l'Intelligence Artificielle pour comprendre les cellules?

Jérémie Kalfon
Gabriel Peyré
Laura Cantini

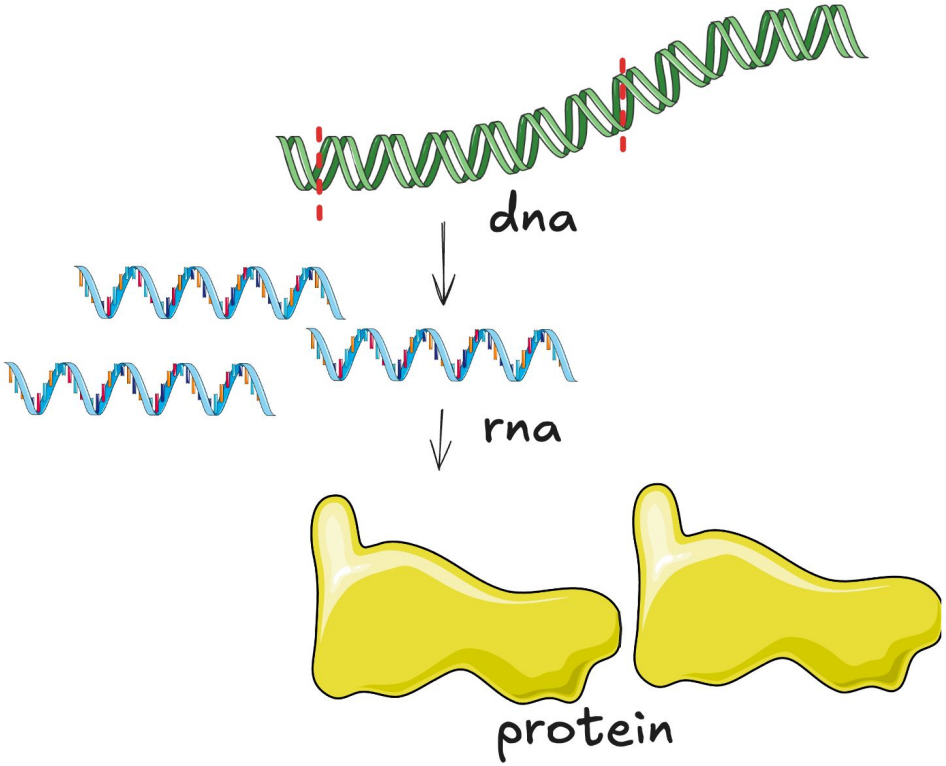
INSTITUT Pasteur

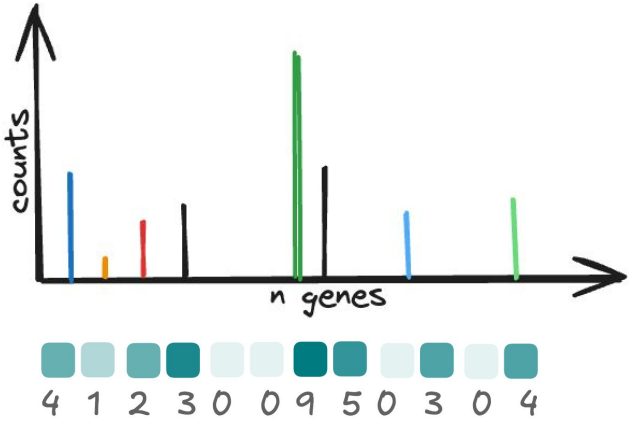
ENS PSL

Introduction

1.1. Biologie cellulaire 101: le dogme

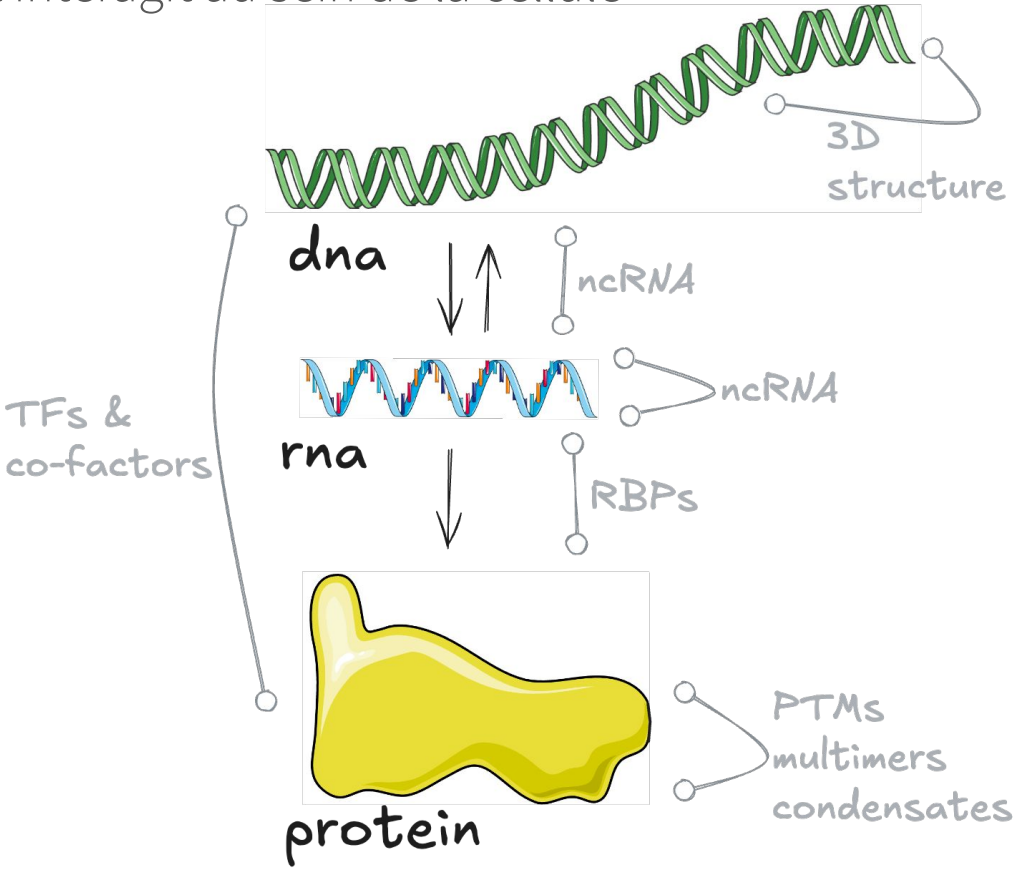
Mesurer l'ARN par cellule unique nous permet de comprendre ce que fait la cellule





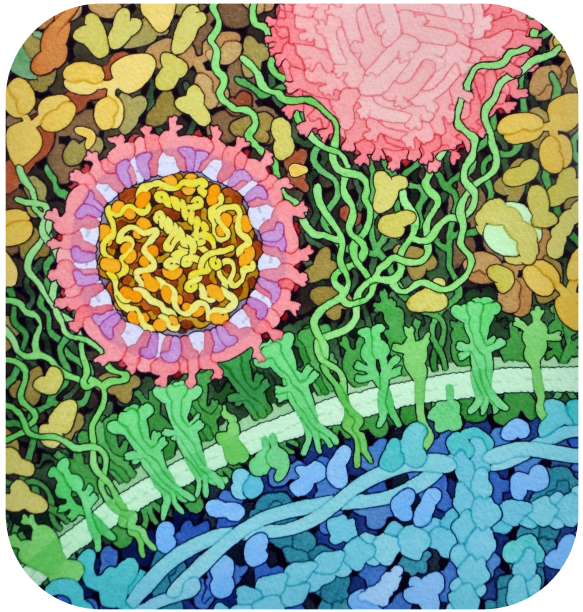
1.2. Du dogme à la réalité

Quasiment tout interagit au sein de la cellule

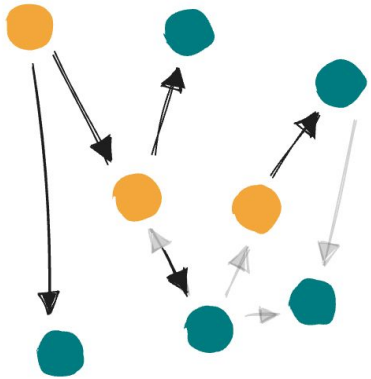


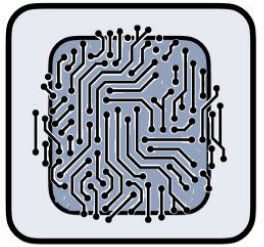
1.3. Réseau génétique et représentation de la cellule

Les cellules sont faits des complexes de trilliards de macromolécules



1





ML

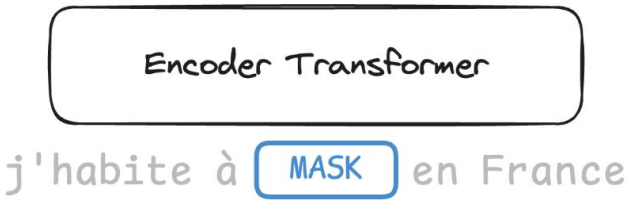
1: Gene regulatory network inference in the era of single-cell multi-omics | Nature Reviews Genetics

Dans le domaine du Machine learning…

Les LLMs ont présentés de nouvelles idées pour apprendre des intéractions sans supervision

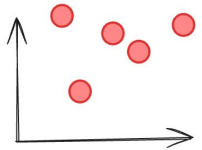
1.4. Les transformer sont pré-entraînés avec du **SSL**

les méthodes classiques d'apprentissage **s**emi-**s**upervisé



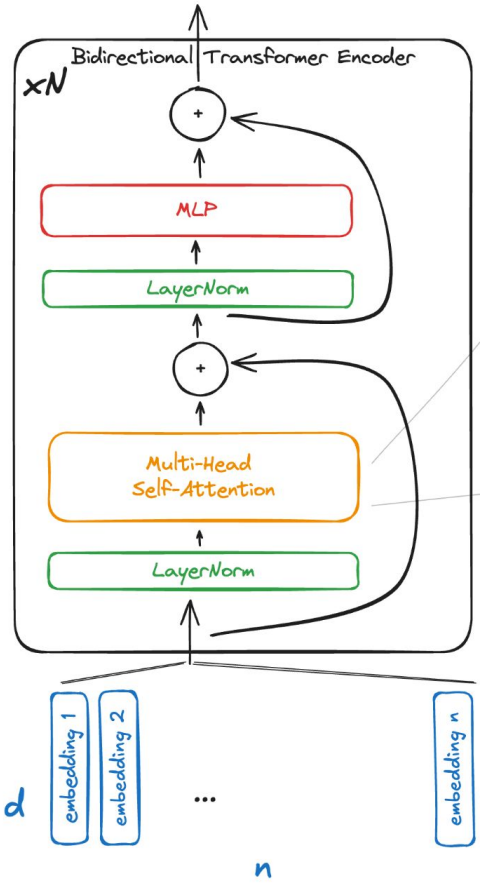
Bidirectional / Masking

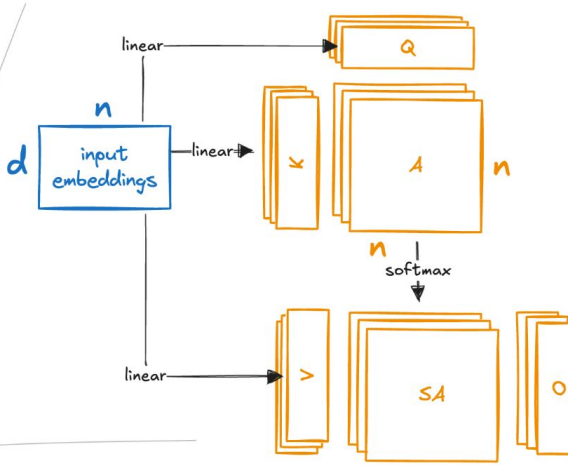
Decoding / Autoregressive

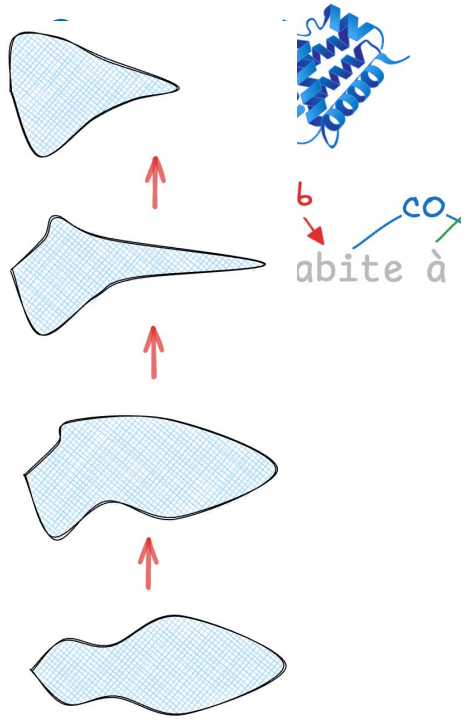


[1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In-context Learning and Induction Head

1.5. Transformer utilisent le mécanisme d'**Attention**



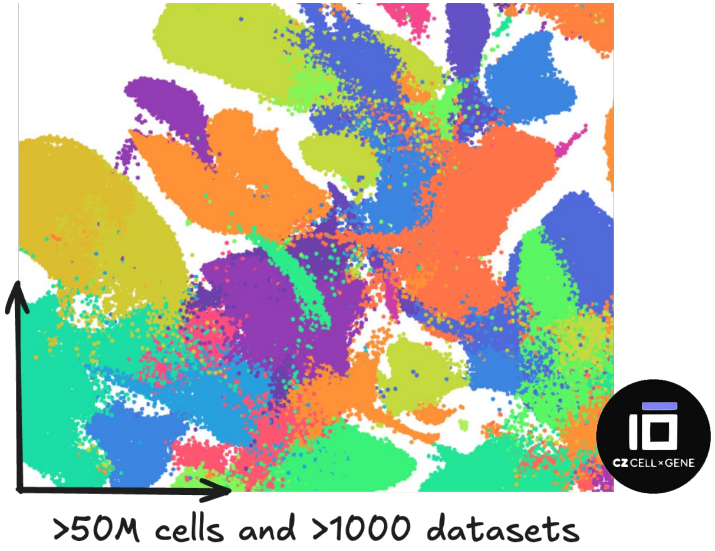


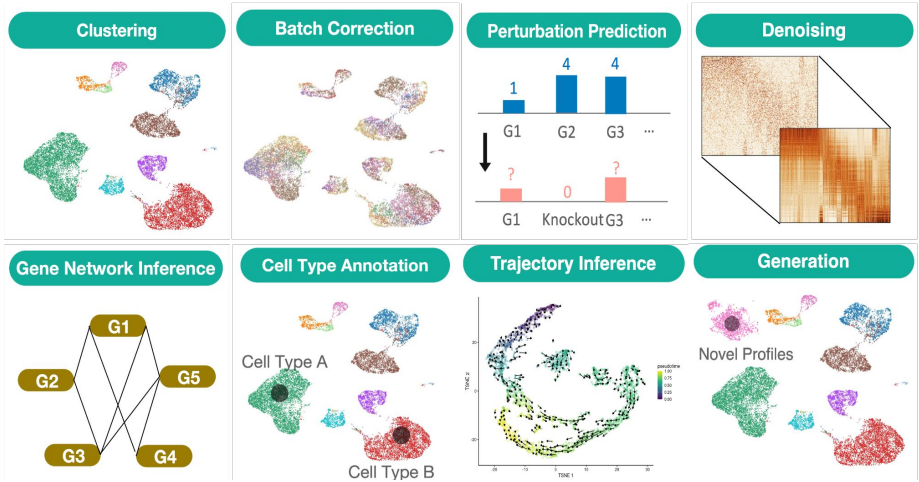


[1706.03762] Attention Is All You Need, Accurate structure prediction of biomolecular interactions with AlphaFold 3 | Nature

1.6. Geneformer et scGPT: single-cell foundation models

premiers résultats en entraînant des LLMs sur du scRNAseq (scFMs)

