ICML 2026

FLAG

Foundation-model representation with Latent diffusion Alignment via Graph for spatial gene expression prediction

Reframing histology-to-gene-expression as structure-aware distribution modeling.

Qi Si* · **Penglei Wang*** · **Yushuai Wu** · **Yifeng Jiao** · **Xuyang Liu** · **Xin Guo** · **Yuan Qi** · **Yuan Cheng**

Shanghai Academy of AI for Science · Shanghai Jiao Tong University · Fudan University · Zhongshan Hospital

</> github.com/darkflash03/FLAG

BACKGROUND

Predicting molecular maps from routine pathology

Spatial transcriptomics (ST) measures gene expression while preserving where each cell sits — revealing tissue microenvironments and disease niches. But ST sequencing is **expensive and low-throughput.**

H&E whole-slide images are cheap and present in **every clinical workflow.** Goal: predict spot-level gene expression directly from H&E.

From a single stained image, recover a high-dimensional expression vector at every spot.

Enables **pathway analysis · spatial-domain discovery · microenvironment characterization**

MOTIVATION

Pointwise accuracy ≠ biological validity

Existing models treat gene inference as **independent scalar regressions**, judged only by pointwise metrics (PCC / MSE). These ignore the **structure** that is often the most informative signal.

Gene–gene relationships lost

Regulatory programs and co-expression networks are erased when genes are predicted in isolation.
→ **broken pathway analysis.**

Spatial organization lost

Regression averages out one-to-many mappings, producing over-smoothed maps that don't match tissue morphology.
→ **broken spatial-domain discovery.**

MOTIVATION

Two structural metrics to measure what matters

We make biological structure explicitly measurable — beyond per-gene accuracy.

GSC Gene Structural Correlation

Does the predicted gene–gene correlation network match the truth?
Pearson correlation between the vectorized gene-gene correlation matrices of prediction vs. ground truth.

High GSC → regulatory networks preserved

SSC Spatial Structural Correlation

Is the spatial pattern of each gene reproduced?
Correlation between predicted and true Moran's I (spatial autocorrelation) across genes.

High SSC → tissue architecture preserved

MOTIVATION → IDEA

From deterministic regression to distribution modeling

Regression

learns the conditional mean

Histology→expression is stochastic & one-to-many. Averaging collapses variance → **over-smoothed maps** with no internal structure.

High PCC, but structure is gone

Diffusion

learns the full joint distribution

Approximates the high-dimensional probability manifold instead of just its mean.
Preserves intrinsic variance **and joint gene correlations** that pointwise objectives ignore.

Structured outputs → GSC & SSC preserved

MOTIVATING ATTEMPT

First try: diffuse the whole spatial graph

Treat spots as **nodes** (expression X) and their functional correlations as **edges** (A), then jointly model p(X, A | C) with graph diffusion + a novel Edge-Modulated Attention.

Why bother with edges?

Pilot (G=50): adding oracle spot-spot gene correlations lifts PCC from ~0.68 to ~0.72.

Latent functional topology is a strong constraint

…so we make edges a generative target too

How it works

Joint score-matching denoises both nodes & edges on a fully-connected tissue graph.

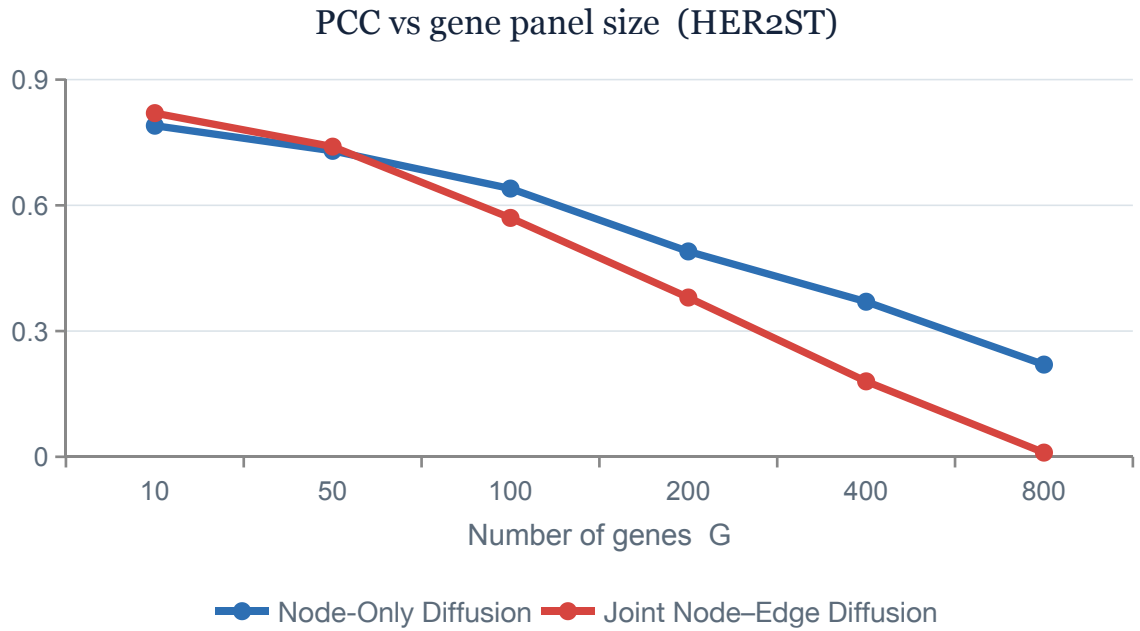
Consistency loss ties generated edges to the correlations implied by the denoised expression — preventing degenerate graphs.

