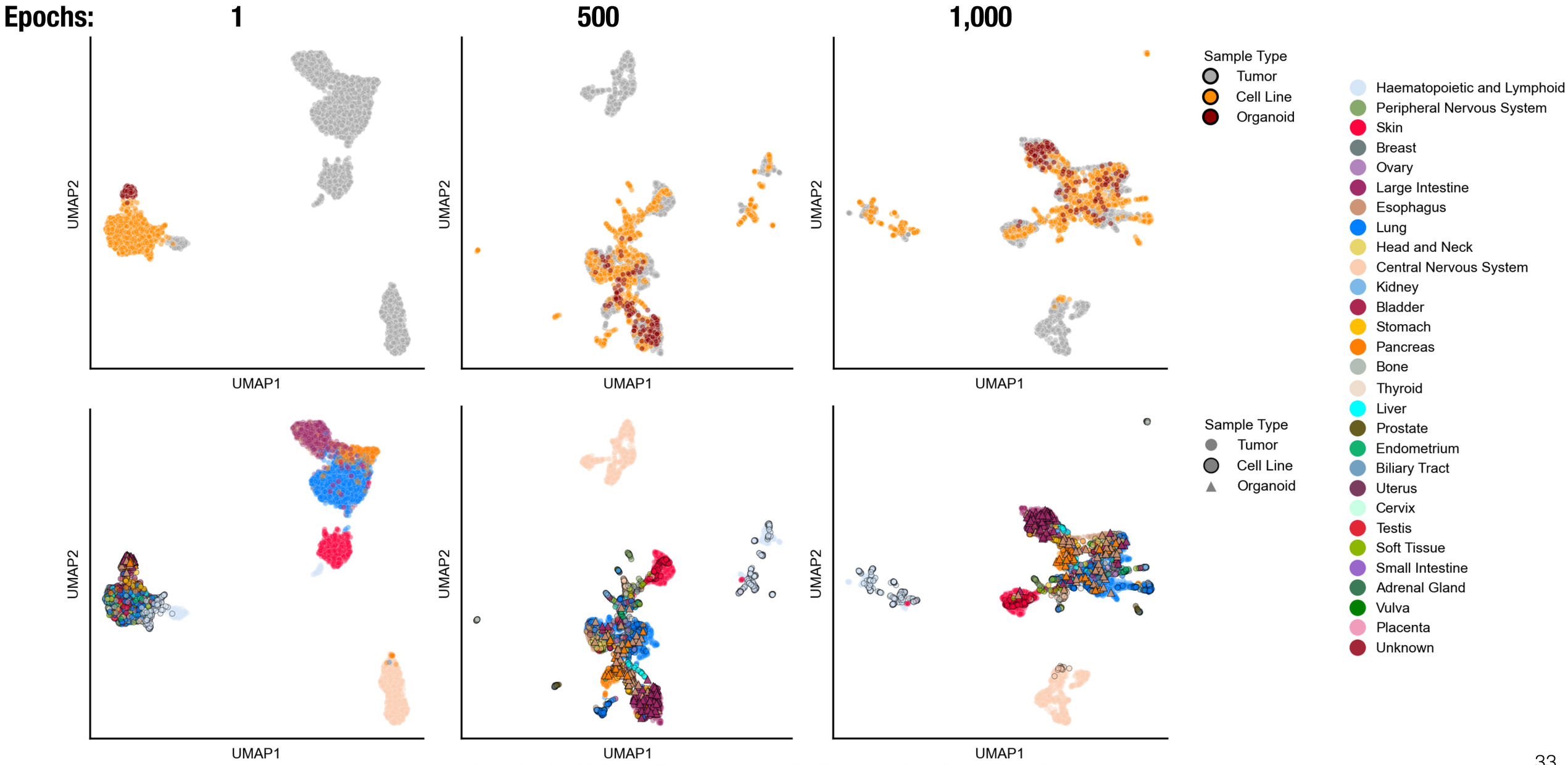


Sample Type
- Tumor
- Cell Line
- Organoid

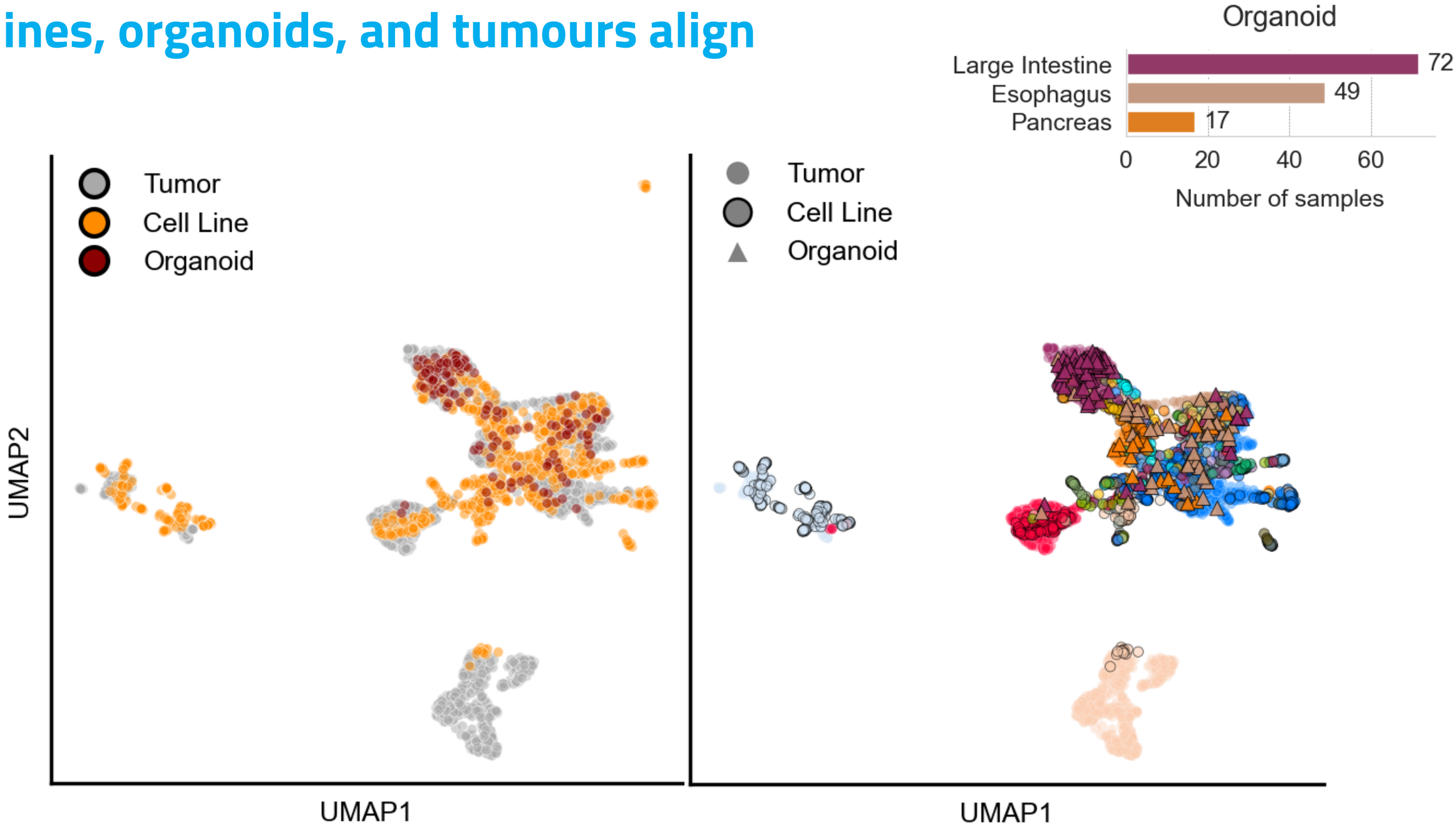
Sample Type
- Tumor
- Cell Line
- Organoid

- Haematopoietic and Lymphoid
- Peripheral Nervous System
- Skin
- Breast
- Ovary
- Large Intestine
- Esophagus
- Lung
- Head and Neck
- Central Nervous System
- Kidney
- Bladder
- Stomach
- Pancreas
- Bone
- Thyroid
- Liver
- Prostate
- Endometrium
- Biliary Tract
- Uterus
- Cervix
- Testis
- Soft Tissue
- Small Intestine
- Adrenal Gland
- Vulva
- Placenta
- Unknown