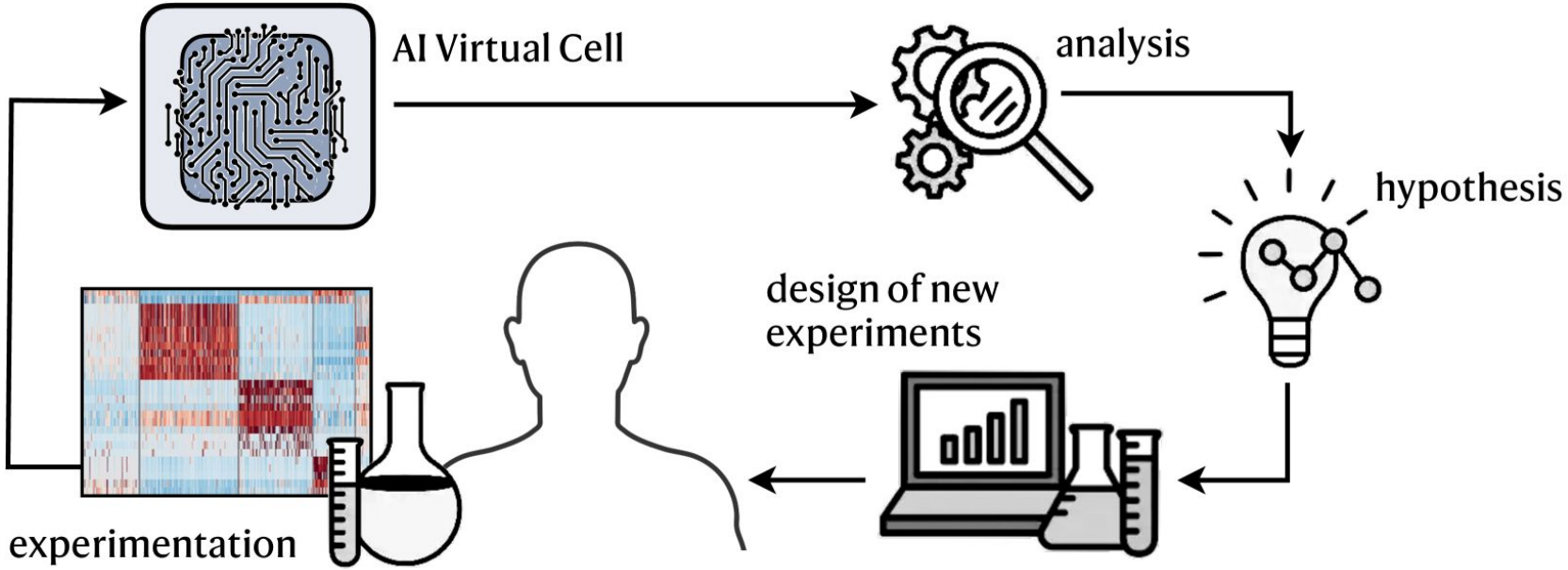




scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods

Pourquoi les modèles de fondation

Modéliser la cellule peut nous permettre de découvrir des médicaments



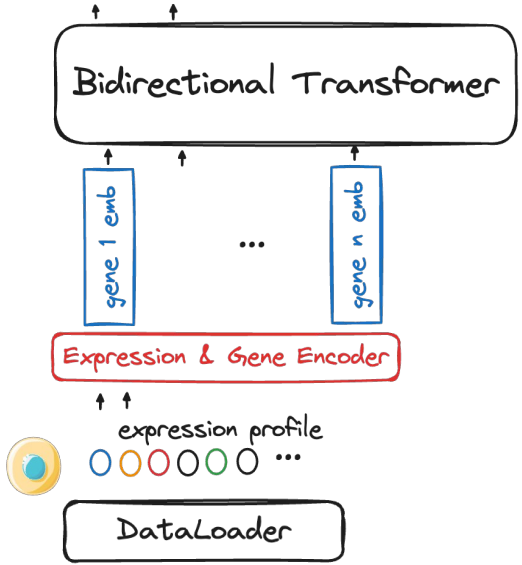
How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities

Qu’est-ce que les scFMs ont appris de la cellule?

Peut-on améliorer les scFMs en utilisant le framework des réseaux de gènes?

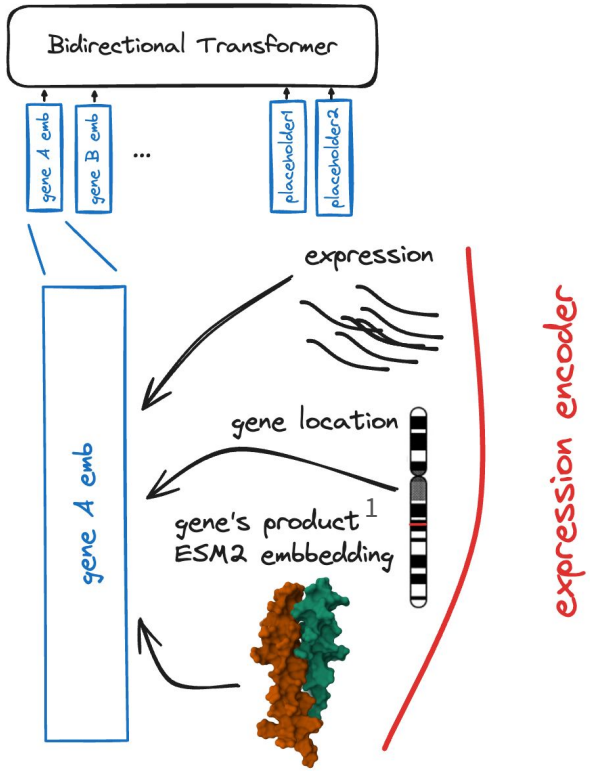
2. scPRINT & scPRINT-2

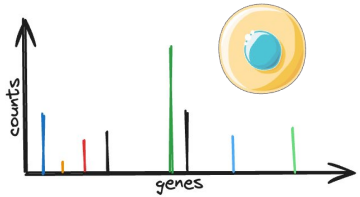
2.1. scPRINT et ses nombreuses contributions



2.2. Encoder les gènes via leur séquence

ajouter des biais inductif pour mieux apprendre les intéractions génétiques

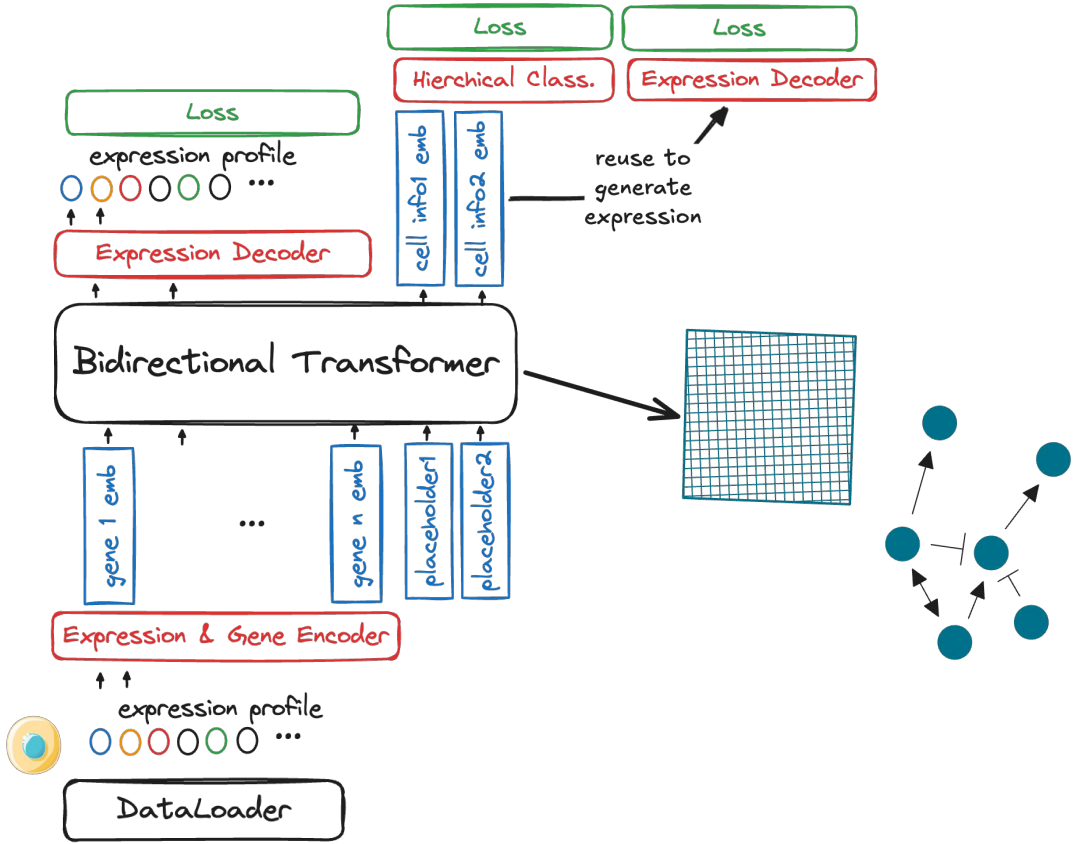




1: Universal Cell Embeddings: A Foundation Model for Cell Biology | bioRxiv

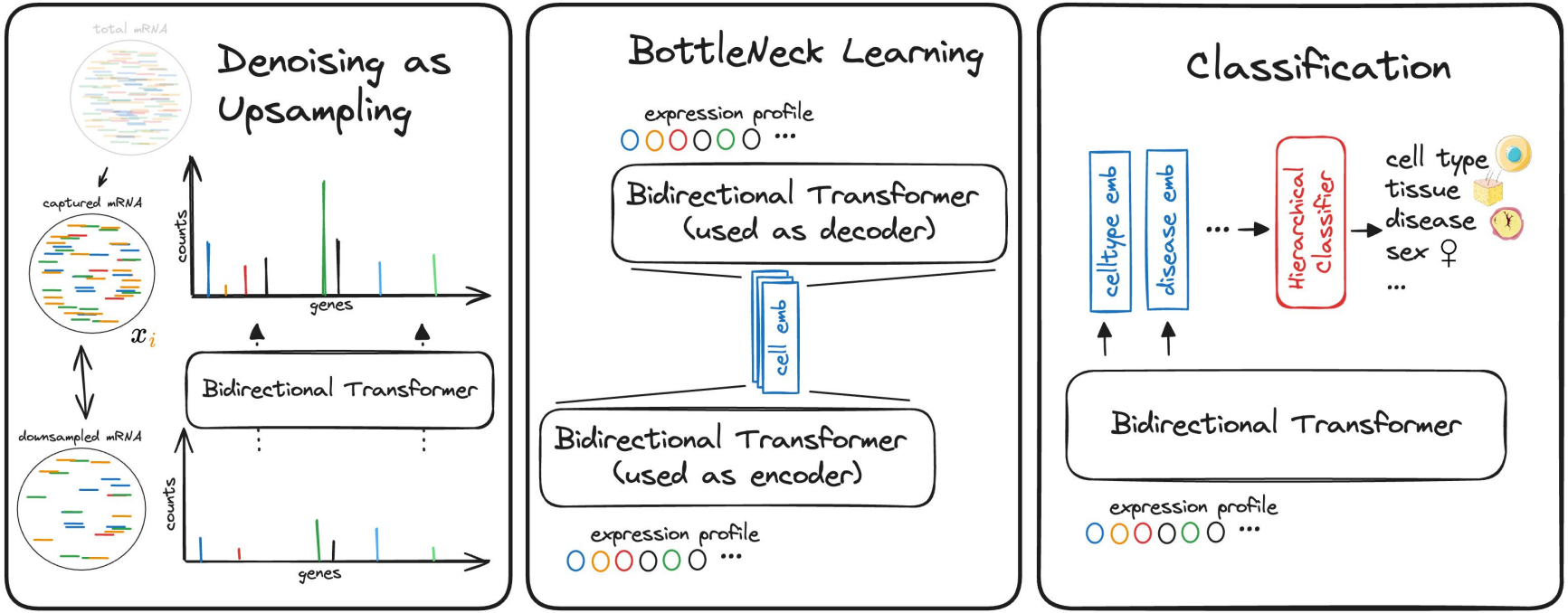
2: Evolutionary-scale prediction of atomic-level protein structure with a language model | Science

2.1. scPRINT et ses nombreuses contributions



2.3. scPRINT et ses **nouvelles** méthodes d'étrainement

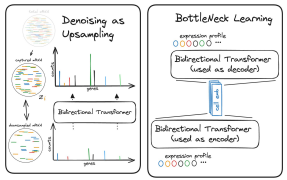
scPRINT spécialisés sont entrainement aux données de scRNA-seq

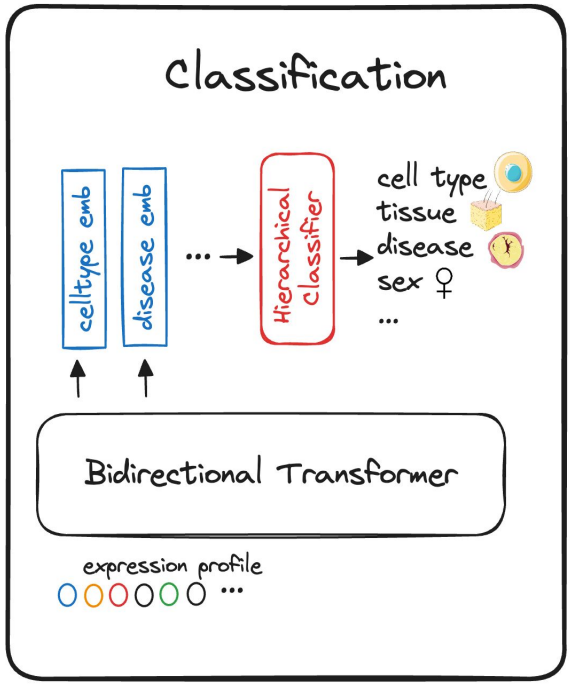


Assessing the limits of zero-shot foundation models in single-cell biology

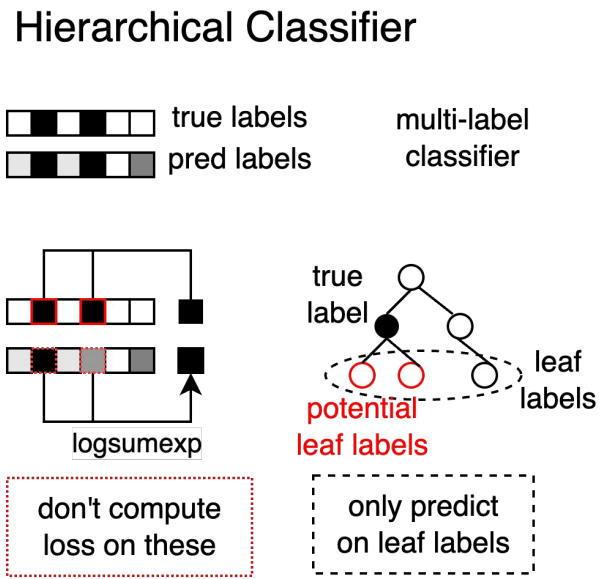
2.4. scPRINT et ses **nouvelles** méthodes d'entraînement

