Effective Biological Representation Learning by Masking Gene Expression

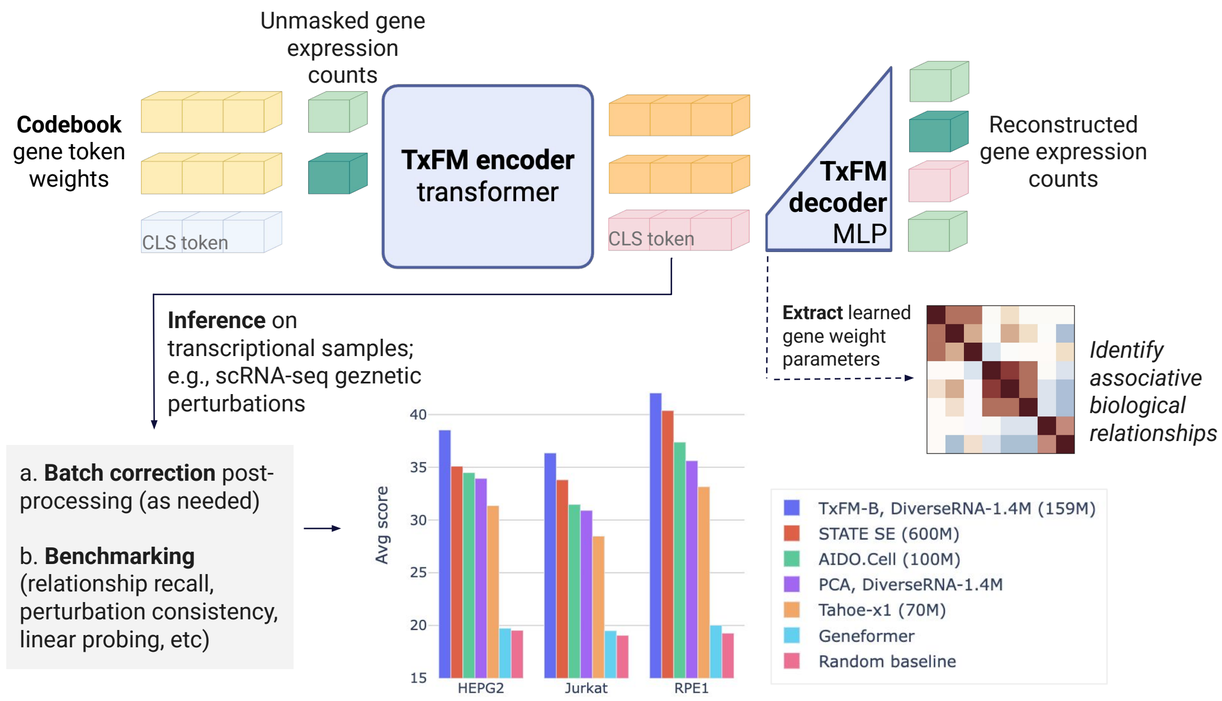
Kian Kenyon-Dean, Ihab Bendidi, Alina Selega, Jordan M. Sorokin, Luca Bertinetto, David Errington, Oren Kraus, Hayley Donnella

Recursion®

Overview

- We present **TxFM**, a SOTA transcriptomic foundation model trained by learning masked gene expression
- TxFM is SSL trained on curated set of 1.4 million bulk and single-cell samples of public data **indicating that data quality can outweigh raw dataset size**
- **SOTA perturbation representation**: TxFM establishes a new state-of-the-art for representing genetic perturbations across three evaluation datasets with held-out cell lines, outperforming 16 public models or FMs and multiple strong baselines
- **Gene-level interpretability**: TxFM's learned gene parameters (encoder token embeddings and decoder reconstruction weights) recover many known gene-gene relationships (e.g. pathways, protein complexes) without direct supervision.