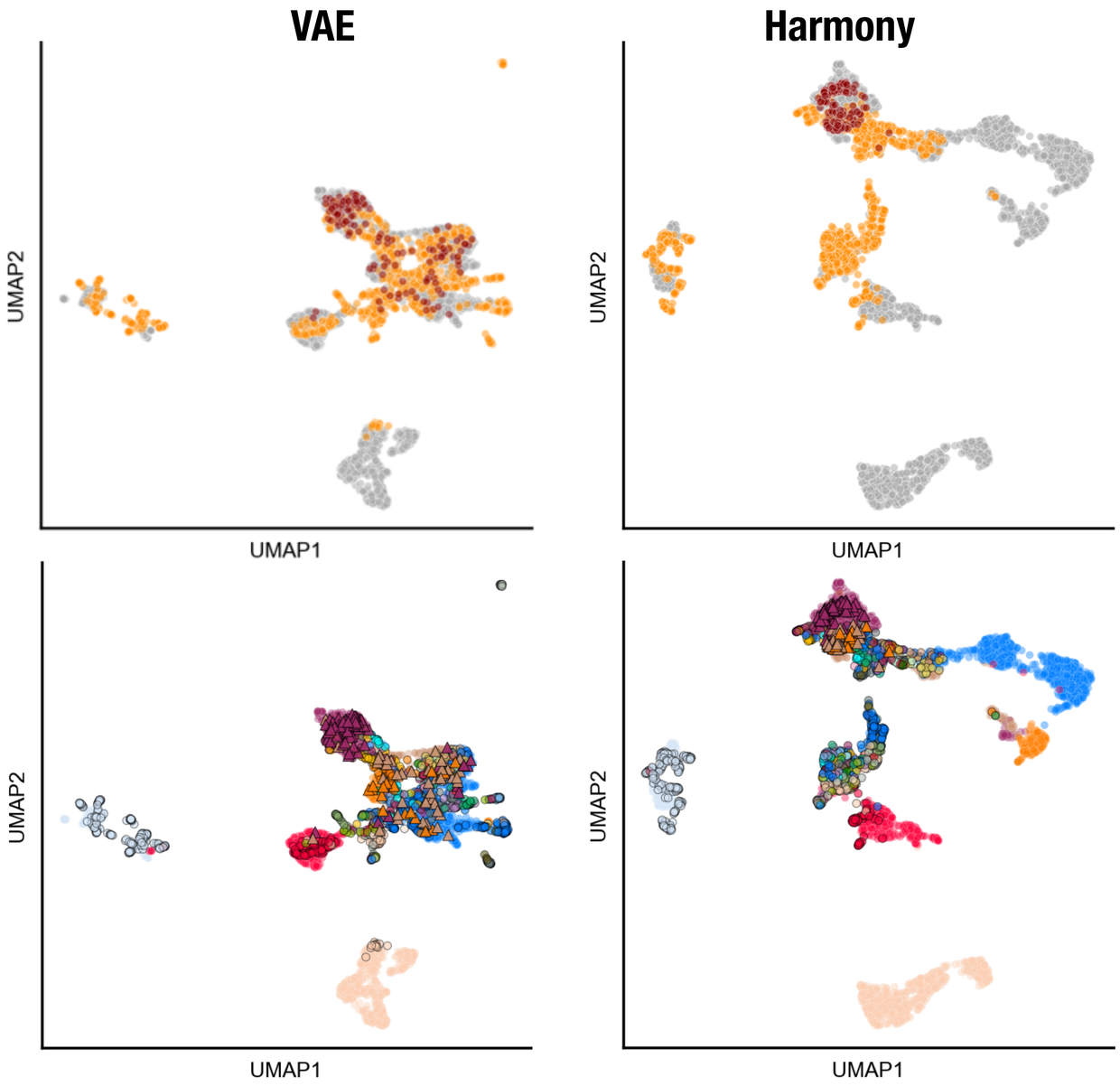
Machine learning across cancer cell models is possible

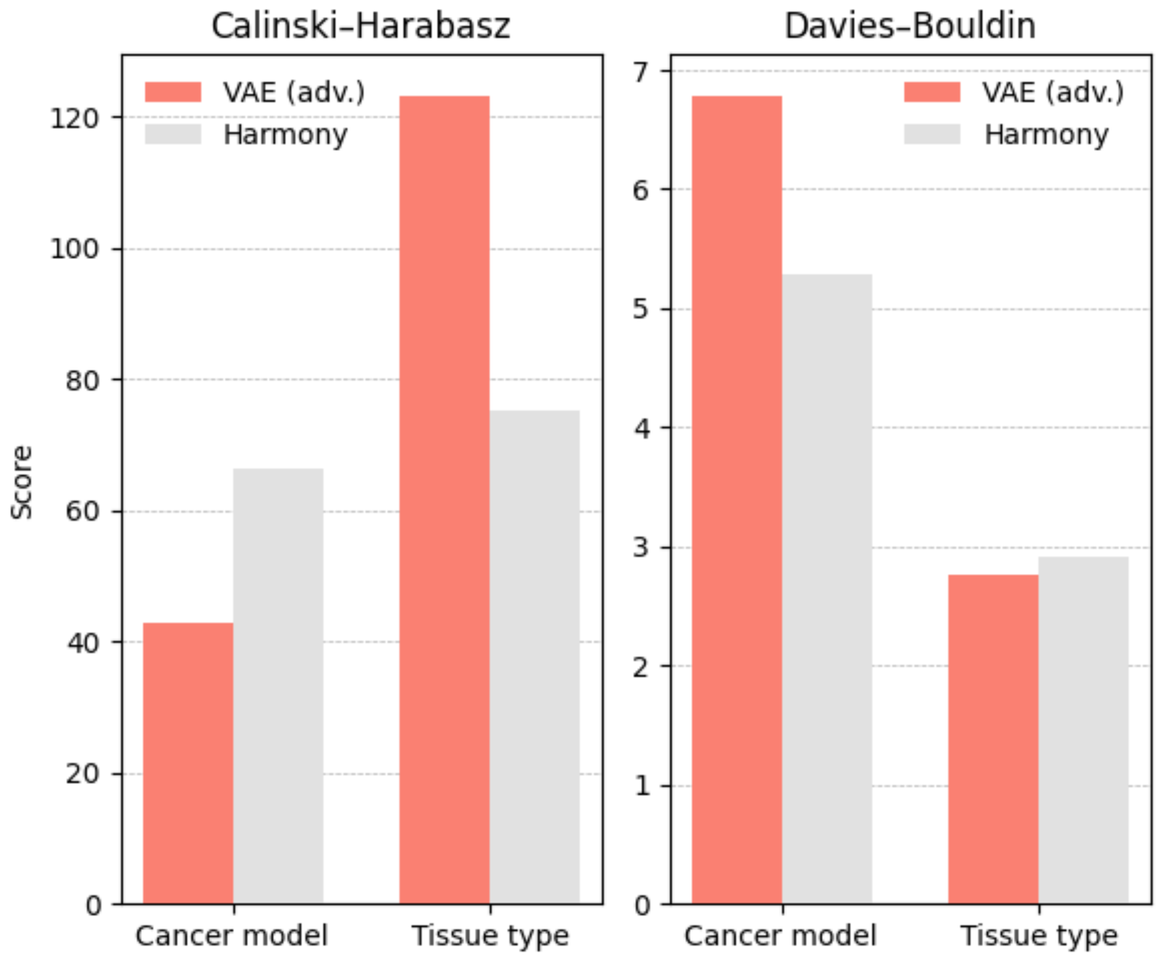


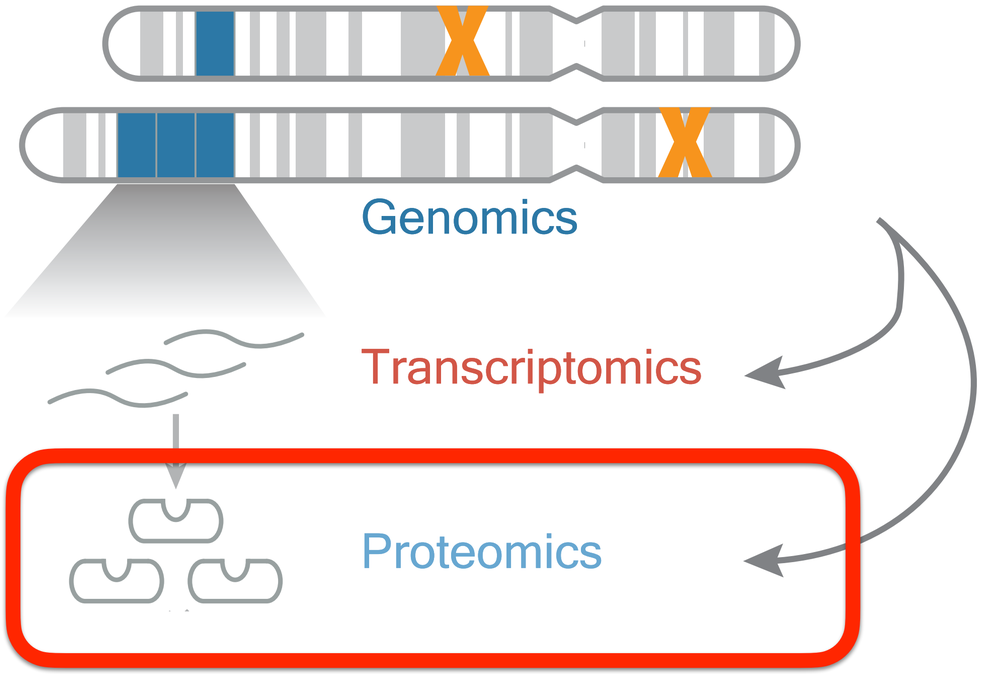
Cell lines, organoids, and tumours align