$$L = L_{graph} + \lambda \cdot L_{cons}$$

THE KEY OBSTACLE

The Gene Dimension Curse

Joint node–edge diffusion works at small G, but **collapses** as the gene panel grows — exactly the regime real ST panels live in.



Why it collapses

As G grows, empirical spot-spot correlations **concentrate sharply** → the consistency manifold becomes razor-thin → the edge score field needs **huge gradients** that exceed finite-network capacity.

Optimization lower bound

$$L^*_{joint}(G) - L^*_{node} \geq \Omega(G)$$

OUR SOLUTION

FLAG: factorize structure, don't diffuse it

Stop generating the high-dimensional edge graph. Inject structure through two stable, decoupled priors.

1. Graph as a spatial encoder

Fix the tissue topology with reliable priors (distance + histology). **Encode spot-spot context once** into a compact signal H_spatial — stable in G.

Spot–spot structure, no variance explosion

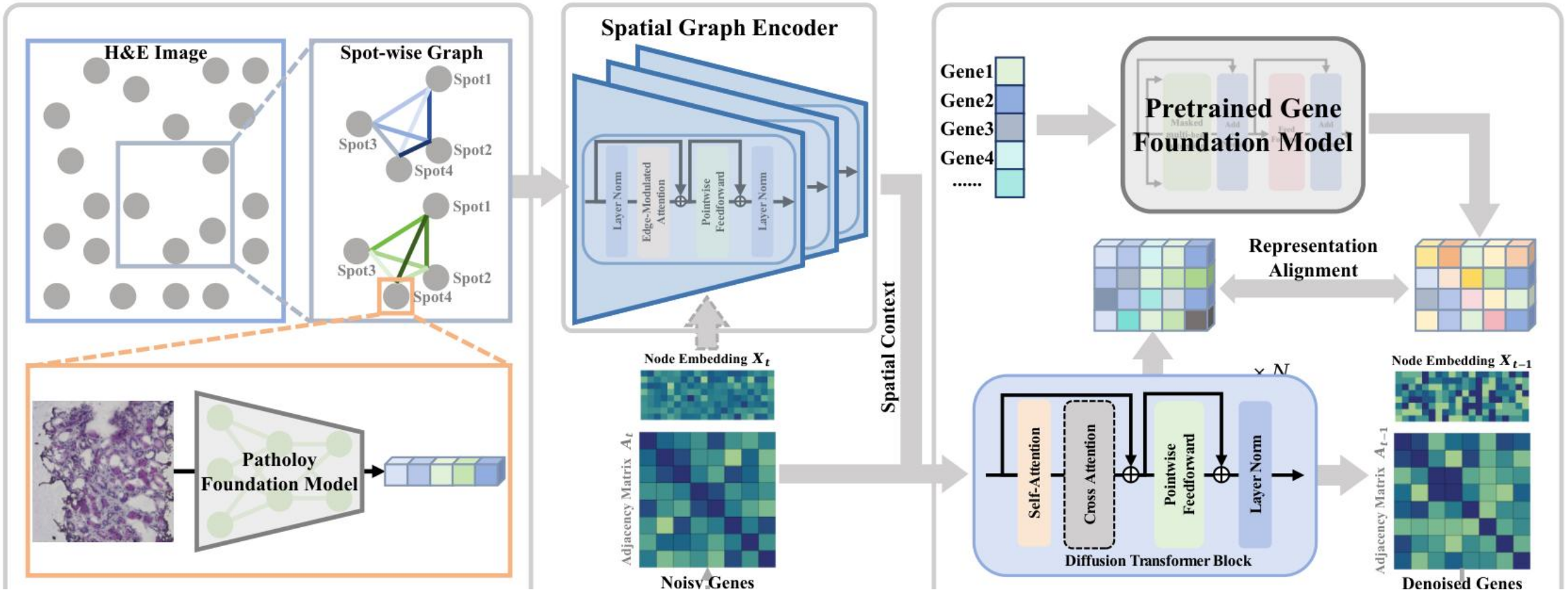
2. Gene Foundation Model alignment

ST slides are too small to learn reliable gene–gene covariance. **Align diffusion features to a frozen GFM** (Geneformer) for robust gene semantics.