La classification pour déconvoluer les sources de variations d'expression



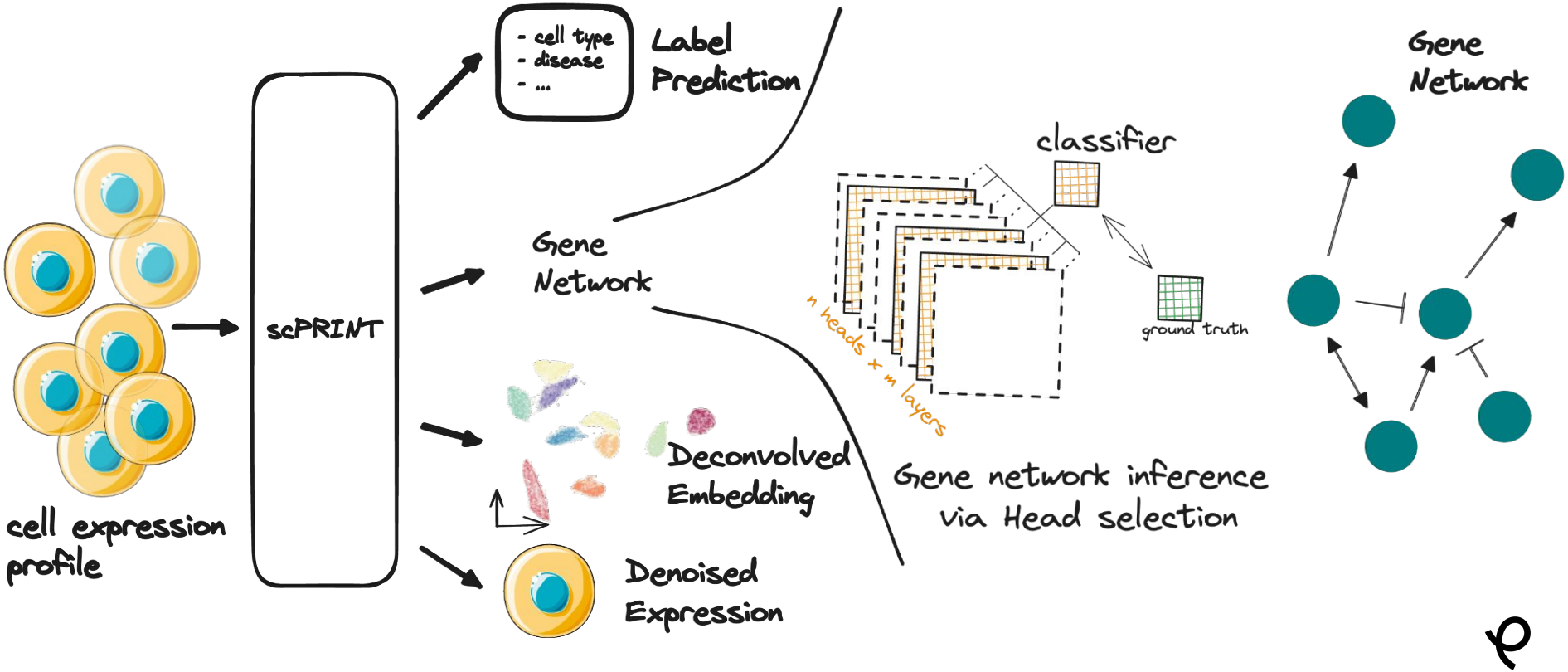


neuron
↓
sensory neuron
↓
olfactory neuron
↓
mitral neuron



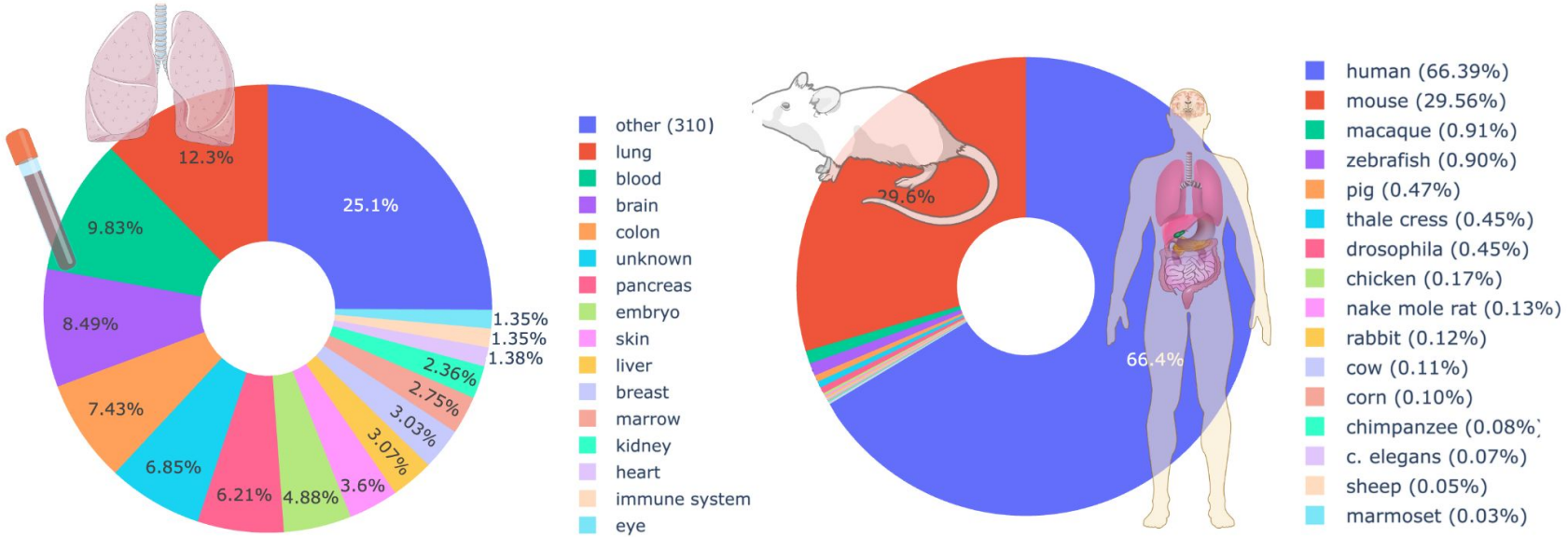
2.5. un scFM avec des capacités diverses

scPRINT does denoising, predicts cell labels, cell embeddings, and gene networks



2.6. Le corpus scPRINT: 350 millions de cellules

La plus grande base de donnée d'expression génétique par cellule seule

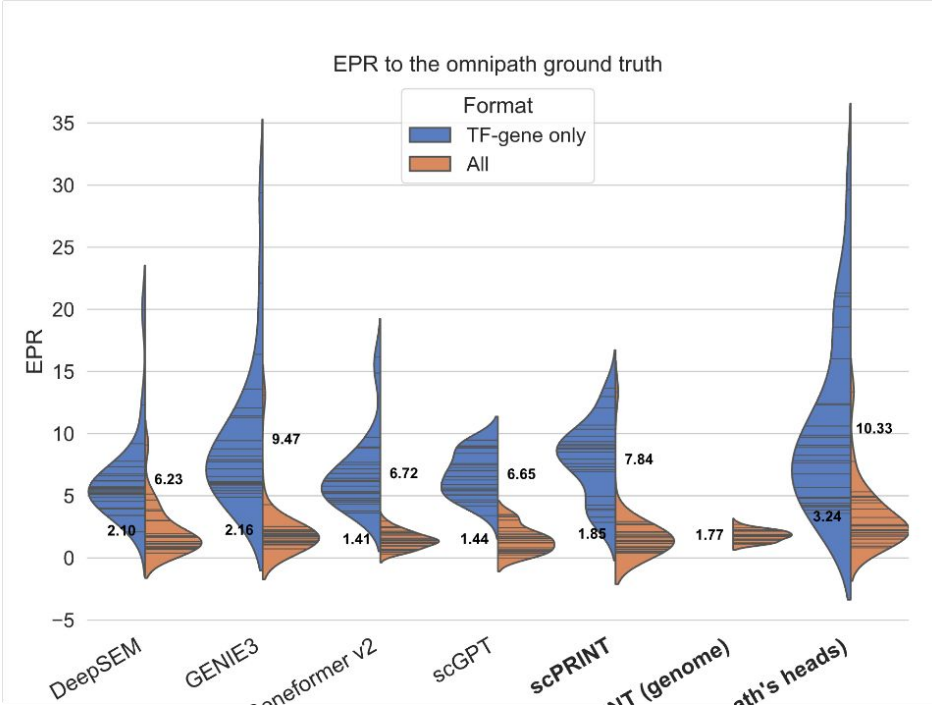


scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

3. Benchmarks et Resultats

3.1. LCMs ont une représentation des interactions

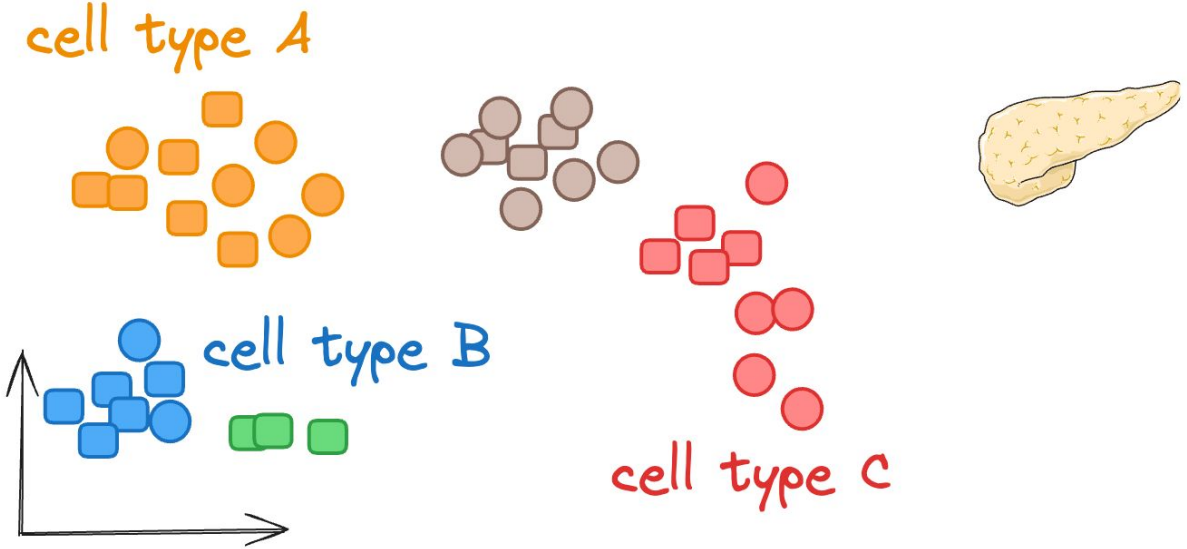
Qui devient meilleurs grâce à nos biais inductifs



1: https://r.omnipathdb.org/

3.2. Prédire les types cellulaires sur de nouvelles données

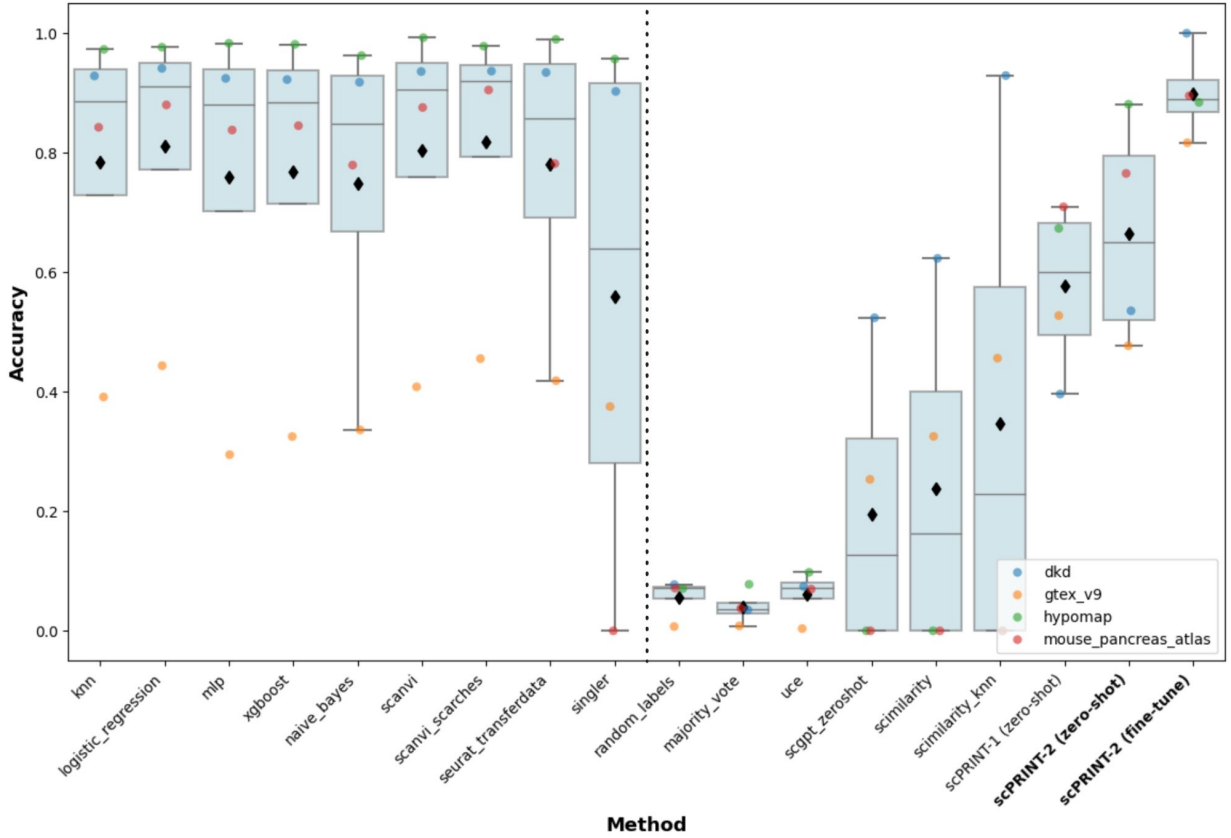
scPRINT prédit: cell type, sex, sequencer, disease, ethnicity sans ré-entraînement



https://openproblems.bio/results/label_projection/

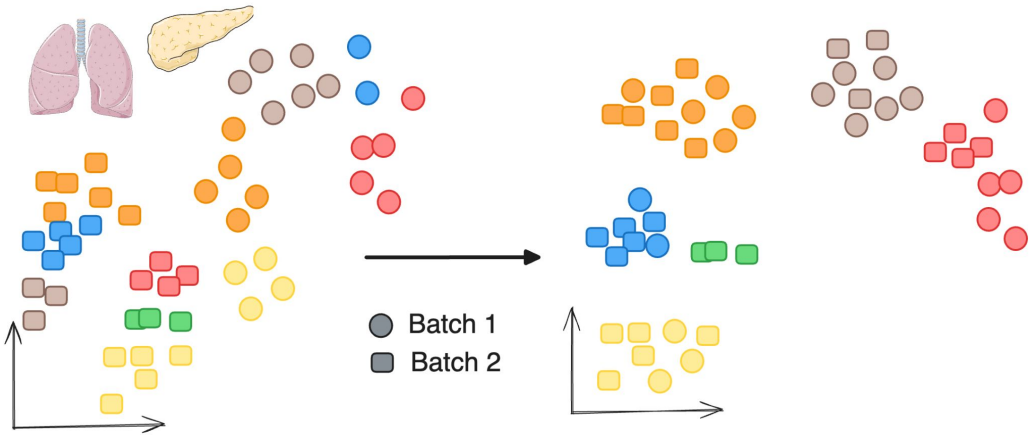
3.2. Les modèles de fondation sont SOTA en classification

Quand mesurés et entraînés correctements



3.3. scPRINT peut retirer les effets de batch

scPRINT learned to disentangle biological signal from sequencing signal

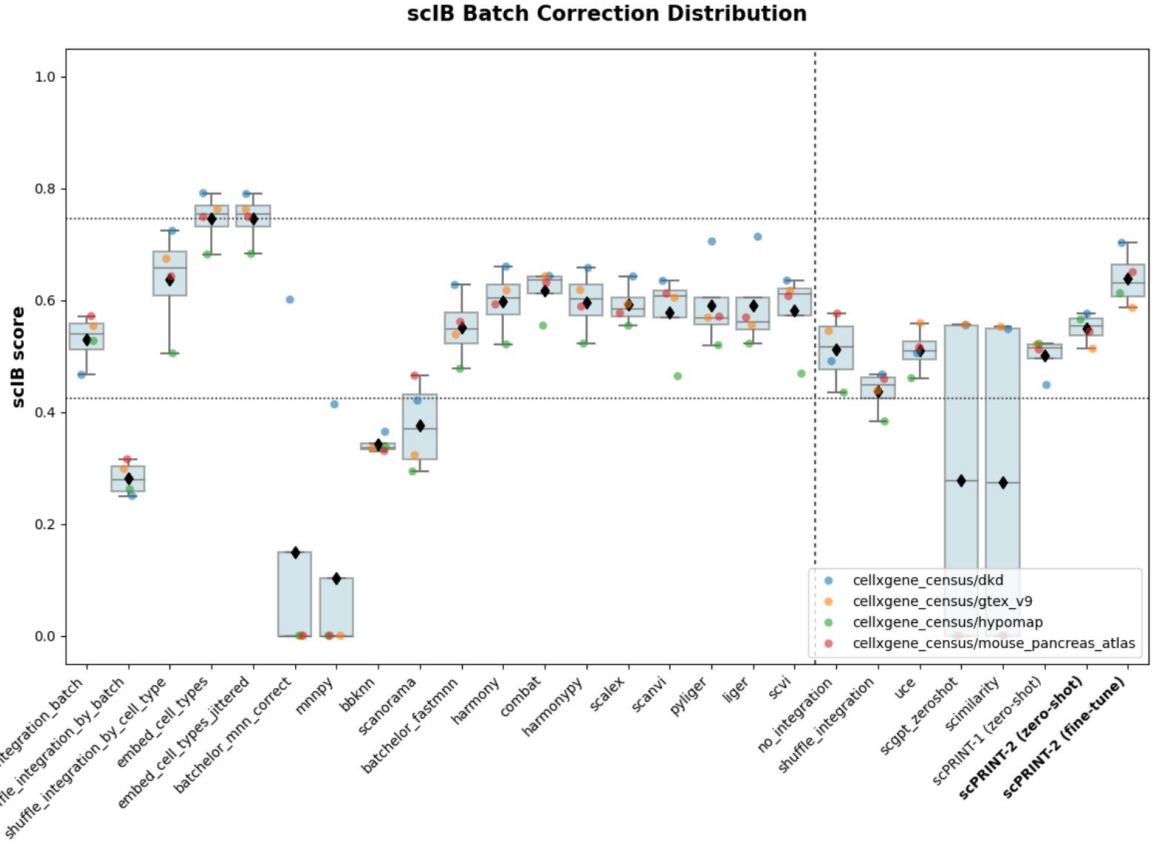



```
pip install scib
```

Benchmarking atlas-level data integration in single-cell genomics | Nature Methods