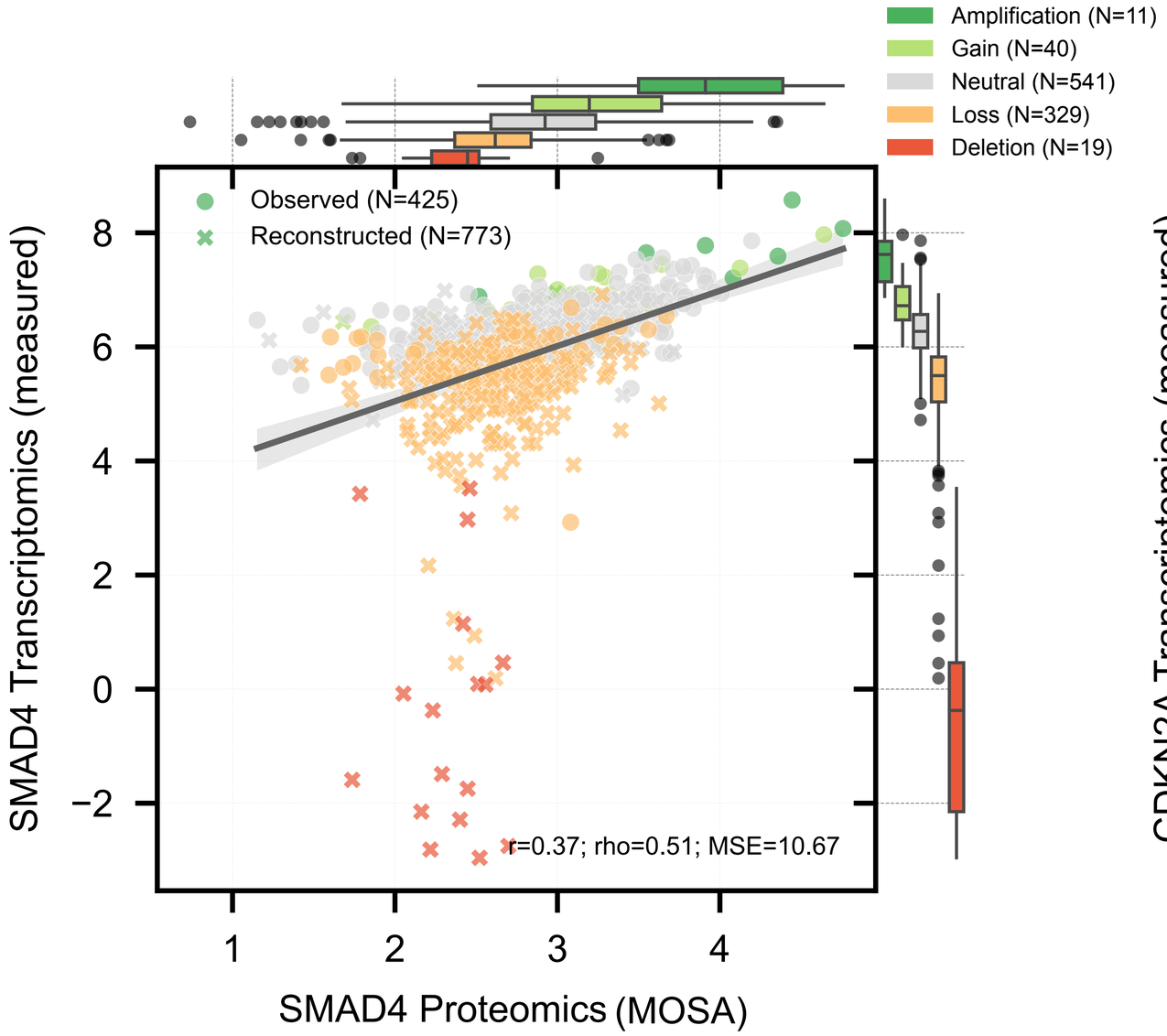
Comparison with existing approaches

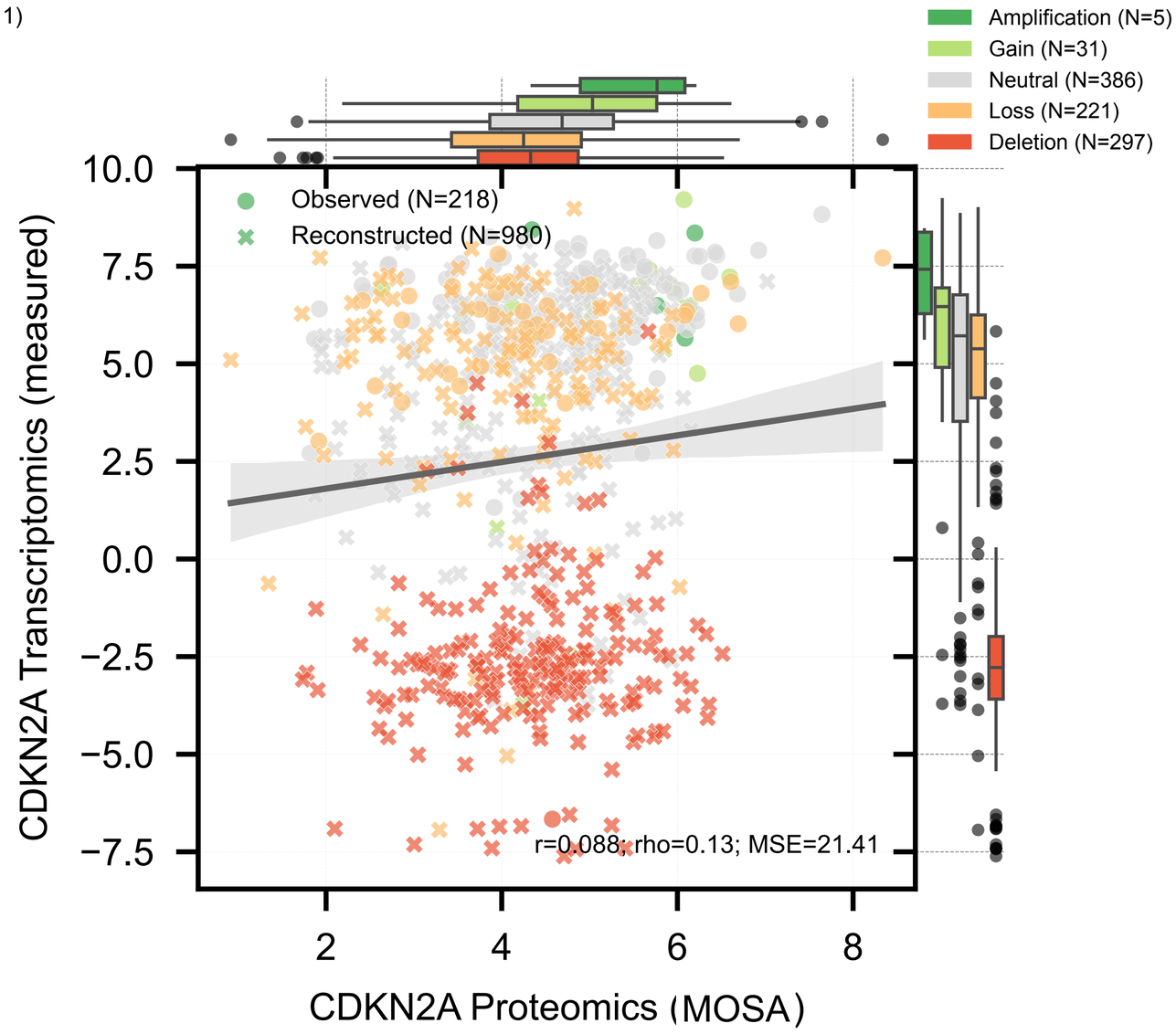





Genomics
Transcriptomics
Proteomics

Limitations - Reconstructing proteomics is more challenging





Limitations - Reconstructing proteomics is more challenging

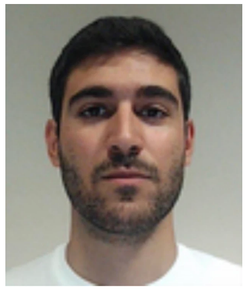
Proteomics dataset is too sparse and more data is needed to train better predictive models.

GainPro: Generative Proteomics Models for Missing Value Imputation

PyTorch implementation and wrapping of several missing value imputation methods (e.g., GAIN, MissForest, and MICE) specifically adapted and benchmarked for proteomics datasets.

Great collaboration with PRIDE (EMBL-EBI).