Gene–gene structure from large-scale priors

METHOD

The FLAG framework



H&E → graph

Pathology FM encodes tiles; spots form a fixed spatial graph.

Spatial Graph Encoder

Edge-modulated attention aggregates neighborhood context.

Gene DiT + GFM align

Diffusion denoises genes, conditioned & GFM-regularized.

METHOD

One factorization, two complementary views

Best of both attention regimes

- **Graph encoder** → spot-to-spot message passing (spatial).
- **Gene DiT** → gene-to-gene attention within each spot.
- **FLAG** captures both — spatial context conditions gene generation.

Tissue topology is a fixed deterministic prior C_e (distance + histology) — never a sampled latent.

Training objective

$$L_{total} = L_{score} + \lambda \cdot L_{align}$$

L_score — score-matching denoises gene expression under spatial conditioning.

L_align — pulls intermediate DiT features toward frozen GFM gene embeddings.

Result: rich spatial + biological priors without the high-dimensional optimization penalty — spatially and biologically coherent expression fields.

EXPERIMENTS

Setup

3 HEST-1k cohorts

HER2ST

36 slides · 13.6k spots

KIDNEY

23 slides · 25.9k spots

PRAD

23 slides · 62.7k spots

7:2:1 slide-level split · Top-200 HMHVG panel

5 SOTA baselines

Discriminative

HisToGene · BLEEP · TRIPLEX

Generative

Stem · STFlow

Covers deterministic & generative paradigms for a complete landscape.

Dual evaluation

Pointwise accuracy

PCC ↑ · **MSE ↓**

Structural fidelity

GSC ↑ (gene–gene)

SSC ↑ (gene–spatial)

Single NVIDIA H800 GPU.

RESULTS

Competitive accuracy, dominant structural fidelity

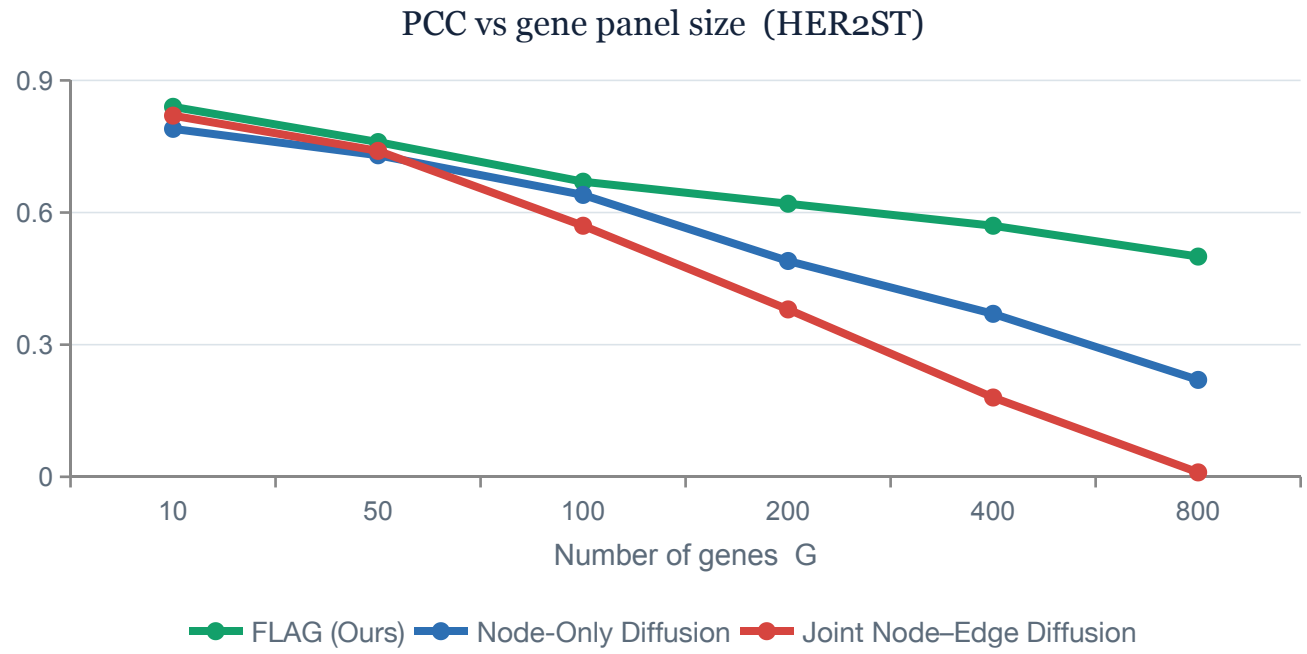
Top-200 HMHVG across three cohorts. FLAG leads on **GSC and SSC everywhere** while staying competitive on PCC / MSE.