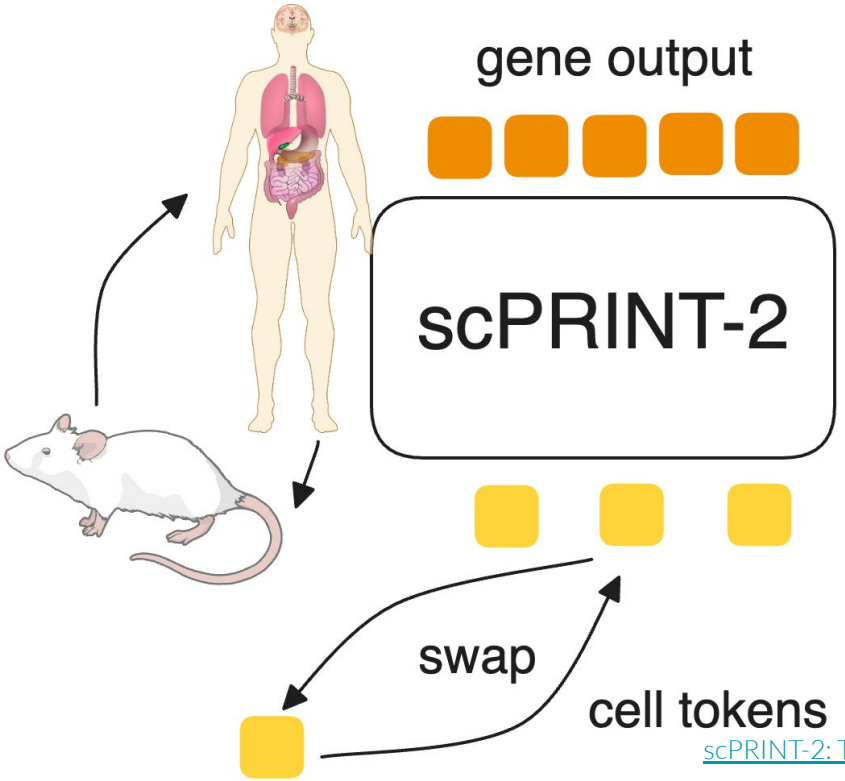
3.3. Batch: scPRINT-2 beats scPRINT-1 and becomes SOTA

thanks to these specialized embeddings (increase scIB batch corr)



3.4. Les LCMs peuvent généraliser à de nouvelles tâches

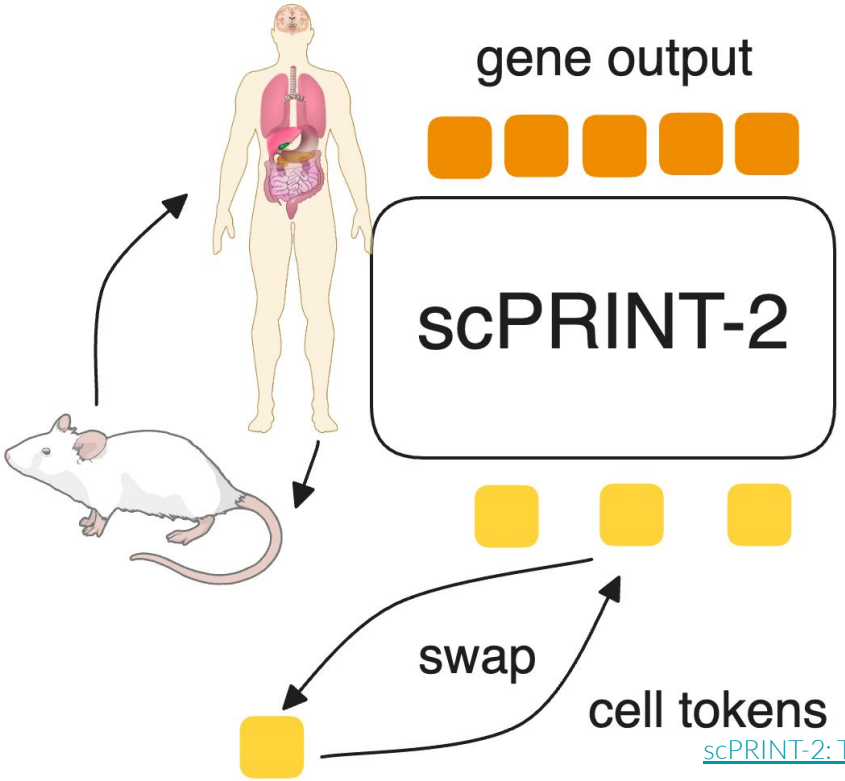
Example: Humaniser les données précliniques de souris

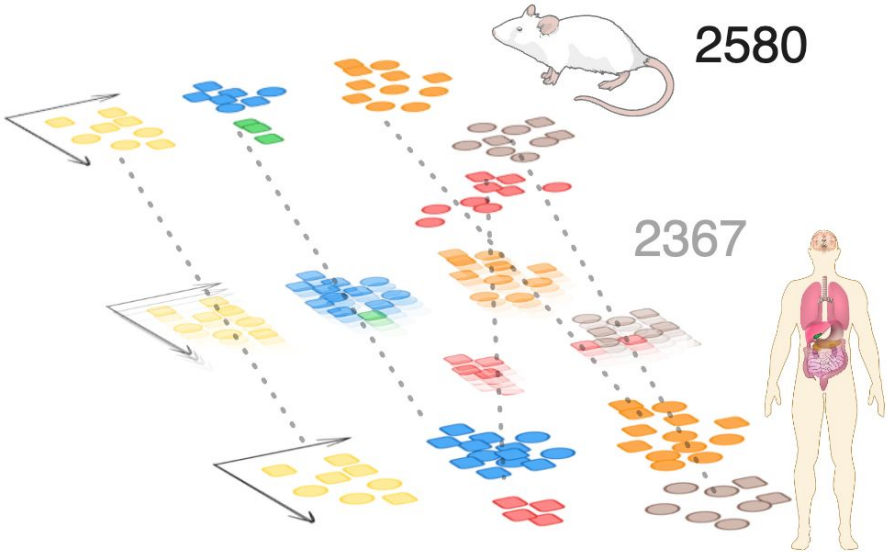


scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

3.4. Les LCMs peuvent généraliser à de nouvelles tâches

Example: Humaniser les données précliniques de souris

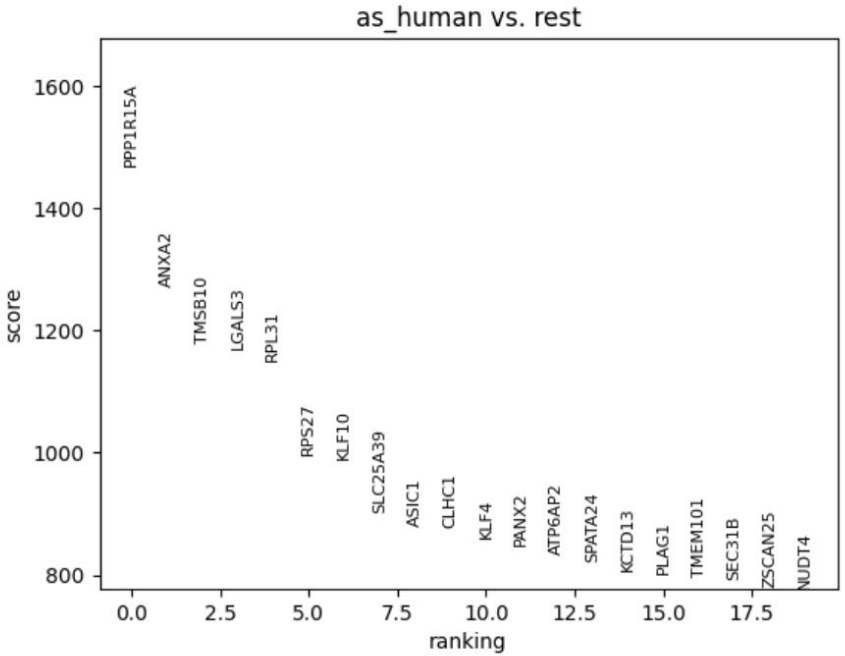




scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

3.4. Confirmé sur de l'analyse d'expression différentielle

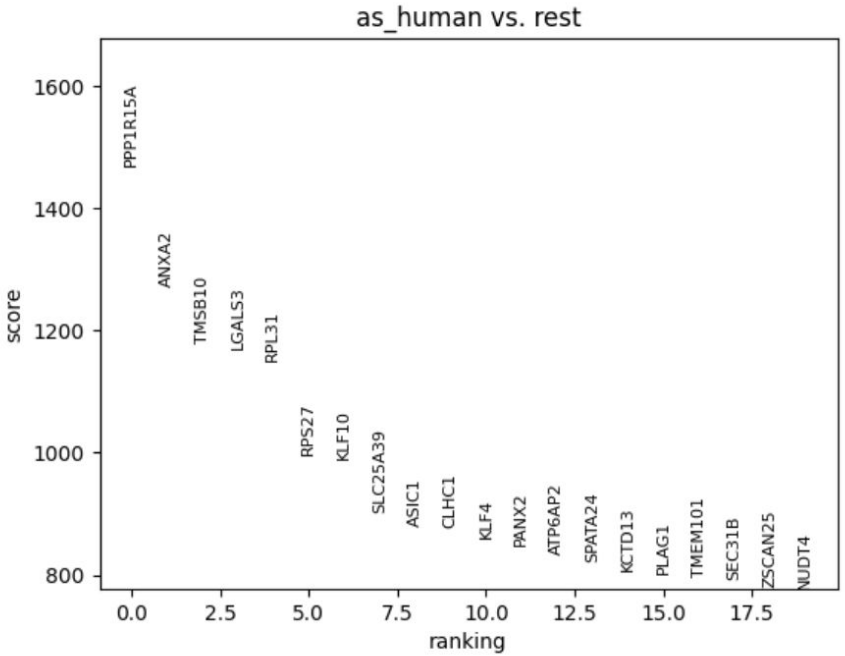
… top differentially expressed genes are the correct ones!

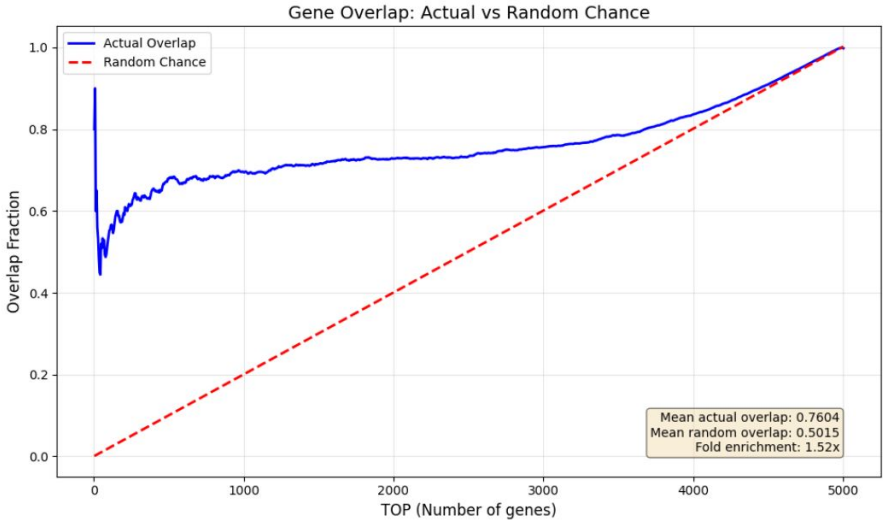


scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

3.4. Confirmé sur de l'analyse d'expression différentielle

… top differentially expressed genes are the correct ones!

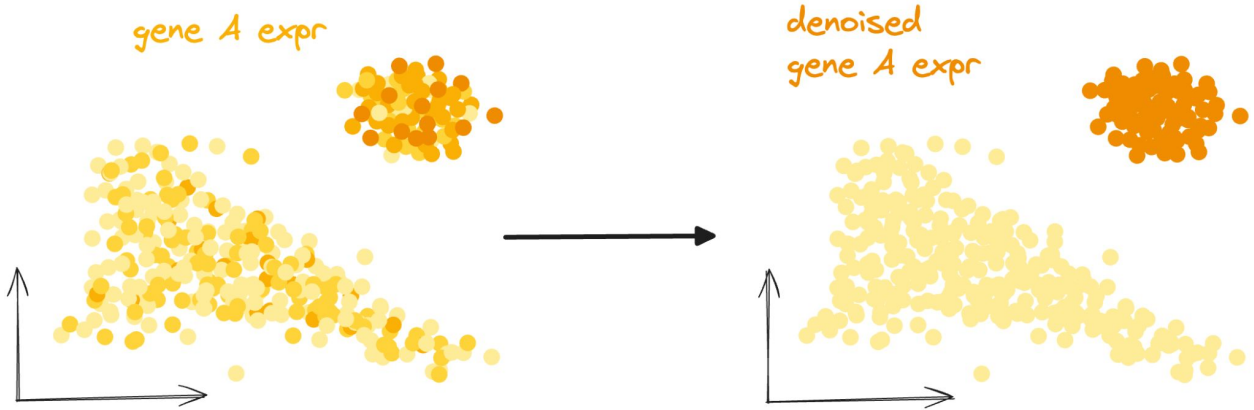




scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

3.5. scPRINT peut augmenter la qualité des données

scPRINT beats state of the art tool MAGIC on challenging rare cell populations

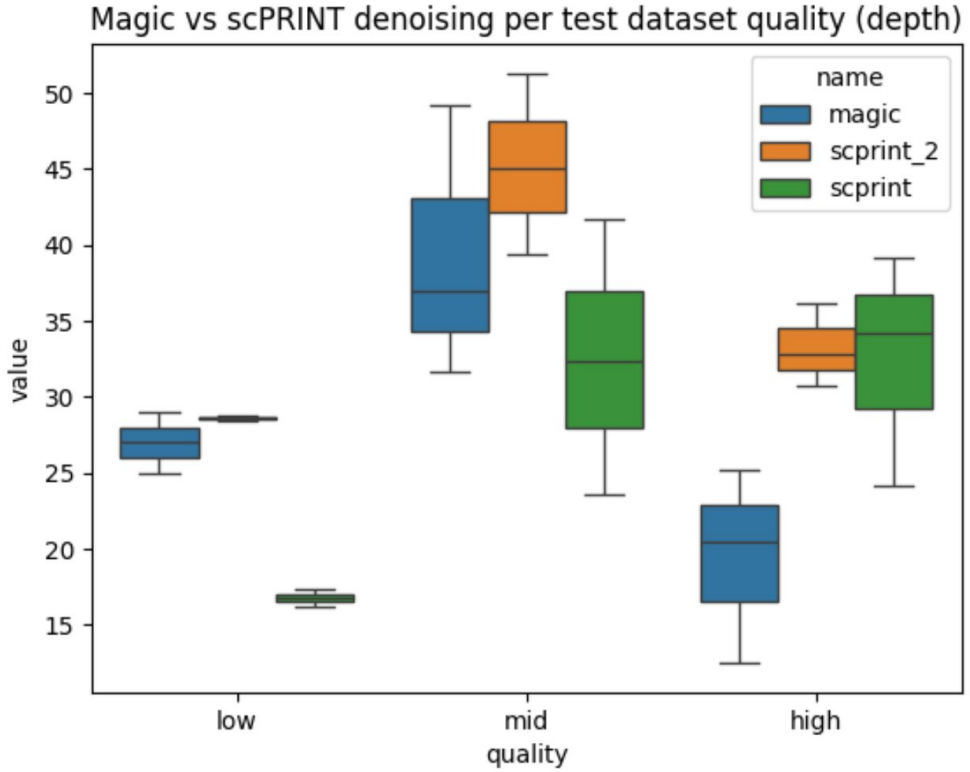


$$\text{score} = r_{pearson}(x_{true}, x_{pred}) - r_{pearson}(x_{true}, x_{noisy})$$

Recovering Gene Interactions from Single-Cell Data Using Data Diffusion

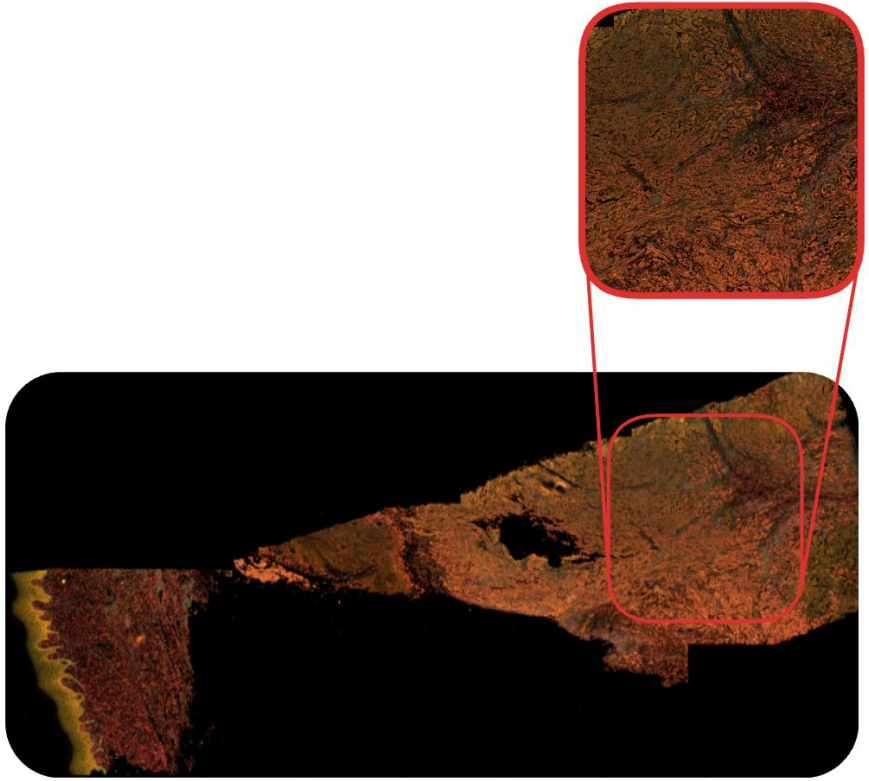
3.5. scPRINT peut augmenter la qualité des données

impact of sequencing quality and bias is reduced thanks to the GNN



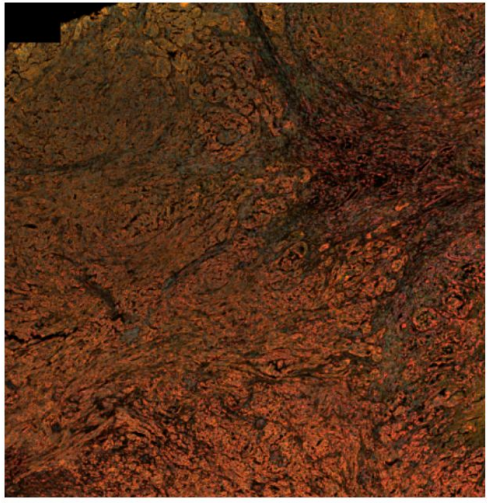
Les LCMS généralisent sur des nouvelles modalités