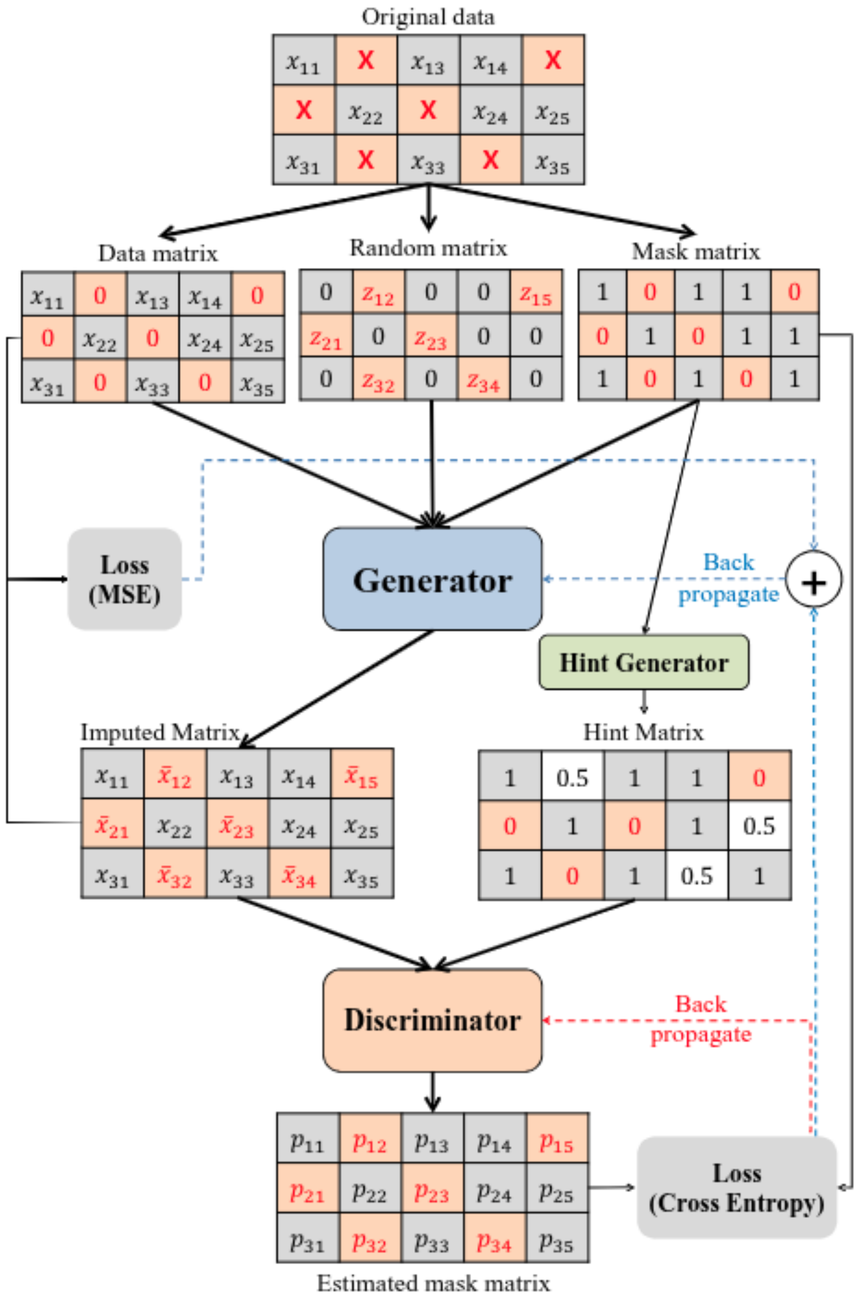
Rita Gama



Diogo Ferreira

EMBL-EBI

Yasset Perez-Riverol

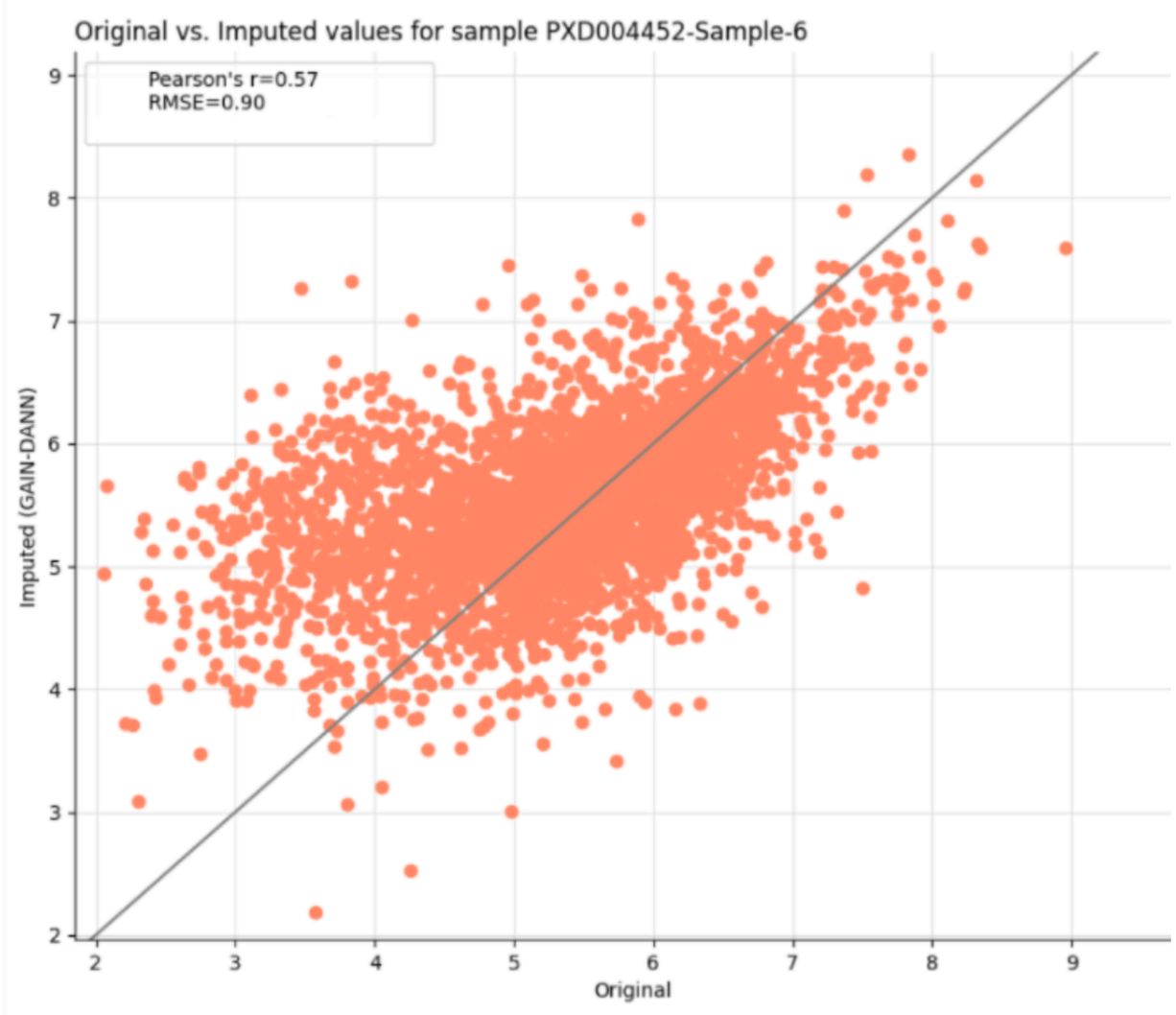


*Yoon J, Jordon J, van der Schaar M (2018) GAIN: Missing Data Imputation using Generative Adversarial Nets. arXiv.

Multiple independent proteomics datasets uniformly processed

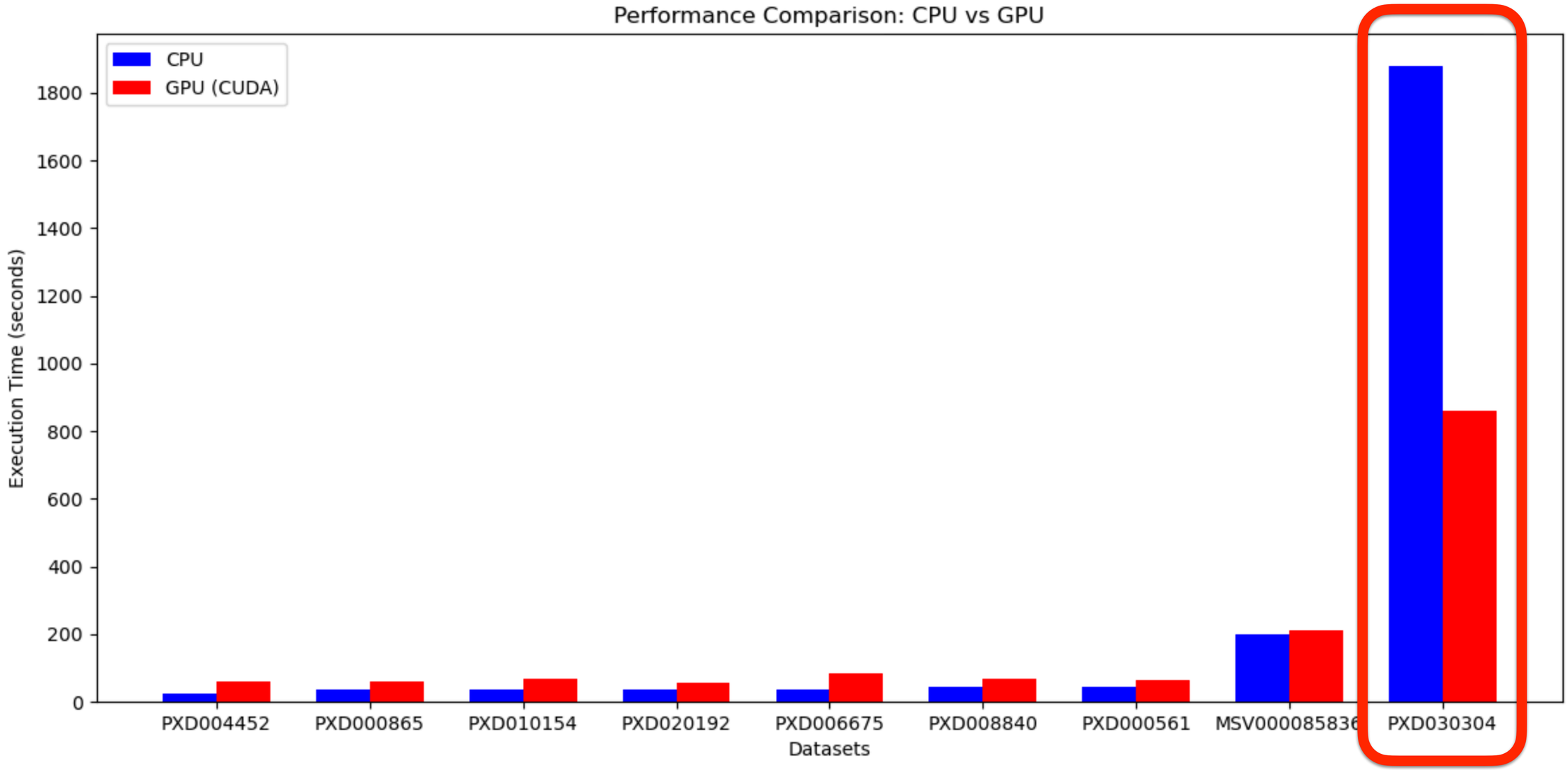
Dataset	Samples	Proteins	Missing Values (%)
PXD004452	4	8,657	17.4425
PXD000865	34	8,680	60.5116
PXD010154	38	14,669	17.1510
PXD020192	46	6,177	34.8818
PXD006675	56	11,135	30.8141
PXD008840	84	9,278	14.2370
PXD000561	85	12,025	80.4310
MSV000085836	460	13,051	29.6256
PXD030304	2012	8,754	39.9235

