Supplementary Information

Evaluation protocol choices dominate reported performance in gene regulatory network benchmarks for single-cell foundation models

Ihor Kendiukhov

Supplementary Tables

Table 1: **Supplementary Table S1: Median AUPR by mapping policy.** Aggregate AUPR across all methods and gold standards for each symbol-mapping ambiguity-resolution policy. Differences are modest in aggregate but can matter in sparse-overlap regimes.

Mapping policy	Median AUPR
Drop ambiguous	0.00679
Drop unmapped	0.00663
Lexicographic	0.00643

Table 2: **Supplementary Table S2: Edge and gene coverage by tissue and gold standard.** Average coverage after symbol normalization. Coverage is uniformly high (>0.99) across all conditions, confirming that the mapping pipeline successfully resolves naming inconsistencies.

Tissue	Gold standard	Edge coverage	Gene coverage
Immune	DoRothEA curated	0.998	0.998
Immune	DoRothEA ChIP-seq	0.999	0.999
Immune	TRRUST v2	0.995	0.995
Kidney	DoRothEA curated	0.998	0.998
Kidney	DoRothEA ChIP-seq	0.999	0.999
Kidney	TRRUST v2	0.995	0.995
Lung	DoRothEA curated	0.998	0.998
Lung	DoRothEA ChIP-seq	0.999	0.999
Lung	TRRUST v2	0.995	0.995

Table 3: **Supplementary Table S3: Base rates by gold standard and candidate-set protocol.** Positive base rates vary over four orders of magnitude depending on gold-standard choice and candidate-set restriction. These differences mechanically drive AUPR differences even when model predictions are identical.

Gold standard	Candidate set	Base rate
TRRUST v2	Universe-aware	2.02e-06
TRRUST v2	TF-source	4.10e-04
TRRUST v2	TF-src+tgt	3.80e-03
DoRothEA curated	Universe-aware	3.23e-05
DoRothEA curated	TF-source	2.01e-03
DoRothEA curated	TF-src+tgt	2.60e-02
DoRothEA ChIP-seq	Universe-aware	2.50e-05
DoRothEA ChIP-seq	TF-source	1.80e-03
DoRothEA ChIP-seq	TF-src+tgt	9.98e-03

Table 4: **Supplementary Table S4: Bootstrap confidence interval widths by candidate-set protocol.** CI widths increase 77× from universe-aware to TF-source+target restriction, demonstrating that the most commonly used evaluation settings produce the least stable results.

Candidate set	Median CI width	Mean CI width
Universe-aware	1.02e-04	1.15e-04
TF-source	2.16e-03	2.34e-03
TF-src+tgt	7.86e-03	8.12e-03

Table 5: **Supplementary Table S5: AUPR stability under random label noise.** Mean AUPR and standard deviation across five repeats for each noise rate, aggregated across all methods. Variance increases with noise rate and is highest under TF-restricted candidate-set protocols.

Candidate set	Noise 5%	Noise 10%	Noise 20%	Noise 30%
Universe-aware	3.1e-04 ± 2e-05	3.0e-04 ± 3e-05	2.8e-04 ± 4e-05	2.6e-04 ± 5e-05
TF-source	4.2e-03 ± 3e-04	4.0e-03 ± 5e-04	3.6e-03 ± 7e-04	3.2e-03 ± 9e-04
TF-src+tgt	1.3e-02 ± 8e-04	1.2e-02 ± 1e-03	1.1e-02 ± 1.4e-03	9.8e-03 ± 1.6e-03

Table 6: **Supplementary Table S6: Method summary by tissue.** Median AUPR, AUROC, and F1 for each method stratified by tissue, aggregated across gold standards and mapping policies.

Tissue	Method	AUPR	AUROC	F1
Immune	GRNBoost2	0.0180	0.628	0.018
Immune	GENIE3	0.0155	0.614	0.016
Immune	Spearman	0.0058	0.552	0.006
Immune	Pearson	0.0056	0.548	0.006
Immune	Random	0.0015	0.500	0.002
Immune	scGPT attention	0.0009	0.488	0.001
Kidney	GRNBoost2	0.0330	0.665	0.033
Kidney	GENIE3	0.0290	0.651	0.029
Kidney	Spearman	0.0112	0.581	0.011
Kidney	Pearson	0.0108	0.576	0.011
Kidney	Random	0.0030	0.500	0.003
Kidney	scGPT attention	0.0019	0.492	0.002
Lung	GRNBoost2	0.0300	0.658	0.030
Lung	GENIE3	0.0265	0.644	0.027
Lung	Spearman	0.0100	0.574	0.010
Lung	Pearson	0.0096	0.569	0.010
Lung	Random	0.0027	0.500	0.003
Lung	scGPT attention	0.0017	0.490	0.002