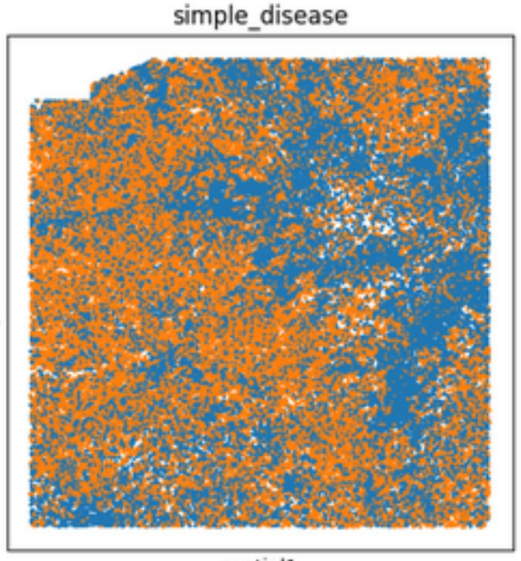
of Xenium 5K human primary skin melanoma

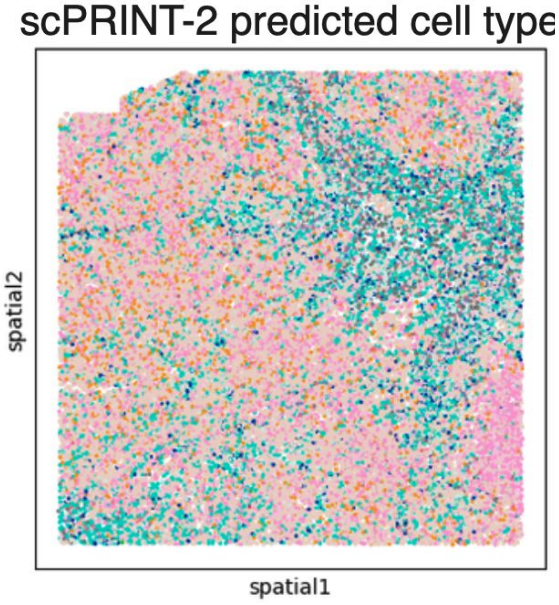


La classification sur des slides Xenium de ST

Mais présente aussi les biais toujours présents dans les données



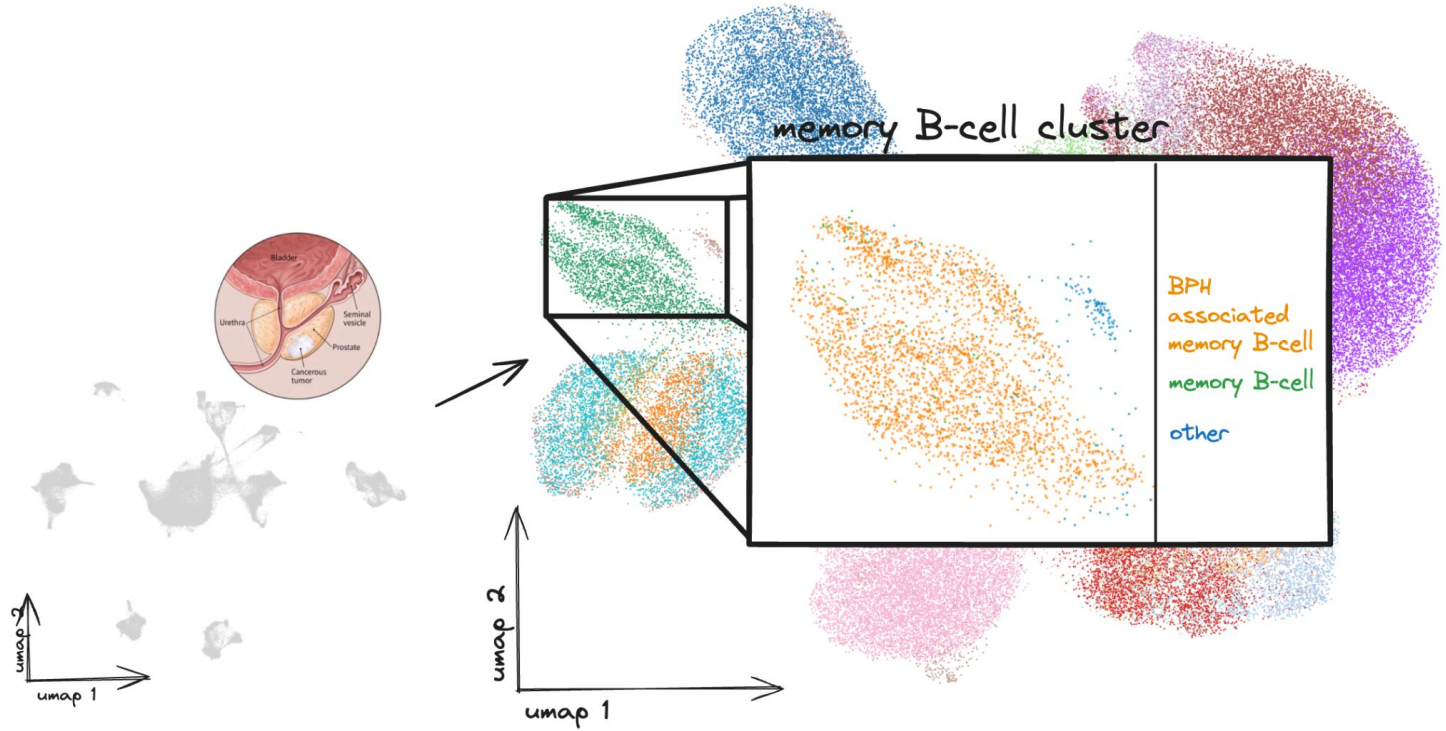




4. étude: Hyperplasie bégnine de la prostate

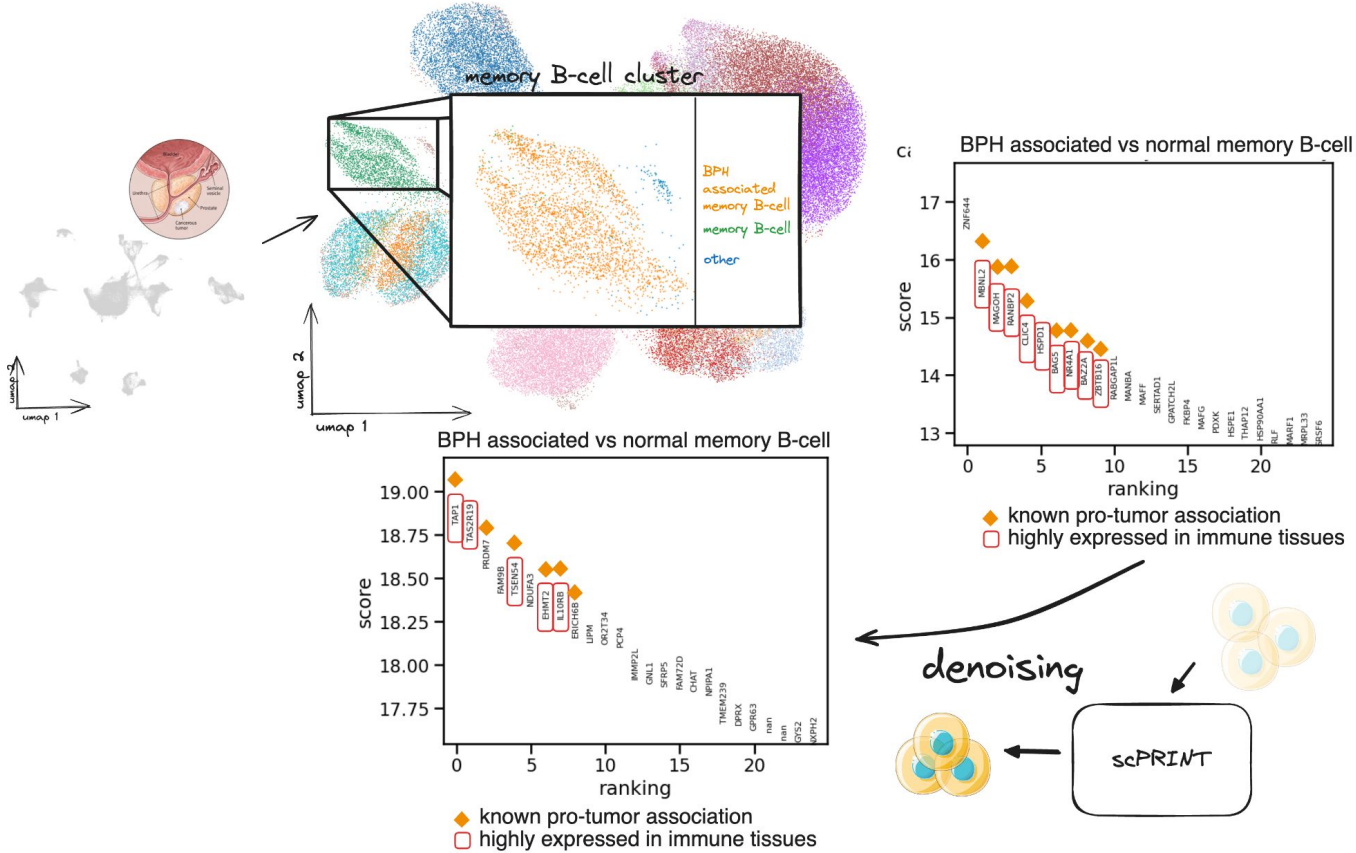
4.1. scPRINT en outil central en bioinformatique

D’un dataset brute à des premiers résultats de target discovery



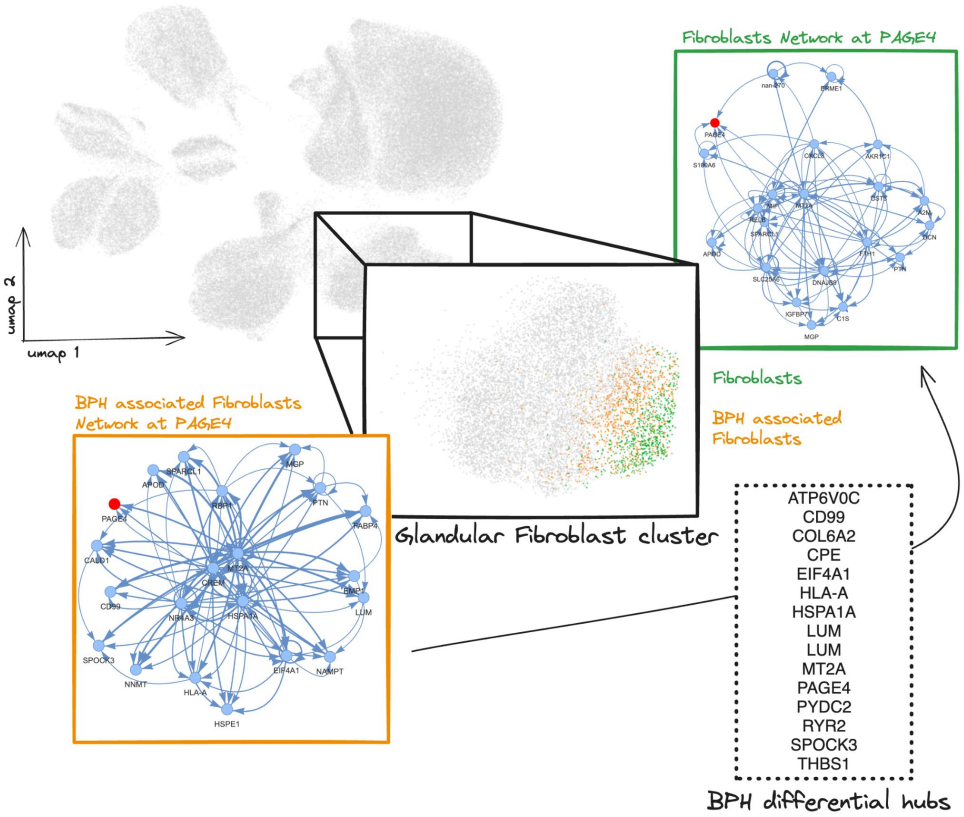
4.1. scPRINT en outil central en bioinformatique

The rare BPH-associated B-cell population predicted by scPRINT contains known markers



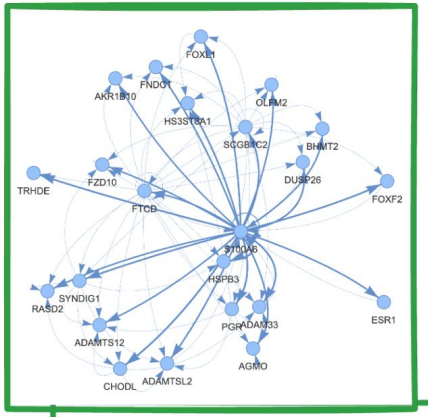
4.2. Trouver des hubs génétiques entre états cellulaires

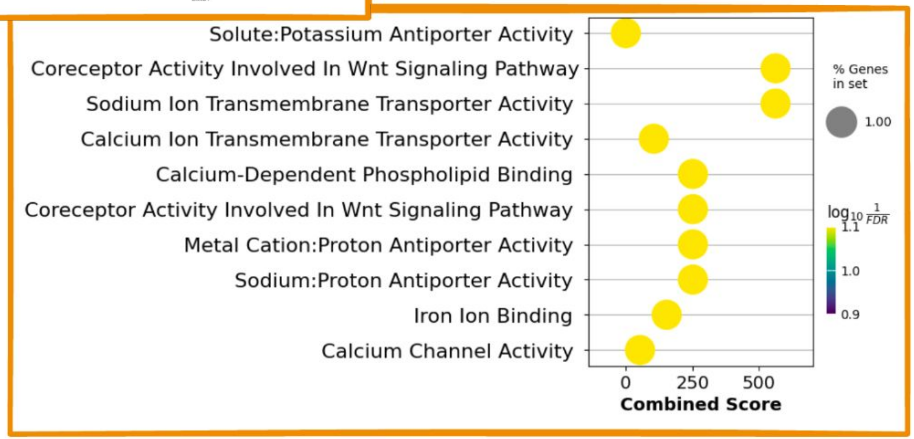
scPRINT networks are cell state specific and explain known biology in BPH