GAIN is accurate …

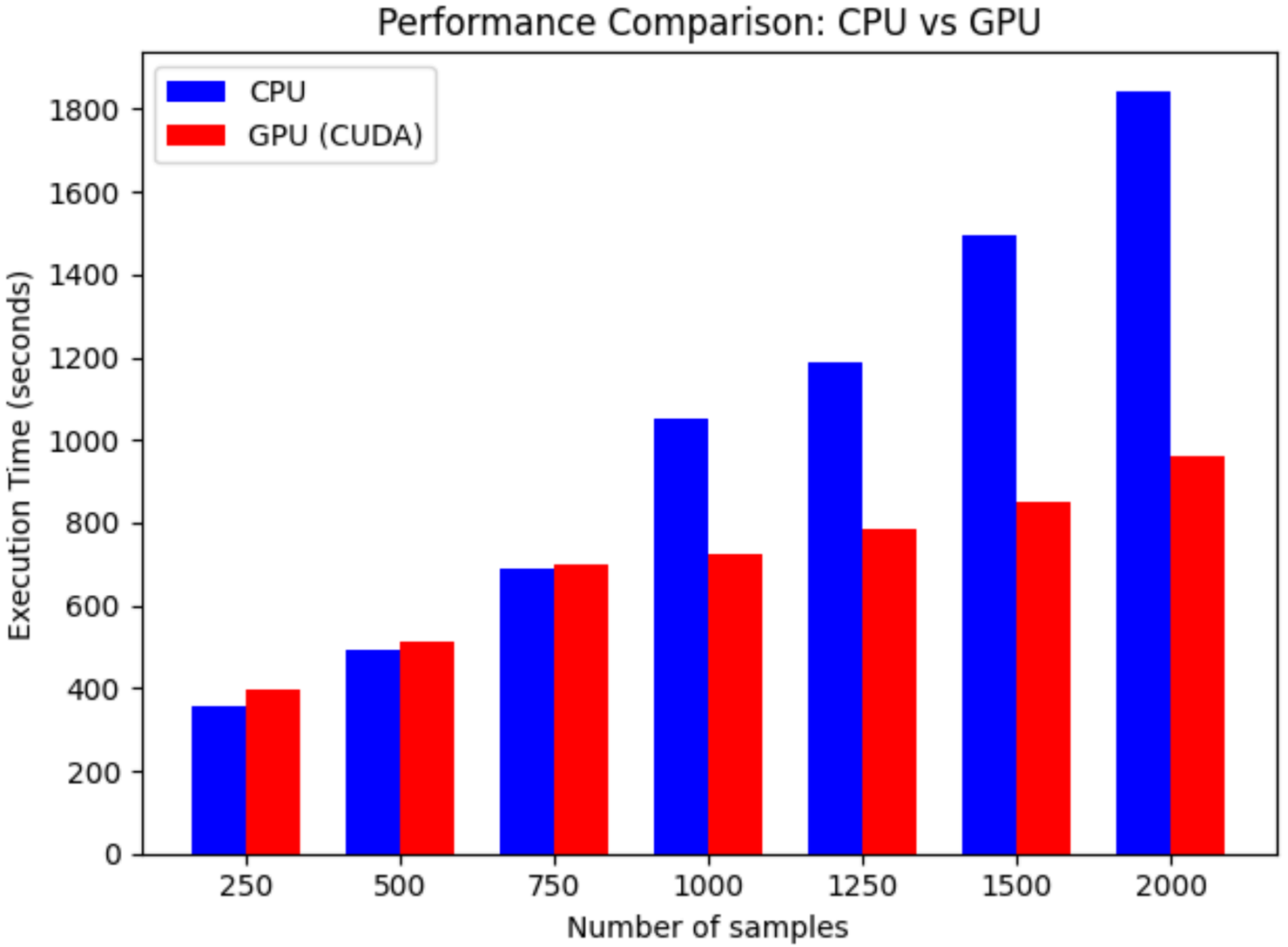


10% randomly introduced missing values (not used for training) and predicted.