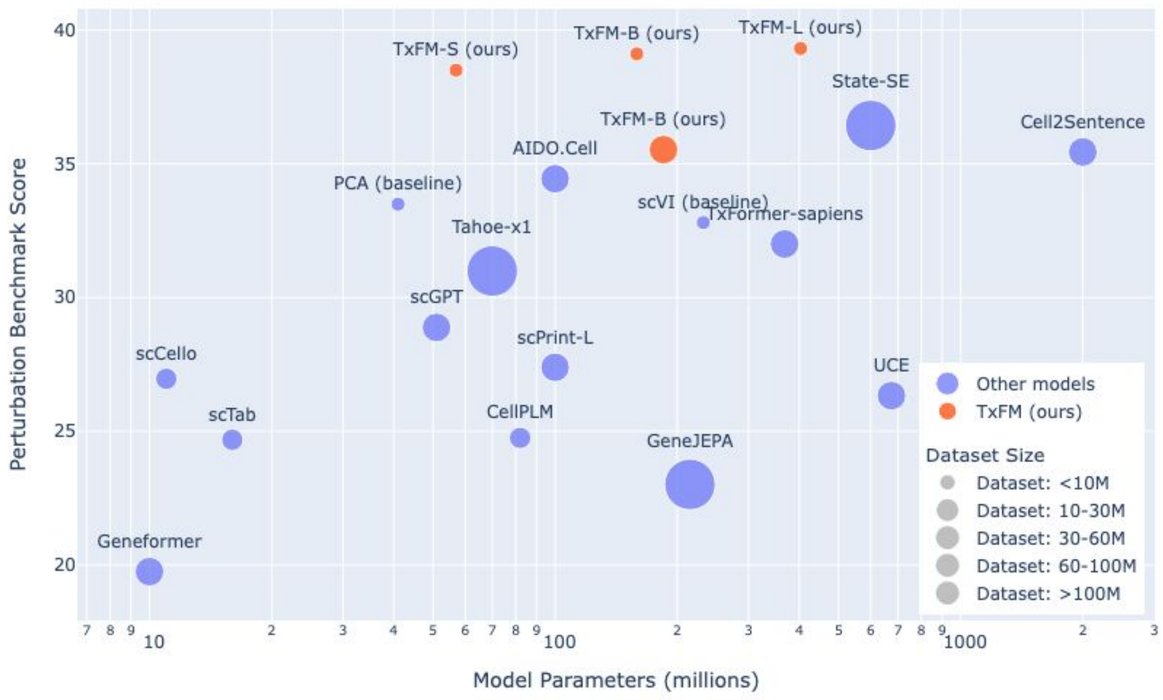
Overview of our approach. TxFM is pre-trained for masked reconstruction of gene expression, and evaluated on downstream sample representation quality and gene–gene association structure encoded in its parameters.



New curated SSL training dataset: DiverseRNA-1.4M. Composition of our default TxFM training dataset afte we apply specialized data curation with an orientation to oncology. Each dataset is publicly available. sc: single-cell.

Dataset reference	mode	# samples	# genes
Replogle et al. phenoprint cells	sc	502,080	8,248
Ruiz-Moreno et al., Glioblastoma	sc	504,929	21,310
McFarland et al.,	sc	102,205	15,438
Srivatsan et al.	sc	99,300	18,486
Nowicki-Osuch et al.	sc	88,399	16,445
Guimarães et al.	sc	71,585	17,619
Wu et al.	sc	31,542	24,712
TCGA Weinstein et al.	bulk	23,733	19,594
GTEX Consortium	bulk	10,526	36,695
DiverseRNA-1.4M	**mixed**	**1,434,299**	**44,349**

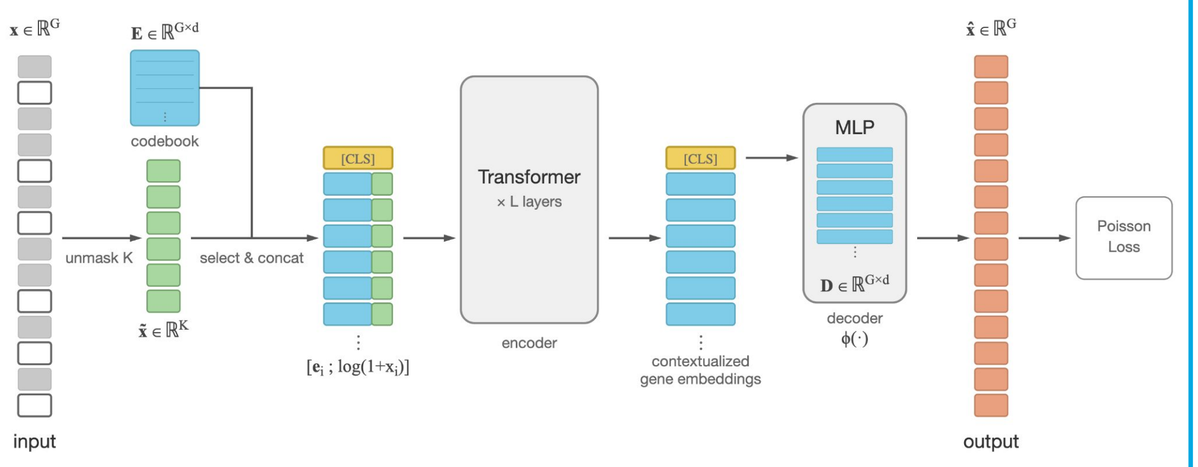
Overview of transcriptomic model complexity and performance. We report the total average perturbation representation score (Bendidi et al., 2024) over the RPE1 (Replogle et al., 2022) and HEPG2 and Jurkat (Nadig et al., 2024) genetic perturbation datasets unseen during training (*inductive*). PCA and scVI baselines are trained on same data as TxFM (DiverseRNA-1.4M).