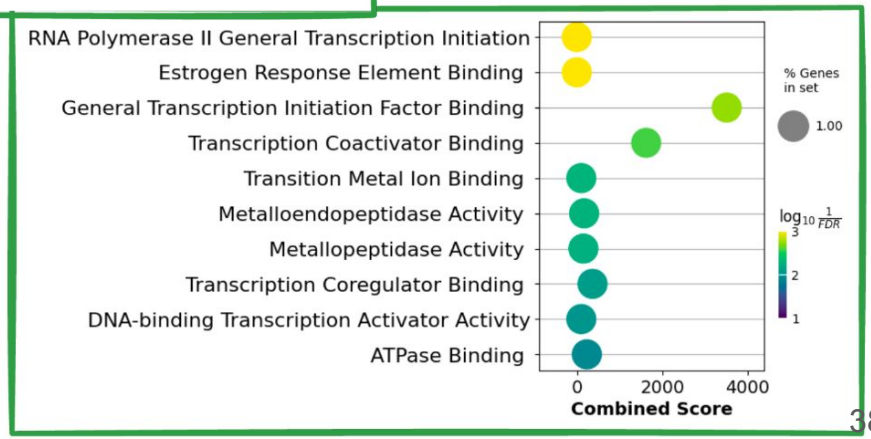


4.3. Découvertes de clusters et de mécanismes

scPRINT networks contain communities (clusters) of genes with enriched functions

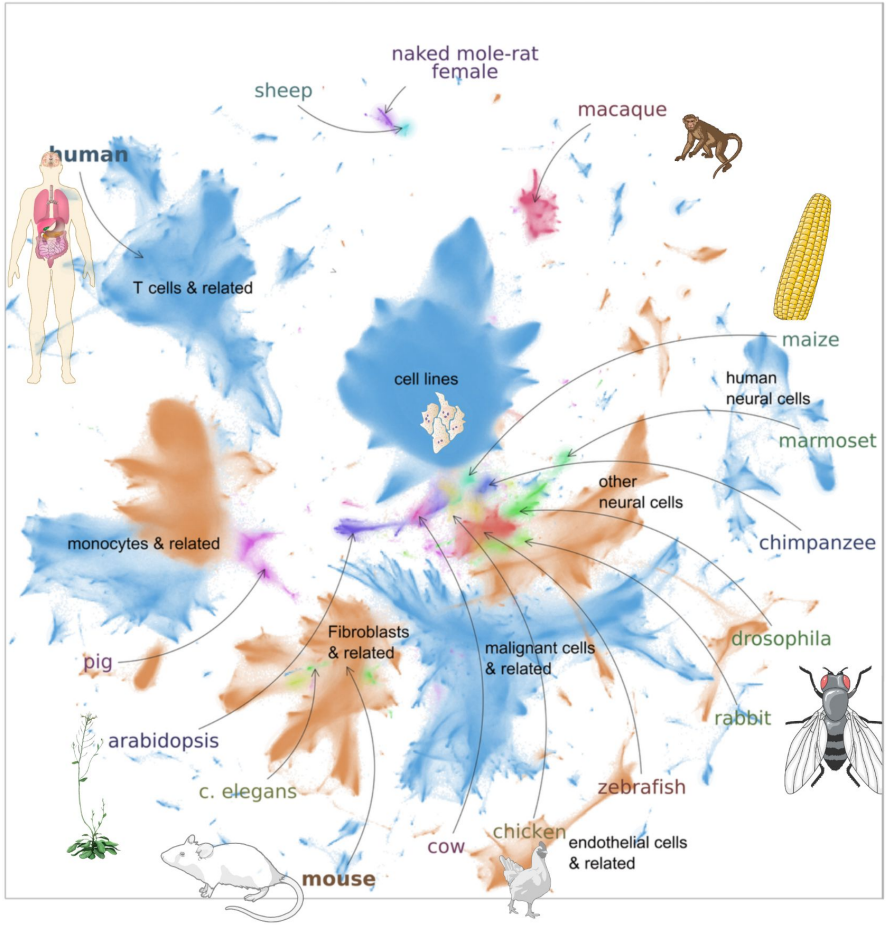




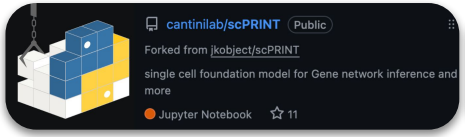


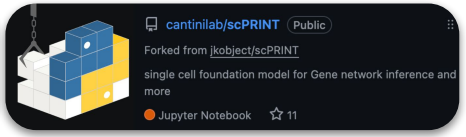
Conclusion

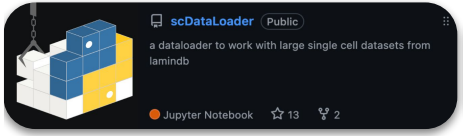
scPRINT & scPRINT-2 sont disponibles et utilisables

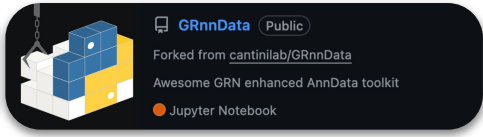


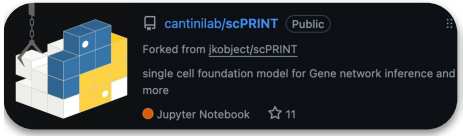
CD **pypi package** **web**
documentation **CI** **hugging face**
command line **dockerfile** **web**
platform **scverse** **wandb logs**

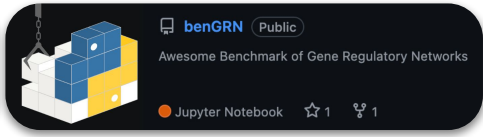








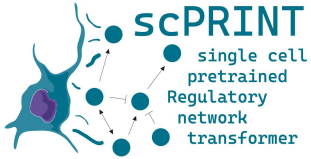




5.2. Contributions… Mais de nombreux problèmes

Better training

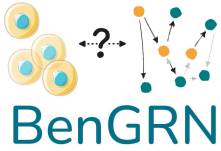
scDataLoader



Better architecture

Xpressor

Better benchmarks



But poor data

cell diversity

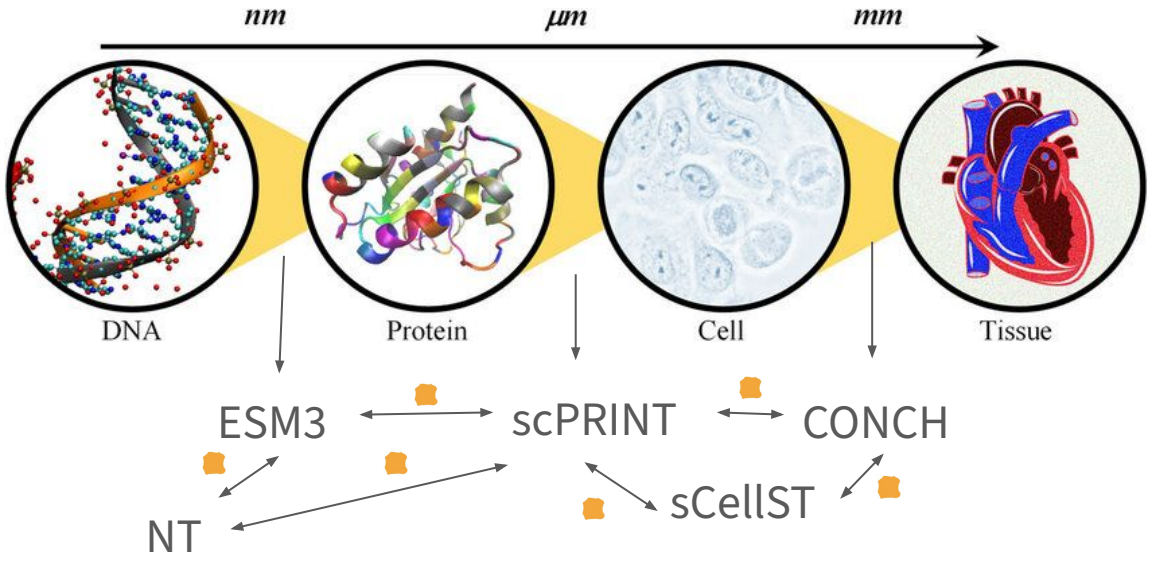
cell complexity (cell models)

sequencing quality
(90 - 99% noise)
worse for some cell types

sequencing complexity
isoforms & non-coding RNAs

spatial & perturbation are
100x worse

5.4. La suite: Créer un modèle multi échelle



Simulating 500 million years of evolution with a language model
Nucleotide Transformer: building and evaluating robust foundation models for human genomics
A visual-language foundation model for computational pathology
sCellST predicts single-cell gene expression from H& E images

1.3. Merci de votre écoute!

An amazing lab in Paris, at Institut Pasteur and at Ecole Normale Supérieure

Laura Cantini
Jérémie Kalfon
Jules Samaran
Daniele Capocefalo
Anna Audit
Clément Weinreich
Anthony Ozier Lafontaine
Lisa Chabrier
Ines Lalou
Lilas Germain
Milena Sophie Mussmann
Ana Boranijasevic
Déborah Philipps

erc MULTIview CELL

INSTITUT Pasteur

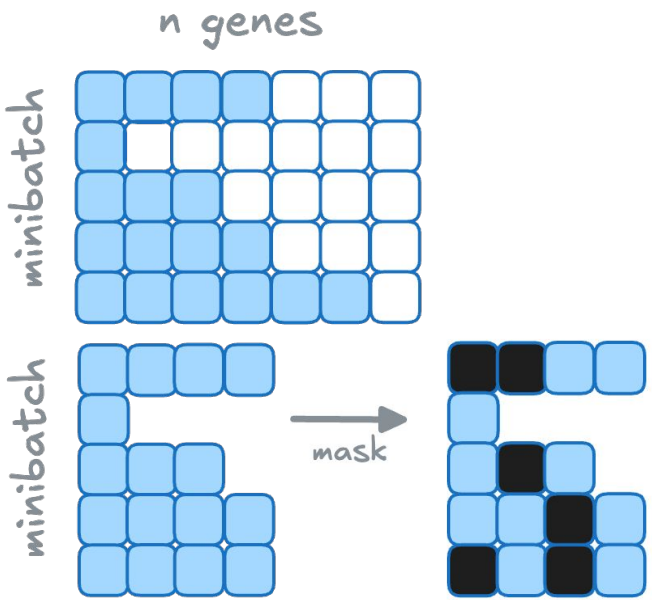
Gabriel Peyré (ENS)

Questions or Feedbacks ?

Do connect on linkedin!

1.7. Geneformer and scGPT's encoding and training

$$x \in \mathbb{R}^n$$
$$n > 20,000$$



scGPT

binned expr embedding

gene A emb + bin i emb → 0-3

gene B emb + bin j emb → 12-40

gene C emb + bin i emb → 7-12

...

geneformer

positional encoding of ranked expr

gene A emb + pos 1 emb → 18

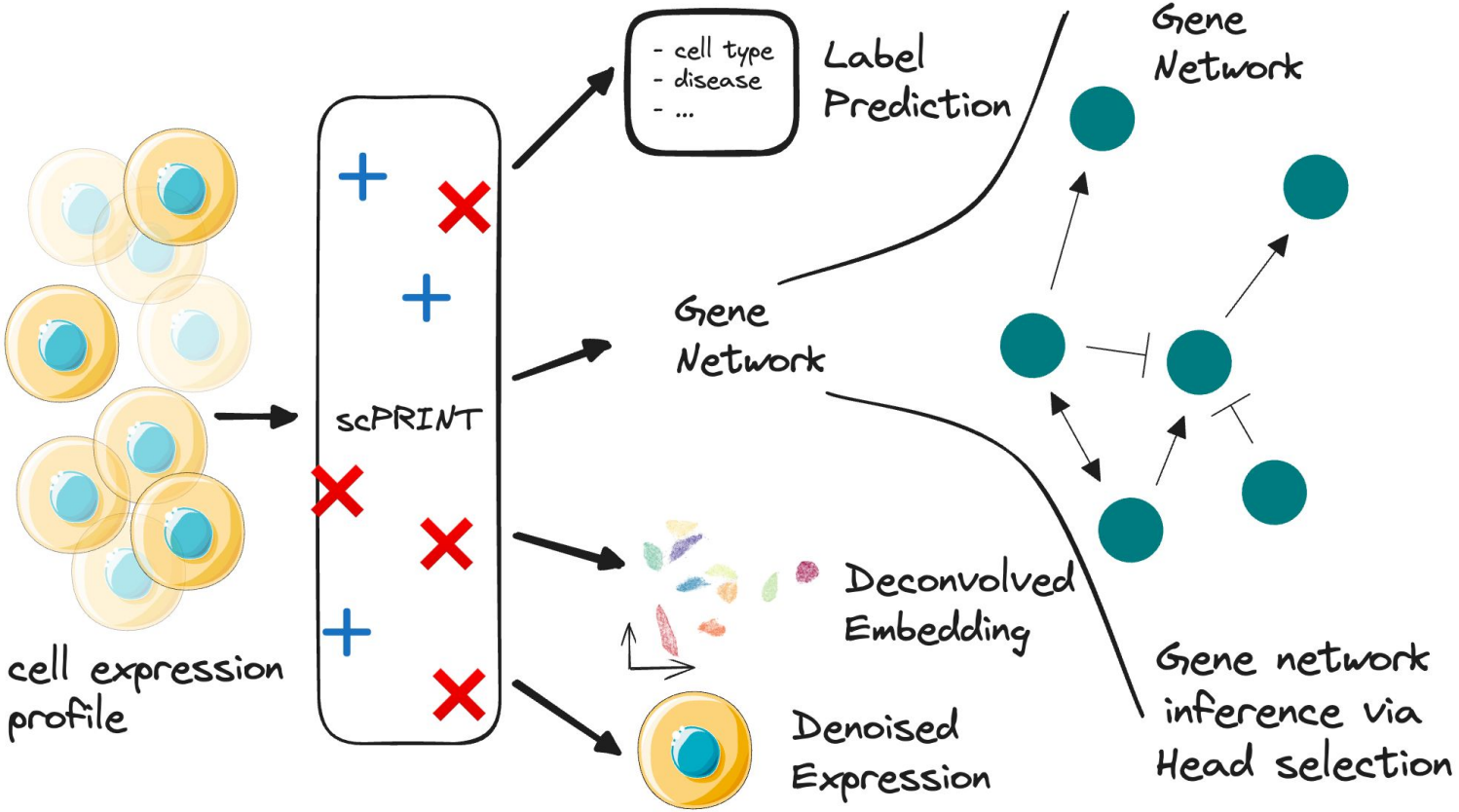
gene B emb + pos 2 emb → 12

gene C emb + pos 3 emb → 5

...

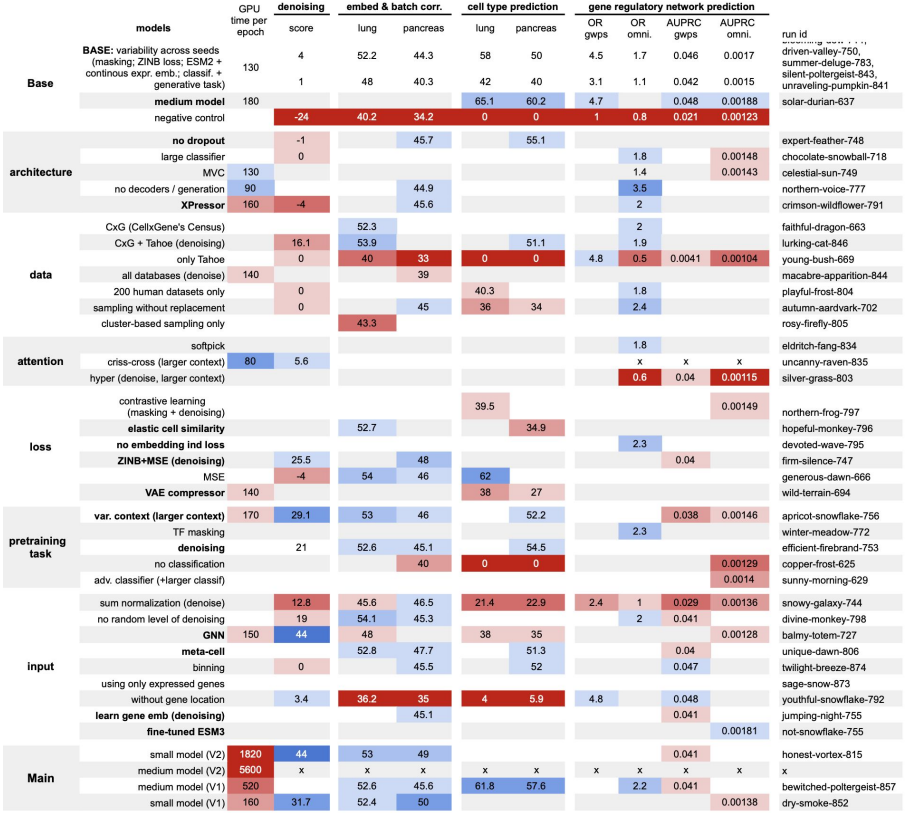
What do we test for?