GAIN is accurate and computational efficient



Downsampling of large dataset

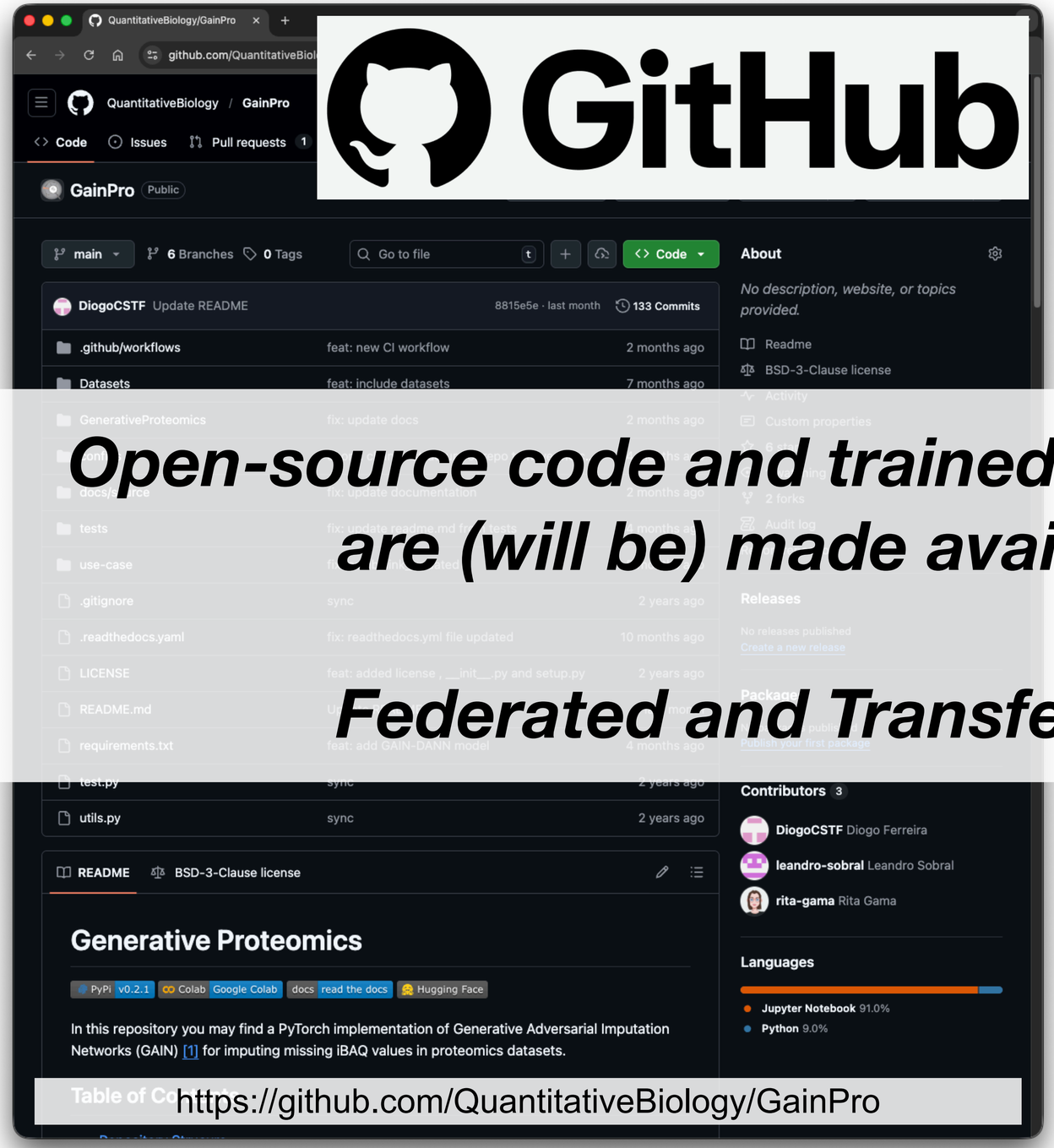


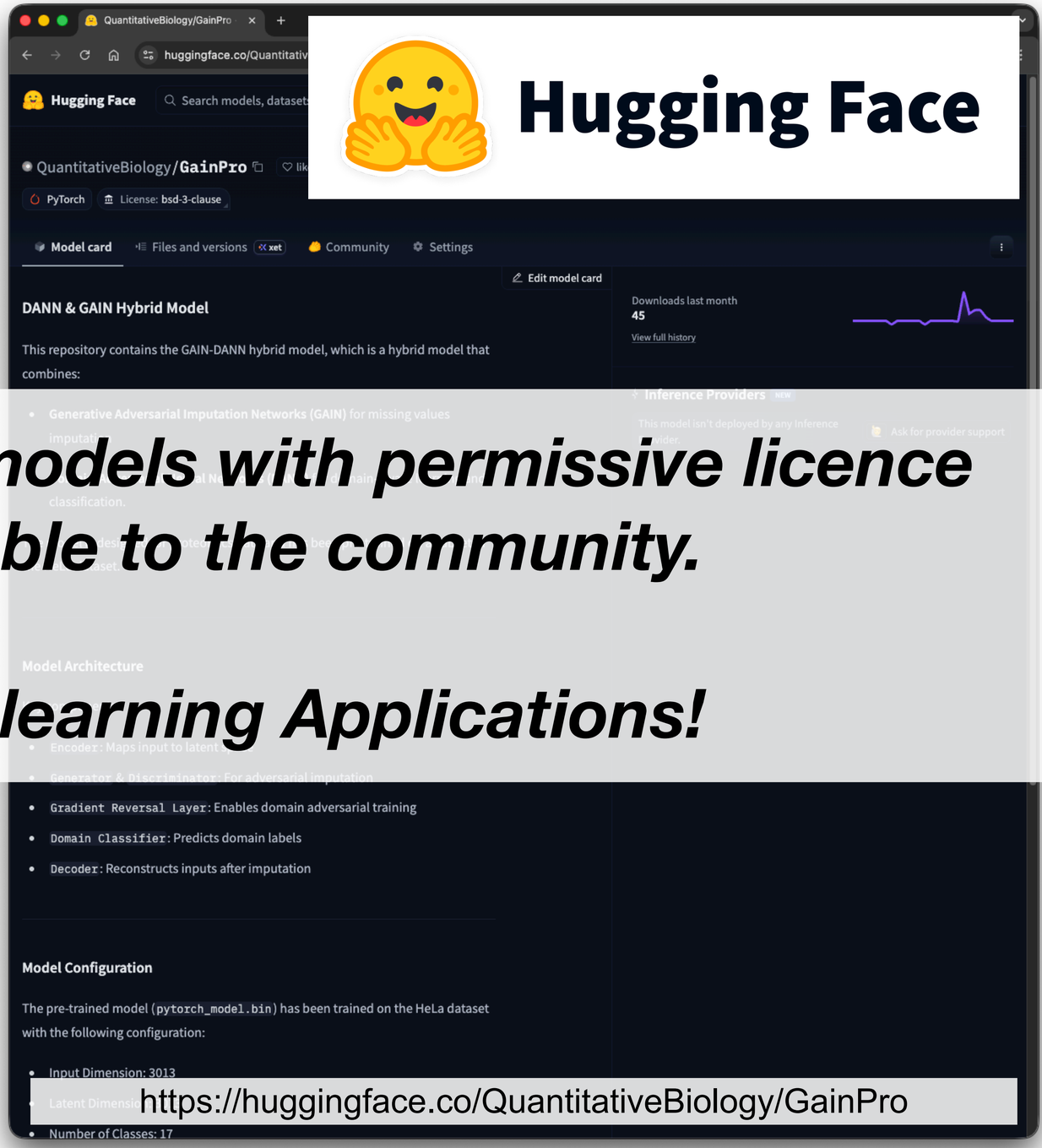
GAIN is accurate and computational efficient

	Miss Forest (10 trees)	Miss Forest (100 trees)	Miss Forest (1000 trees)	GAIN
Execution time (min)	46.15	438.07	2000*	29.16

Using the same dataset (PXD030304) with equivalent reconstruction capacity, GAIN provides >35% shorter training times.

Opens the possibility to train over large proteomics datasets (thousands of samples).





Open-source code and trained models with permissive licence are (will be) made available to the community.

Federated and Transfer learning Applications!

Conclusions and future work

- Deep learning models over a flexible and efficient way to integrate large-scale and heterogenous datasets, providing robust generative capacity.
- Opens the opportunity to develop multi-omics foundation models.

- Feature and latent representations interpretability (e.g. SHAP analysis and disentanglement learning).
- Estimate confidence on which omics/features we can robustly predict.
- Explore augmented datasets, e.g., synthetic lethal interactions.

Ana Rita Baião
André Morgado
Corinne McFarlane
Carolina Pinto
Diogo Ferreira

Gonçalo Gonçalves
Henrique Machando
Joana Correia
José Mendes
Leandro Sobral

Maria Galhardas
Miguel Dinis
Mohamed Emam
Rita Gama
Vasco Paisana

https://web.tecnico.ulisboa.pt/emanuel.v.goncalves/

Mathew Garnett
Saroor Patel
Clare Pacini
Gabriele Picco
Sanger Institute

Roger R. Reddel
Simon Cai
Qing Zhong
Phill J. Robinson
Children's Medical Research Institute