Fine-tuning performance: Average perturbation representation score over the six metrics for bespoke PCA, scVI, and SSL-fine-tuned TxFM-B models trained on RPE1, HEPG2, Jurkat separately. This is a *transductive* representation learning evaluation.

Model / Dataset	RPE1	HEPG2	Jurkat
TxFM-B (ours, SSL-finetuned)	**44.85**	**40.78**	**38.01**
PCA	42.64	38.63	34.89
scVI	37.65	35.63	32.62

TxFM's MAE architecture and ablation study



$$\phi(z) = \log(L+1)\mathrm{ReLU}\left(\tanh\left(\frac{z}{4e}\right)\right) \quad \mathcal{L}_{\text{Poisson}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{G}\sum_{i=1}^{G}(e^{\hat{\mathbf{x}}_i} - \hat{\mathbf{x}}_i e^{\mathbf{x}_i}).$$

loss	perts	enc	dec
MSE	27.2*	31.1*	41.8*
SmoothL1	23.4*	23.2*	38.1*
Poisson	**37.3**	**32.6**	**43.9**
Neg. Bin.	33.9*	29.4*	41.2*

(a) **Reconstruction loss.** Our Poisson-based loss function significantly improves performance.

preprocessing	perts	enc	dec
log1p	27.9*	25.7*	42.2*
LibNorm, log1p	**37.3**	**32.6**	**43.9**

(b) **Count preprocessing.** Normalizing count data for library size helps to learn high-quality perturbational and non-perturbational gene representations.

# unmasked	perts	enc	dec
512	31.9*	26.8*	40.2*
1024	35.2*	31.9*	**44.3**
2048	**37.3**	**32.6**	43.9
4096	35.6*	25.3*	42.8*

(c) **Mask ratio.** Training with 2048 unmasked gene tokens (∼90% mask ratio) yields strong performance despite 4096 having more training compute.

temperature τ	perts	enc	dec
-1	31.3*	32.5	42.7*
-0.5	34.4*	**34.3***	43.6
0 (uniform)	37.3	32.6	**43.9**
0.5	**37.6**	28.9*	42.0*
1.0	36.4	25.7*	40.3*

(d) **Masking strategy.** Masking genes uniformly at random offers a good performance trade-off vs frequency-weighted masking.

decoder	perts	enc	dec
linear	36.5	32.3	43.6
1-layer MLP	36.0	32.3	43.8
4-layer MLP	**37.3**	**32.6**	**43.9**
8-layer MLP	36.5	32.0*	43.6

(e) **Decoder depth.** A linear or MLP decoder works effectively with minimal significant differences.

activation	perts	enc	dec
ReLU (diverges)	34.4*	26.7*	43.9
clamped ReLU	34.6*	32.4	**44.3**
rect. tanh	**37.3**	**32.6**	43.9
softmax	27.6*	16.2*	15.3*

(f) **Decoder count activation.** Our rectified tanh activation function prevents divergence and improves representation quality.

backbone	perts	enc	dec
-S (57M)	33.2*	29.2*	43.9
-B (159M)	**37.3**	**32.6**	**43.9**
-L (403M)	37.2	28.2*	43.1*

(g) **Encoder backbone.** Scaling TxFM to Base architecture on DiverseRNA-1.4M performed best on these tasks.

train data (# samples) [compute]	perts	enc	dec
DiRNA w/o K562 cells (932K)	31.4*	31.5*	42.0*
DiRNA + K562 controls (1.0M)	34.3*	32.0*	42.2*
DiRNA + curated K562 (1.4M)	**37.3**	32.6	43.9
DiRNA + full K562 (2.8M)	35.6*	31.2*	**44.9***
DiRNA + full K562 (2.8M) [2x]	36.7	29.8*	44.7*
TF-Sapiens data (57M) [4x]	30.3*	**35.2***	40.3*

(h) **Dataset curation.** A phenoprints-oriented data curation strategy of DiverseRNA-1.4M is an effective way to train TxFM-B.

Codebook and decoder map relationship recall during SSL training: Gene-gene relationship recall (95th percentile threshold) across 5 databases and all combined. During training, TxFM naturally learns biological structure both in its encoder gene tokens and decoder gene reconstruction weights, outperforming scVI's equivalent parameters trained on the same data.