Using a task gymnasium from the scPRINT paper



4.2. A definitive benchmark of scFMs

42 different features from dataset to architectures, losses and more



	models	GPU time per epoch	denoising score	embed & batch corr. lung	embed & batch corr. pancreas	cell type prediction lung	cell type prediction pancreas	gene regulatory network prediction OR gwps	gene regulatory network prediction OR omni.	gene regulatory network prediction AUPRC gwps	gene regulatory network prediction AUPRC omni.	run id
Base	**BASE:** variability across seeds (masking; ZINB loss; ESM2 + continous expr. emb.; classif. + generative task)	130	4	52.2	44.3	58	50	4.5	1.7	0.046	0.0017	driven-valley-750, summer-deluge-783, silent-poltergeist-843, unraveling-pumpkin-841
			1	48	40.3	42	40	3.1	1.1	0.042	0.0015	
	medium model	180				65.1	60.2	4.7		0.048	0.00188	solar-durian-637
	negative control		-24	40.2	34.2	0	0	1	0.8	0.021	0.00123	
architecture	**no dropout**		-1		45.7		55.1					expert-feather-748
	large classifier		0						1.8		0.00148	chocolate-snowball-718
	MVC	130							1.4		0.00143	celestial-sun-749
	no decoders / generation	90			44.9				3.5			northern-voice-777
	XPressor	160	-4		45.6				2			crimson-wildflower-791
data	CxG (CellxGene's Census)			52.3					2			faithful-dragon-663
	CxG + Tahoe (denoising)		16.1	53.9			51.1		1.9			lurking-cat-846
	only Tahoe		0	40	33	0	0	4.8	0.5	0.0041	0.00104	young-bush-669
	all databases (denoise)	140			39							macabre-apparition-844
	200 human datasets only		0			40.3			1.8			playful-frost-804
	sampling without replacement		0		45	36	34		2.4			autumn-aardvark-702
	cluster-based sampling only			43.3								rosy-firefly-805
attention	softpick								1.8			eldritch-fang-834
	criss-cross (larger context)	80	5.6						x	x	x	uncanny-raven-835
	hyper (denoise, larger context)								0.6	0.04	0.00115	silver-grass-803
loss	contrastive learning (masking + denoising)					39.5					0.00149	northern-frog-797
	elastic cell similarity			52.7			34.9					hopeful-monkey-796
	no embedding ind loss								2.3			devoted-wave-795
	ZINB+MSE (denoising)		25.5		48					0.04		firm-silence-747
	MSE		-4	54	46	62						generous-dawn-666
	VAE compressor	140				38	27					wild-terrain-694
pretraining task	**var. context (larger context)**	170	29.1	53	46		52.2			0.038	0.00146	apricot-snowflake-756
	TF masking								2.3			winter-meadow-772
	denoising		21	52.6	45.1		54.5					efficient-firebrand-753
	no classification				40	0	0				0.00129	copper-frost-625
	adv. classifier (+larger classif)										0.0014	sunny-morning-629
input	sum normalization (denoise)		12.8	45.6	46.5	21.4	22.9	2.4	1	0.029	0.00136	snowy-galaxy-744
	no random level of denoising		19	54.1	45.3				2	0.041		divine-monkey-798
	GNN	150	44	48		38	35				0.00128	balmy-totem-727
	meta-cell			52.8	47.7		51.3			0.04		unique-dawn-806
	binning		0		45.5		52			0.047		twilight-breeze-874
	using only expressed genes											sage-snow-873
	without gene location		3.4	36.2	35	4	5.9	4.8		0.048		youthful-snowflake-792
	learn gene emb (denoising)				45.1					0.041		jumping-night-755
	fine-tuned ESM3										0.00181	not-snowflake-755
Main	small model (V2)	1820	44	53	49					0.041		honest-vortex-815
	medium model (V2)	5600	x	x	x	x	x	x	x	x	x	x
	medium model (V1)	520		52.6	45.6	61.8	57.6		2.2	0.041		bewitched-poltergeist-857
	small model (V1)	160	31.7	52.4	50						0.00138	dry-smoke-852

high score — low score

What do we test for?

42 different elements from dataset to architectures, losses and more

names	GPU time per epoch	denoising	embed & batch corr.		cell type prediction		gene regulatory network prediction			
		score	lung	pancreas	lung	pancreas	OR gwps	OR omni.	AUPRC gwps	AUPRC omni.
variability across seeds	130	2,5 +-1,5	50.1 +-2.1	42.3 +-2	50 +-8	45 +-5	3,8. +-0.7	1,4 +-0.3	0,044 +-0.002	0,00165 +-0.00015
medium model	180	3	50.3	42.8	65.1	60.2	4.7	1.15	0.048	0.00188
negative control	0	-24	40.2	34.2	0	0	1	0.8	0.021	0.00123
no dropout	130	-1	49.9	45.7	47	55.1	3.9	1.5	0.044	0.00157
large classifier	130	0	52	43	53	49	4.4	1.8	0.046	0.00148
MVC	130	1	51.7	44.4	54.8	45.8	3.9	1.4	0.043	0.00143
no decoders / generation	90	2	52.2	44.9	53.2	47.7	4	3.5	0.044	0.00166
XPressor	160	-4	50.4	45.6	47.3	46.8	4.2	2	0.046	0.00163
only Tahoe	130	0	40	33	0	0	4.8	0.5	0.0041	0.00104
CZI + Tahoe (denoise)	130	16.1	53.9	43.1	52.6	51.1	4.5	1.9	0.046	0.00143

scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

4.3. We recover and discover enabling features for scFMs

Denoising ✅ (scFoundation)

Xpressor architecture ✅

Fine-tuning ESM3 ✅

Graph neural net encoding / meta-cell ✅

ZINB+MSE loss ✅

Generative decoder ❌ (UCE)

Contrastive learning ❌ (cellPLM)

…

… but questions remains

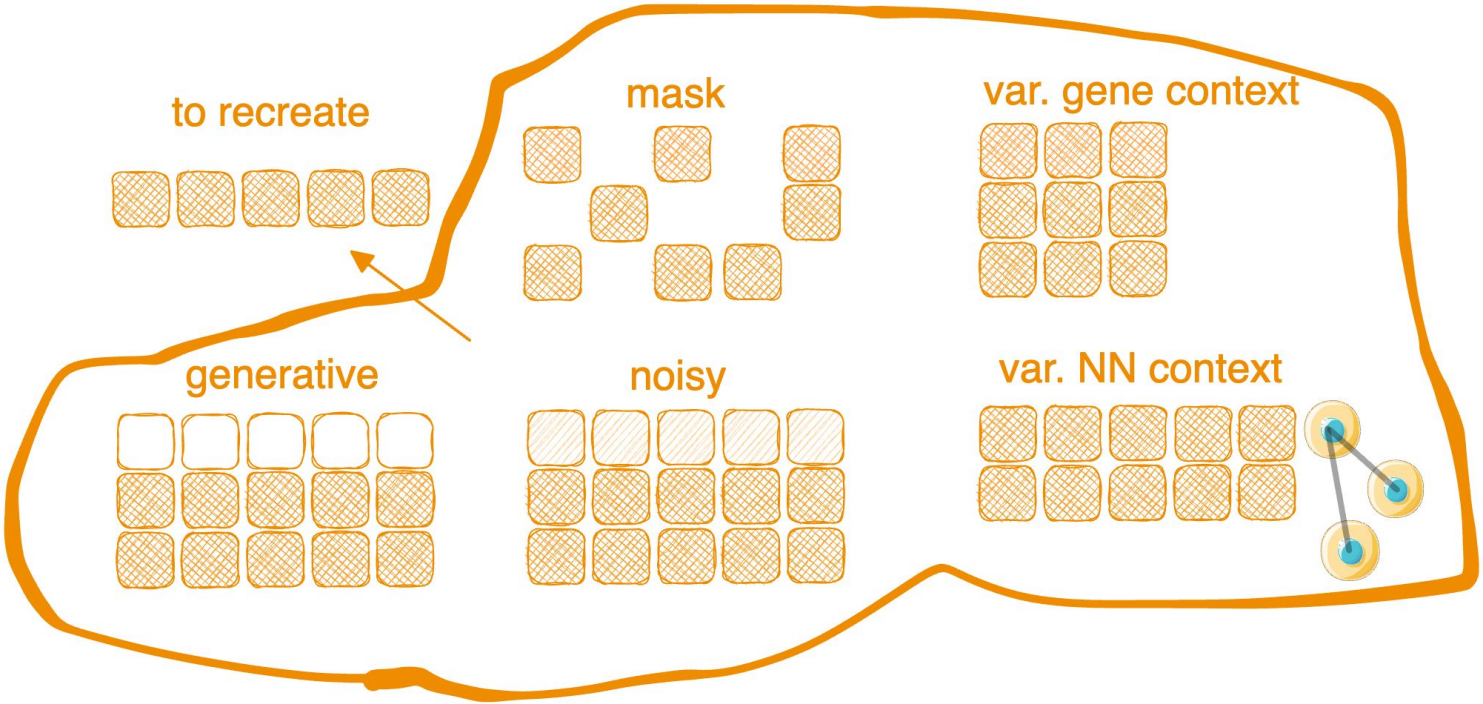
scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

We found the best set of features

We use them to create scPRINT-2 and showcase
1. Its generative & generalization abilities
2. The many issues in current benchmarks

Multi-objective training mechanism

can be combined with contrastive learning although it did not seem to help



scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

4.6. The 13 tasks of a next-generation scFM

classical

embedding quality (SotA)
cell type prediction (SotA)
Accurate predictions on Unseen **Xenium** modality

novel scores

denoising (SotA)
cell-specific **gene tokenization** (Beats scPRINT & scGPT)
GN inference (Beats scPRINT thanks to **new extraction** method)

new to scFMs

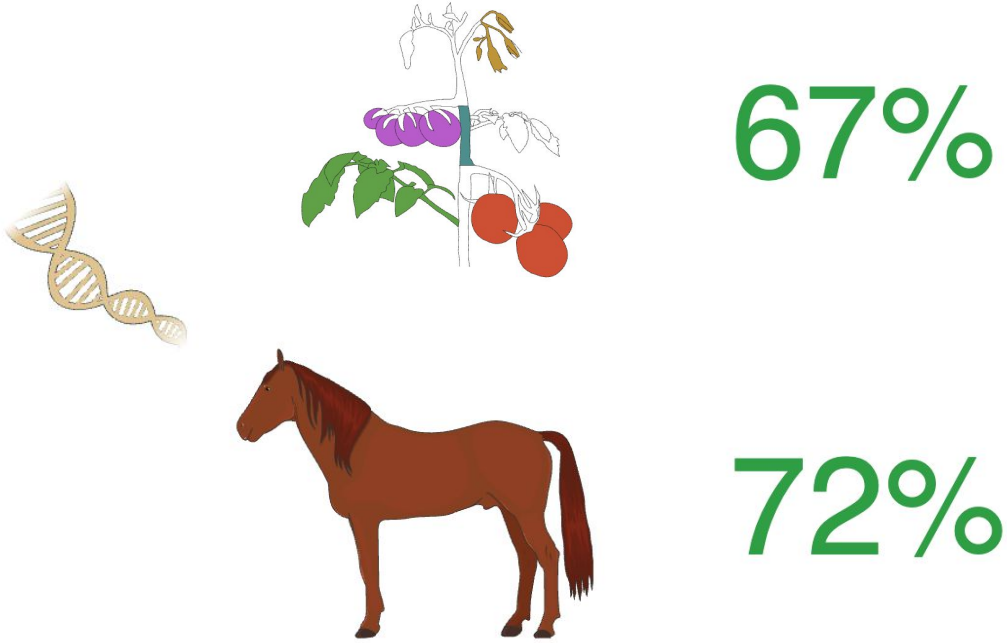
Unseen task: unseen-species-integration
Unseen gene imputation for Xenium data
unseen species cell type prediction (beats experts)

never seen before tasks

Generative cell profile given neighbors
Counterfactual reasoning to humanize a mouse dataset
Merged with **AF-MM for PPI** networks prediction
unseen organism **clade** prediction

scPRINT-2 can even infer species from unseen genetics

using ESM3 embeddings of species it has never seen it correctly associate to nearby organisms

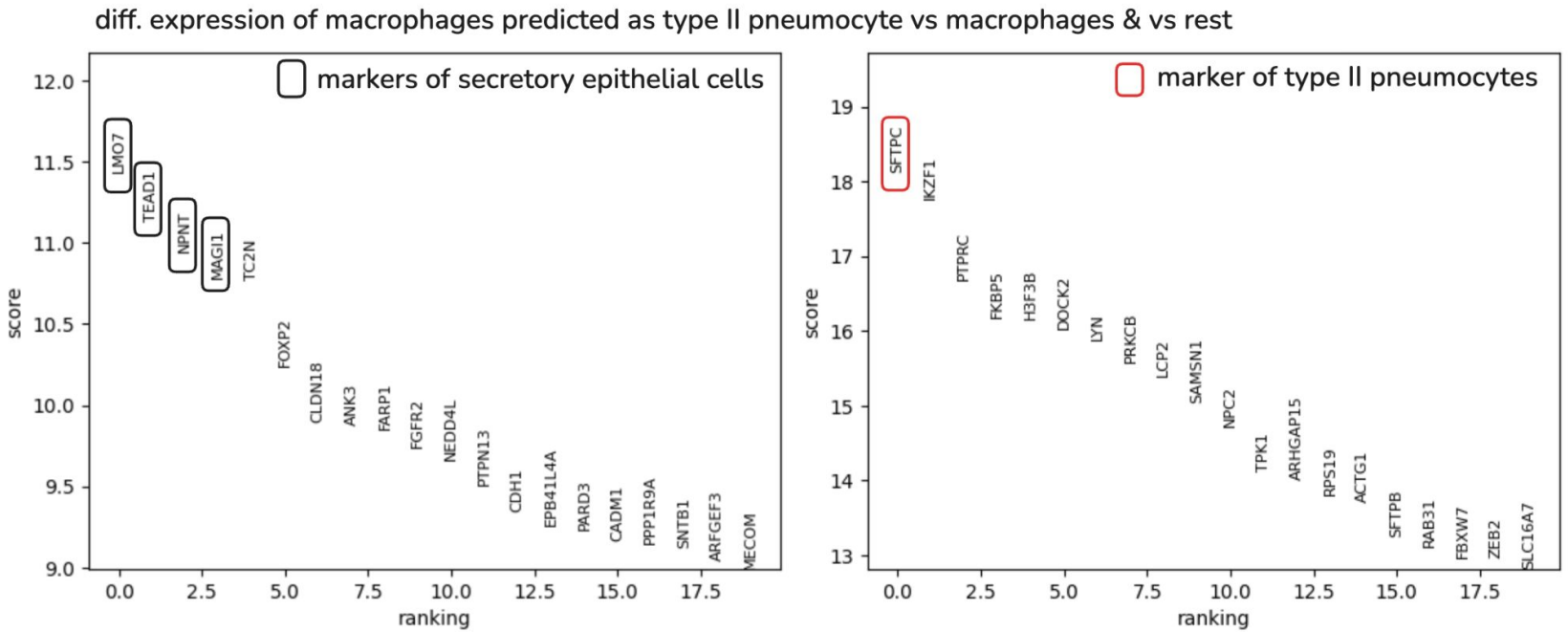


scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

scPRINT-2 can predict cell type of unseen species

scPRINT-2 can correct expert annotations

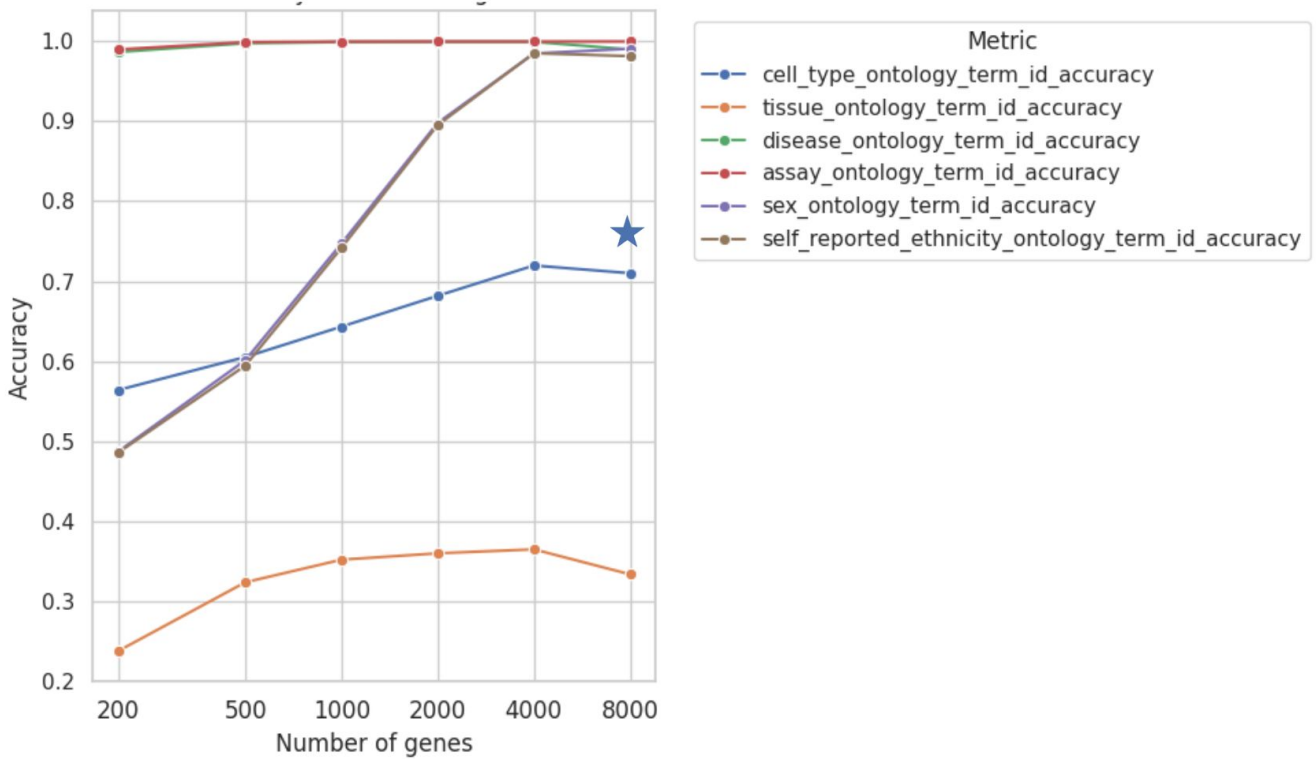
on species it has no training data for



scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

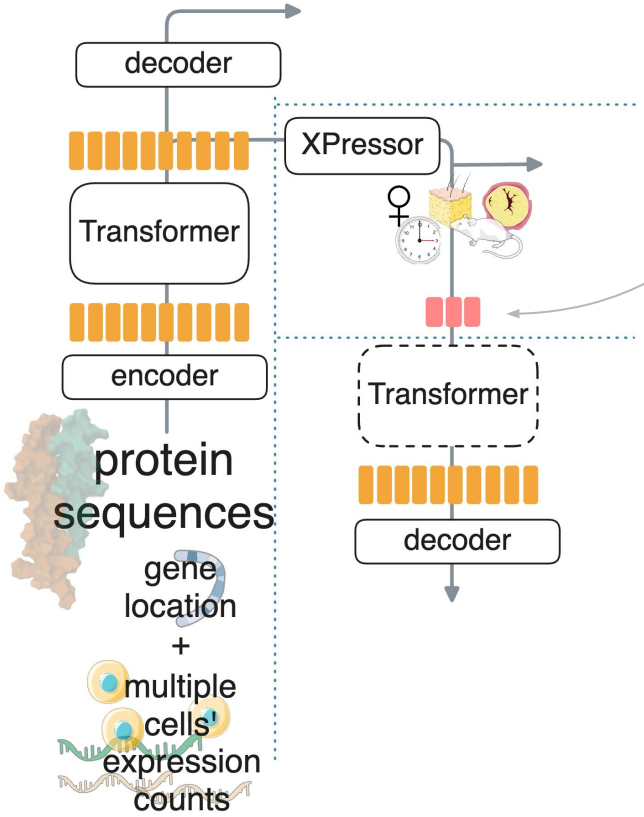
Impact of the number of genes in context