John Doench
Ganna Reint
Ellie Kaplan
Jessie Li
Berta Escude
Broad Institute

Margarida Amaral
Inês Pankonien
Cláudia Rodrigues
Universidade de Lisboa

Ana Luisa Correia
Champalimaud Foundation

Yasset Perez-Riverol
EMBL-EBI

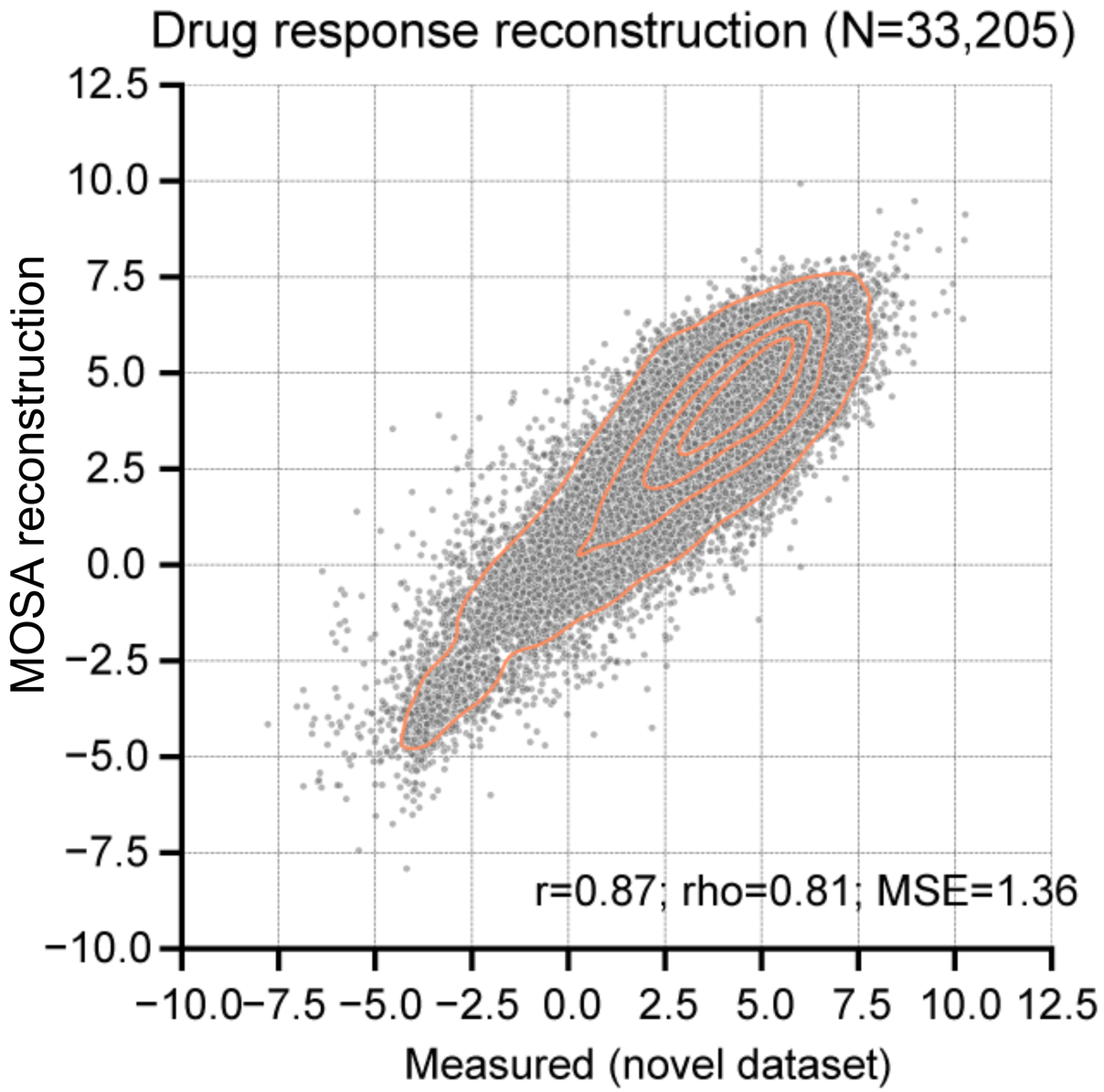
Fundação para a Ciência e a Tecnologia

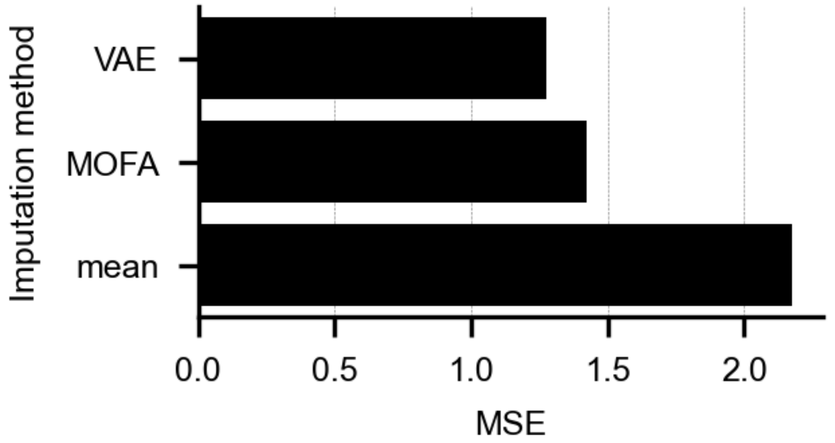
Funded by the European Union

FULBRIGHT Portugal

Fondazione per la Ricerca sulla Fibrosi Cistica - ETS
fibrosicisticaricerca.it

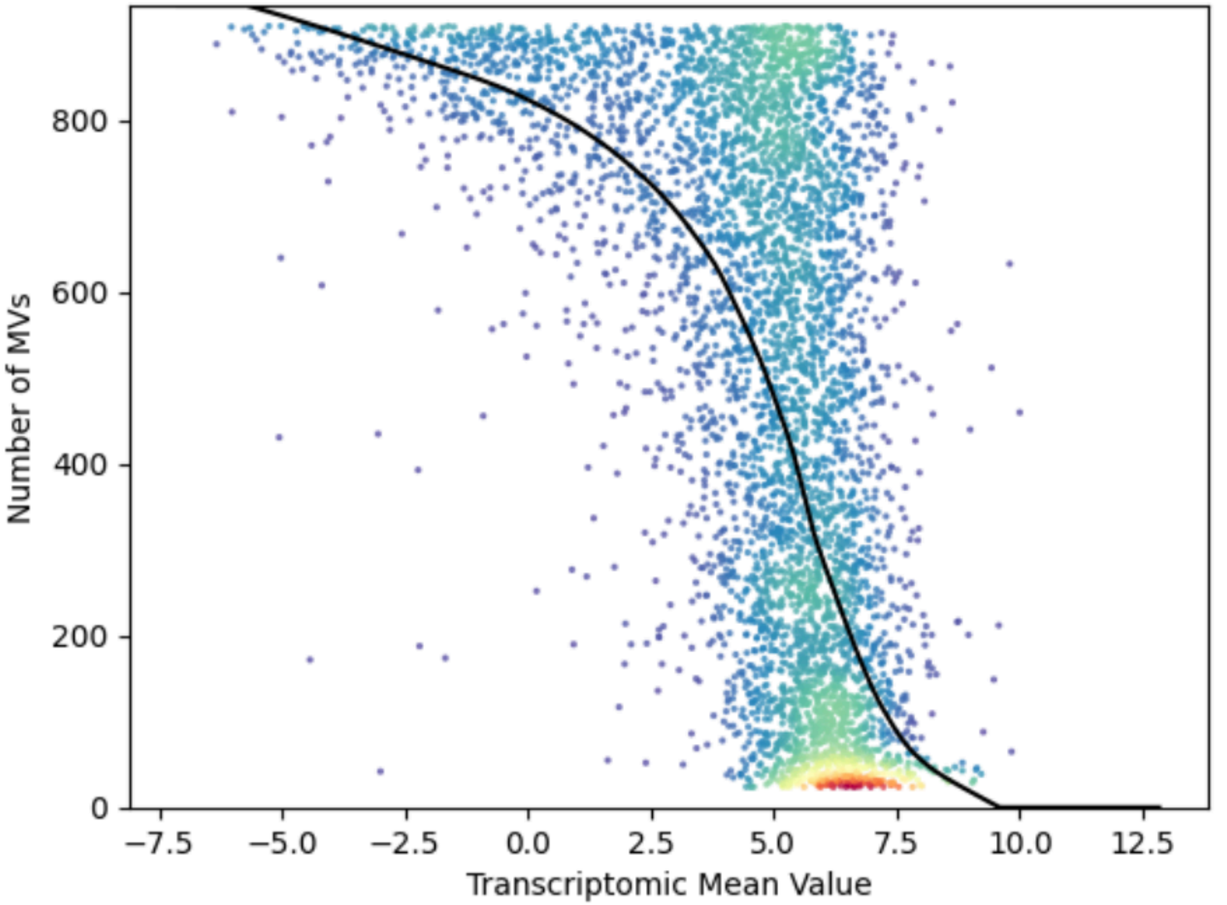
Successful reconstruction of new drug response screens



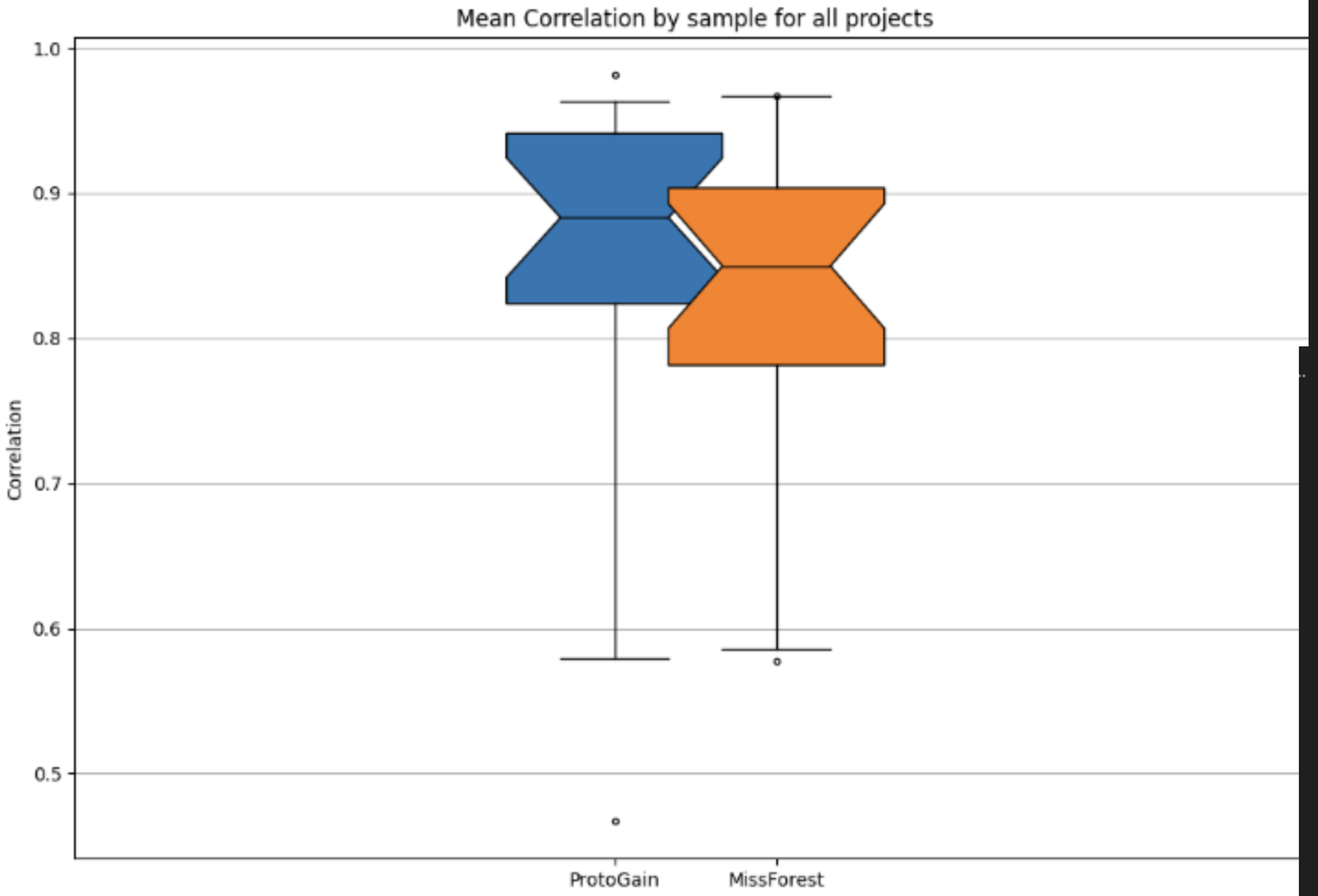


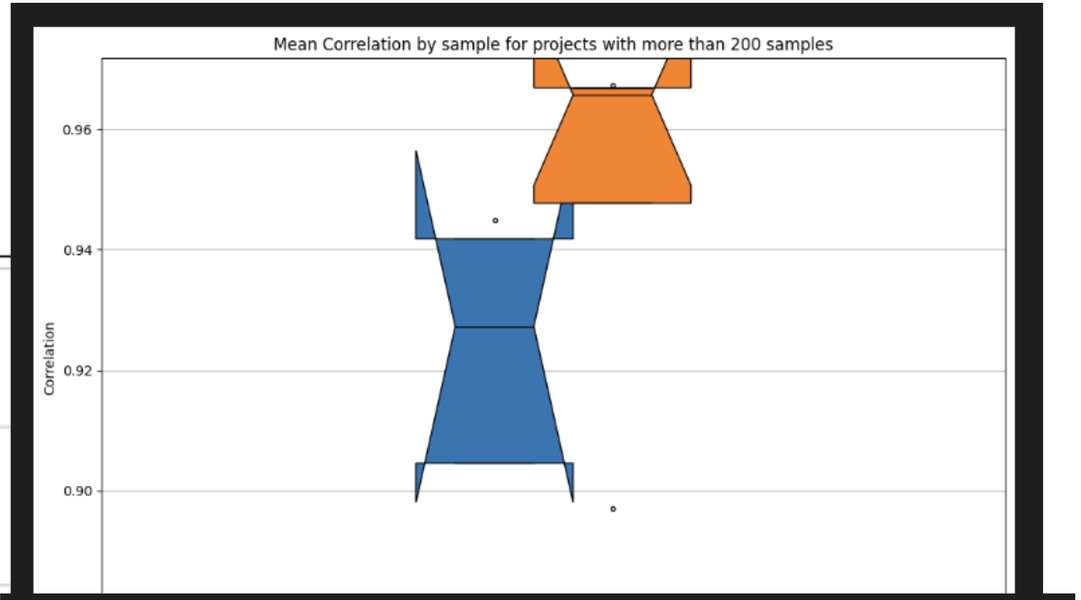
New single-agent drug screens were robustly predicted.

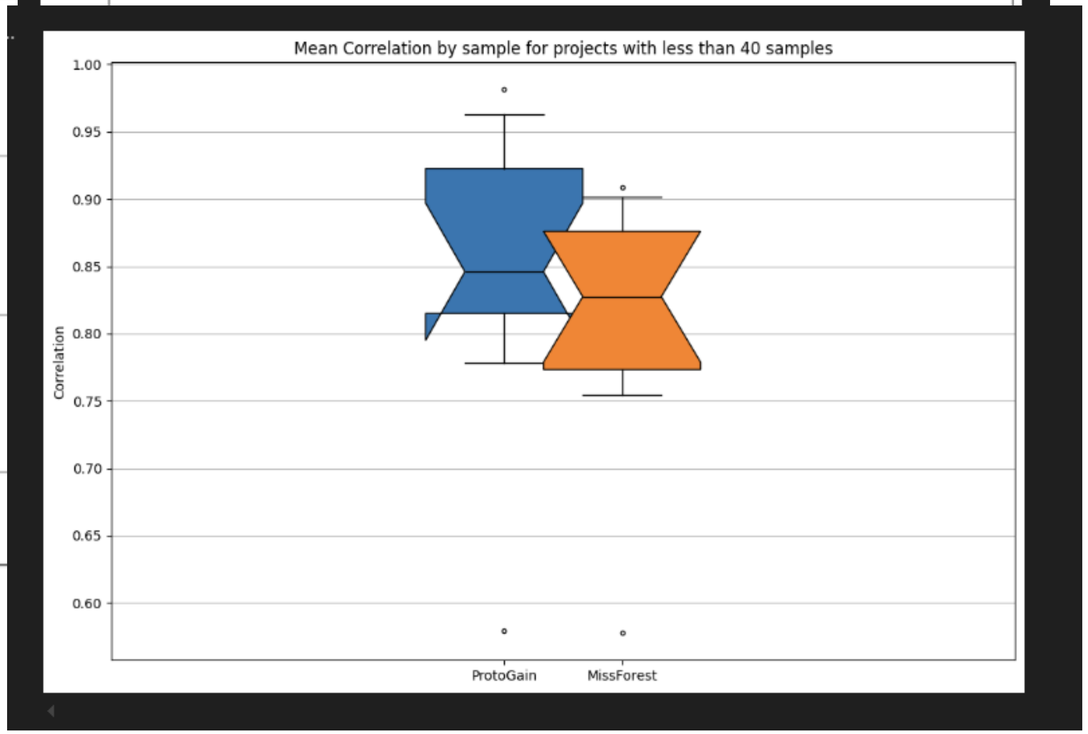
VAE model outperformed more classical linear factor model.


Number of MVs
800
600
400
200
0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
Transcriptomic Mean Value

	Mean	MF T10	MF T100	MF T1000	*Optuna* GAIN
Computational Time (min)	-	46.15	438.07	2000*	29.16
RMSE	5.61E-2	4.65E-2	4.39E-2	4.36E-2	4.72E-2


Mean Correlation by sample for all projects
Correlation
1.0
0.9
0.8
0.7
0.6
0.5
ProtoGain
MissForest


Mean Correlation by sample for projects with more than 200 samples
Correlation
0.96
0.94
0.92
0.90


Mean Correlation by sample for projects with less than 40 samples
Correlation
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
ProtoGain
MissForest

project	samples	proteins	missingrate (%)	mean_correlation_protogain	mean_correlation_mf	project_title	project_description	project_sample_description	project_data_description	project_pubmed_id	organism
PXD005736	2	8470	2,101534829	0,9631576312	0,7768586995	Proteomic Analysis of Human Cardiac Organoids	The mammalian heart undergoes major transitions	Nine hCO from either CTRL medium or MM conditions were	Peptides were analysed on a Dionex 3500RS coupled to a Q-	28916735	['Homo sapien
PXD002179	3	8836	3,380111664	0,9814026183	0,9016778847	CN-CHPP 2015 Testis Project - Tissue-Based Proteogenomics Reveals that Human Testis	The testis tissue-based trans-omics was a feasible	Extracted proteins were resolved by a regular SDS-PAGE (10%)	The raw MS/MS data was converted into MGF format and	26282447	['Homo sapien
PXD004452	4	8657	17,44253205	0,9031571496	0,7544930869	HeLa proteome of 12,250 protein-coding genes	We developed an optimized multi-shot proteomics	Protein concentration was estimated by Bradford assay	All raw LC–MS/MS data were analyzed by MaxQuant v1.5.3.6	28601559	['Homo sapien
PXD005445	7	9165	7,421089549	0,9423432215	0,9089278491	CEGS Proteomics - A multiregional proteomic survey of the postnatal human brain	Here we exploit label-free quantitative tandem mass-	Sample preparation Frozen brain samples were weighed and added	Data analysis Mass spectra were processed using MaxQuant35	29184206	['Homo sapien
PXD005445	7	9165	7,421089549	0,9423432215	0,9089278491	CEGS Proteomics - A multiregional proteomic survey of the postnatal human brain	Here we exploit label-free quantitative tandem mass-	Sample preparation Frozen brain samples were weighed and added	Data analysis Mass spectra were processed using MaxQuant35	29184206	['Homo sapien
PXD000612	11	8782	39,62340324	0,8132514899	0,8425212718	Ultra-deep human phosphoproteome reveals different regulatory nature of Tyr and Ser/Thr-	Regulatory protein phosphorylation controls	HeLa S3 cells were subjected to a double thymidine block in	Raw mass spectrometric data was analyzed in the MaxQuant	25159151	['Homo sapien
PXD019909	12	11700	18,62037037	0,7782226585	0,7696694919	Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin	Human skin provides both physical integrity and	Sample preparation for mass spectrometry Samples of skin	MS data analysis DIA raw files were analyzed using Spectronaut	33154365	['Homo sapien
PXD019909	12	11700	18,62037037	0,7782226585	0,7696694919	Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin	Human skin provides both physical integrity and	Sample preparation for mass spectrometry Samples of skin	MS data analysis DIA raw files were analyzed using Spectronaut	33154365	['Homo sapien
PXD012131	26	10188	6,769653589	0,8989361704	0,8976360077	Proteomic Atlas of the Human Brain in Alzheimer's Disease	Here we provide a proteomic resource	Tissue Extraction. Individual brains were extracted and rapidly	All raw files were searched together in the software	30735395	['homo sapien
PXD000865	34	8680	60,51165628	0,5791171964	0,5776350031	Mass spectrometry based draft of the human proteome	This PXD project contains two projects published on	Human tissue specimens were obtained from the bio bank of the	For peptide identification, tandem mass spectra were processed	24870543	['Homo sapien
PXD010154	38	14669	17,15109917	0,8452636296	0,8544015012	A deep proteome and transcriptome abundance atlas of 29 healthy human tissues	We generated a systematic, quantitative and deep	Fresh frozen human tissue samples were homogenized in	For peptide and protein identification and label free	30777892	['homo sapien
PXD020192	46	6177	34,88185485	0,826088113	0,8369621933	Proteomic Profiling of the Human Tissue and Biological Fluid Proteome	Here, we performed label-free liquid chromatography	Tissue protein extraction Fresh frozen human tissue sections	Data analysis XCalibur software v.2.0.6 (Thermo Fisher Scientific)	33107741	['Homo sapien
PXD006675	56	11135	30,81419591	0,8538012623	0,8449440789	Region and cell-type resolved quantitative proteomic map of the human heart and its	The heart is a central human organ and its diseases are the	All 16 heart regions and cell types were processed using in	Tandem mass spectra were searched against the 2015 Uniprot	29133944	['Homo sapien
PXD008840	84	9278	14,23708414	0,8890134105	0,8752975968	A proteomic landscape of diffuse-type gastric cancer	Here we present a dataset from 84 DGC patients with	Specimens in dry ice were transferred to Beijing Proteome	Raw files were searched against the human National Center for	29739932	['Homo sapien
PXD000561	85	12025	80,43106274	0,4670283317	0,5857712758	A draft map of the human proteome	The availability of human genome sequence has	17 adult tissues, 7 fetal tissues, and 6 hematopoietic cell types	Mass spectrometry data obtained from all LC-MS analysis were	24669763	['Homo sapien
MSV000085836	460	13051	29,62563255	0,9137362382	0,9667231899						['Homo sapien
PXD030304	2012	8754	39,92359528	0,9297441695	0,9119672466	Pan-cancer proteomic map of 949 human cell lines (ProCan-DepMapSanger)	This pan-cancer cell line proteomic atlas comprises	The cell lines were each prepared using Accelerated Barocycler	DIA-MS data in wiff file format were collected for 6,981 MS runs,	35839778	['Homo sapien