scPRINT-2 can generalize to larger contexts



2.3. scPRINT-2 used novel pre-training modalities

Learn an embedding of the cell to compress and denoise expression



AutoEncoder (AE)

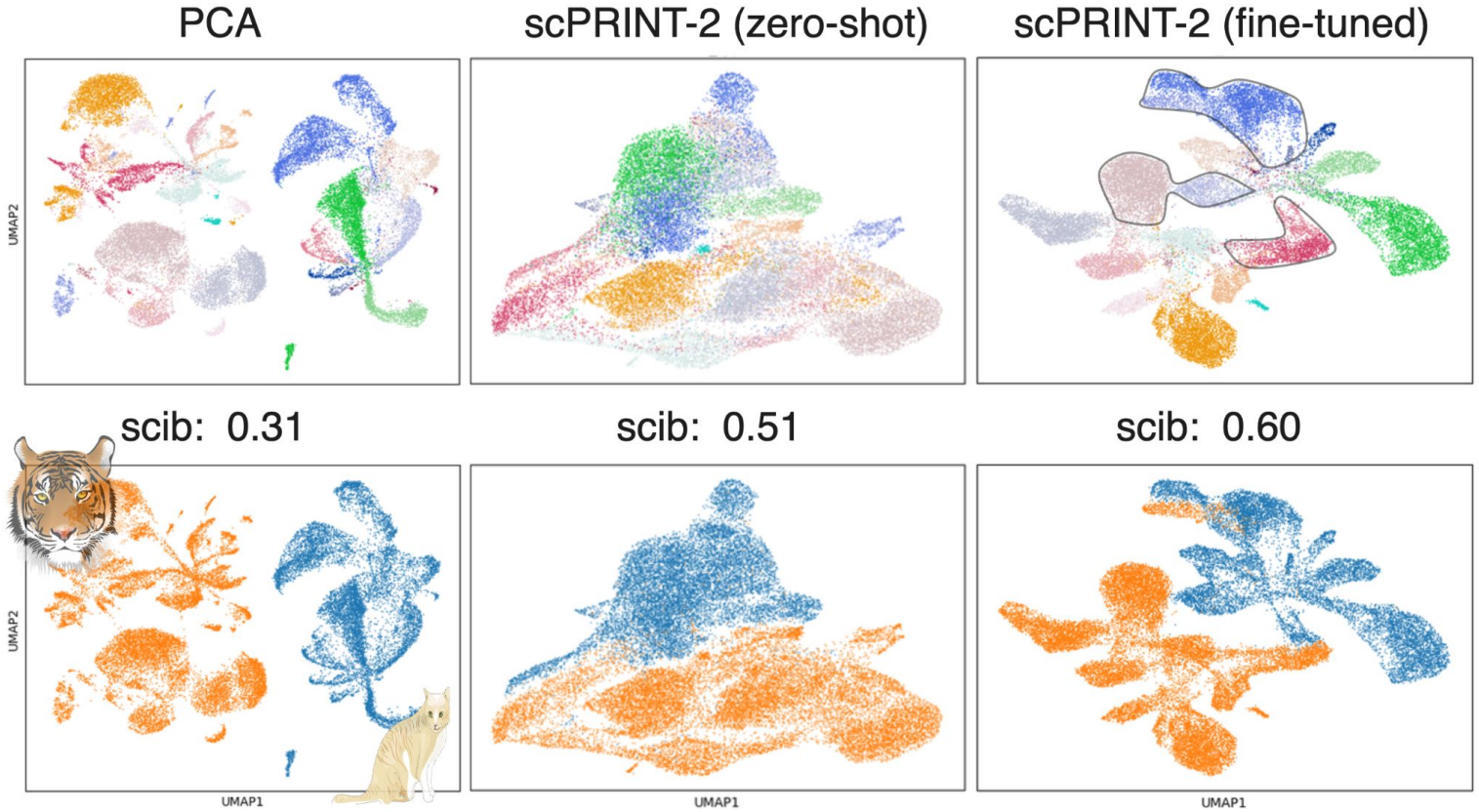
$$\text{emb} = f(\text{genes}, \text{cell})$$

$$\text{cell} = f(\text{genes}, \text{emb})$$

scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

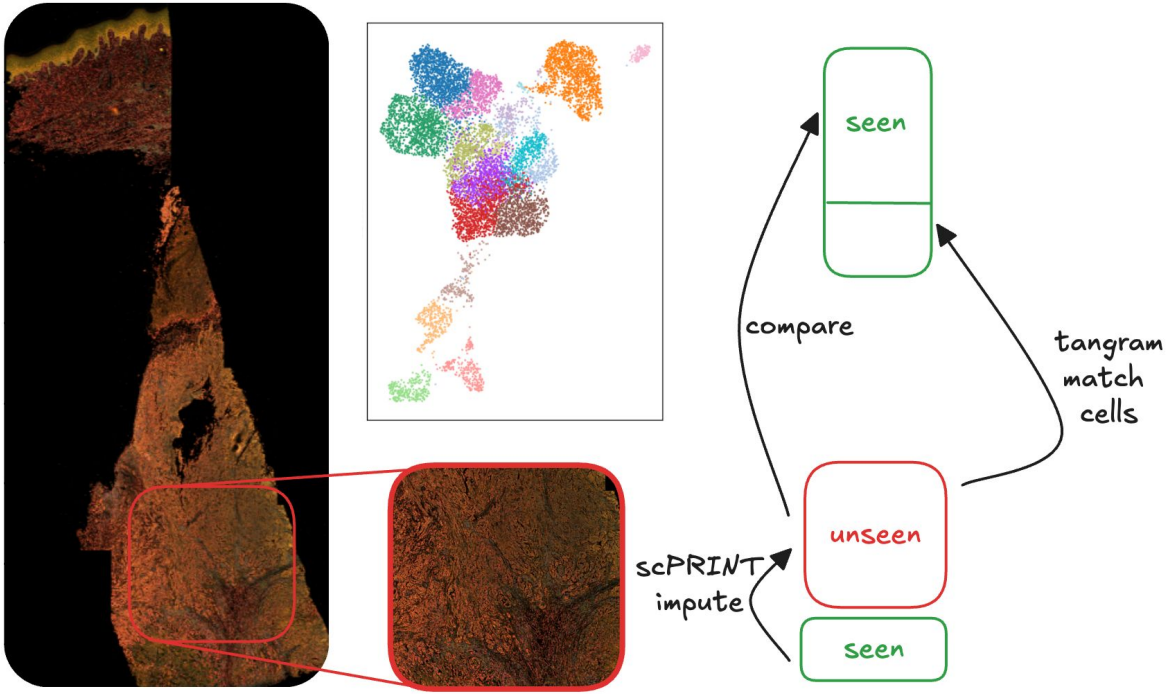
scPRINT-2 can also integrate across species

in both zero-shot and fine-tune mode



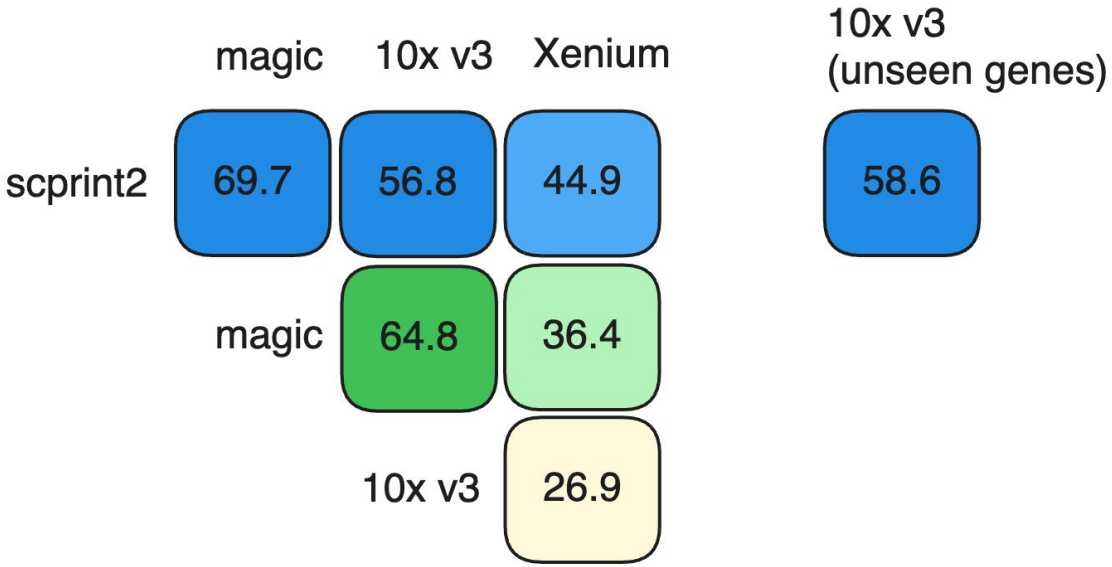
Tangram's OT mapping of another dataset

as a ground truth for our denoising… and imputation



Tangram's OT mapping of another dataset

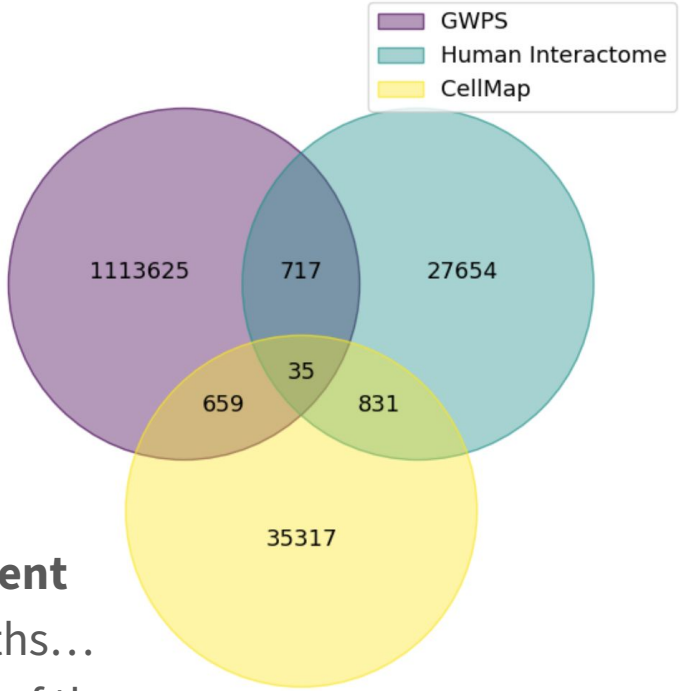
as a ground truth for our denoising… and imputation



scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

Adding 2 news databases for Gene Network inference

mostly related to PPIs, human interactome being the largest to date

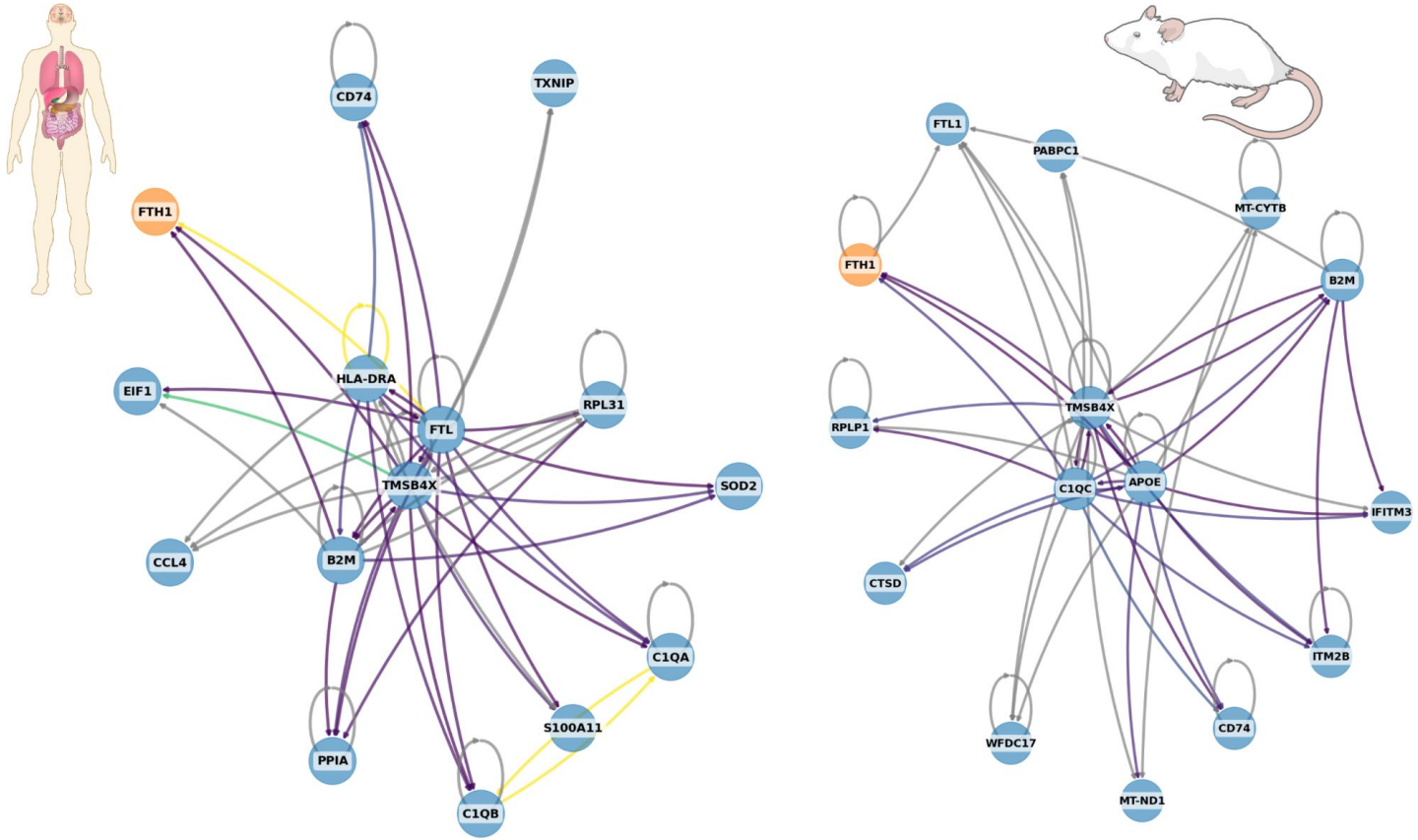


Tremendous **disagreement**
between the ground truths…
We need to assess on all of them

scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. BioRxiv

Looking at mouse / human gene networks

one can see many similarities when looking at genesets and hubs



Improve the ground truth using scPRINT-2

re-compute interactions using AF-multimer discovers new relationships