Method	PCC ↑	MSE ↓	GSC ↑	SSC ↑
HER2ST				
HisToGene (D)	0.4940	1.8459	0.2065	-0.1549
BLEEP (D)	0.4852	1.1516	0.6988	0.1886
TRIPLEX (D)	0.6913	0.6559	0.5593	0.0708
Stem (G)	0.5772	0.9535	0.8322	0.3810
STFlow (G)	0.7058	0.6769	0.7890	0.2890
FLAG (Ours)	**0.6835**	**0.7342**	**0.8926**	**0.6386**
KIDNEY				
HisToGene (D)	0.2318	1.3935	-0.0264	0.1696
BLEEP (D)	0.1471	2.7602	0.5331	0.0889
TRIPLEX (D)	0.3739	1.1454	0.4469	0.2347
Stem (G)	0.3443	1.3828	0.8451	0.1257
STFlow (G)	0.3145	1.2790	0.6857	-0.1007
FLAG (Ours)	**0.3917**	**1.2112**	**0.8713**	**0.3409**
PRAD				
HisToGene (D)	0.2553	1.9681	0.3810	-0.1277
BLEEP (D)	0.0417	5.6868	0.4338	0.2897
TRIPLEX (D)	0.5267	1.5824	0.5533	0.6343
Stem (G)	0.4025	2.0938	0.8216	0.2768
STFlow (G)	0.5337	1.8776	0.7228	0.5638
FLAG (Ours)	**0.5853**	**1.3771**	**0.8775**	**0.7510**

RESULTS

FLAG breaks the Gene Dimension Curse

Same axes as before — now with FLAG. Baselines collapse; FLAG stays high into large gene panels.



0.50 FLAG PCC at G=800

0.22 Node-Only at G=800

≈0 Joint diffusion at G=800

FLAG scales to biologically comprehensive panels — the others do not.

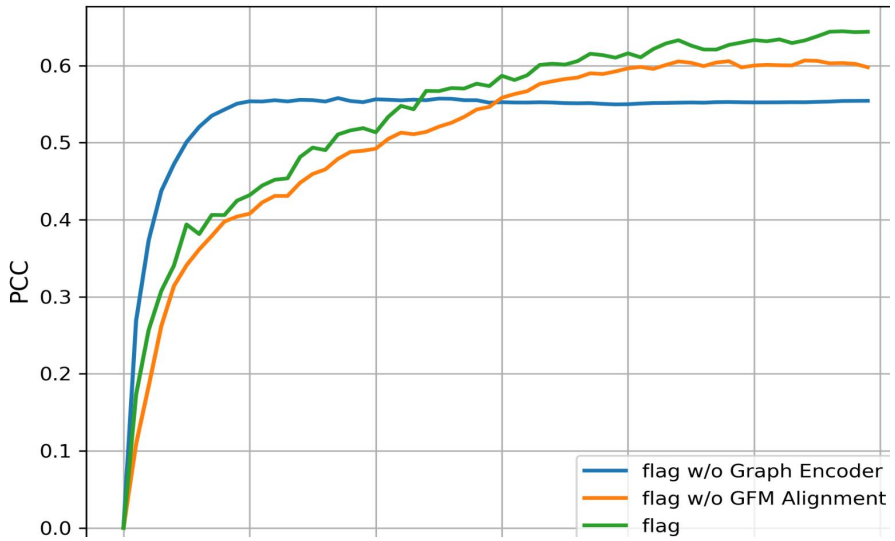
RESULTS

Every component pulls its weight

HER2ST variant	PCC ↑	MSE ↓	GSC ↑	SSC ↑
w/o Diffusion (supervise)	0.6748	0.7864	0.3217	0.5685
w/o GFM Alignment	0.6682	0.7938	0.8713	0.5894
w/o Spatial Graph	0.6297	0.8499	0.9028	0.3399
FLAG (full)	**0.6835**	**0.7342**	**0.8926**	**0.6386**

- **Diffusion** is the generative engine — without it GSC craters (0.32).
- **Spatial graph** drives spatial fidelity — SSC drops to 0.34 without it.
- **GFM alignment** adds accuracy & coherence. All three are synergistic.

Training dynamics (HER2ST, G=200)

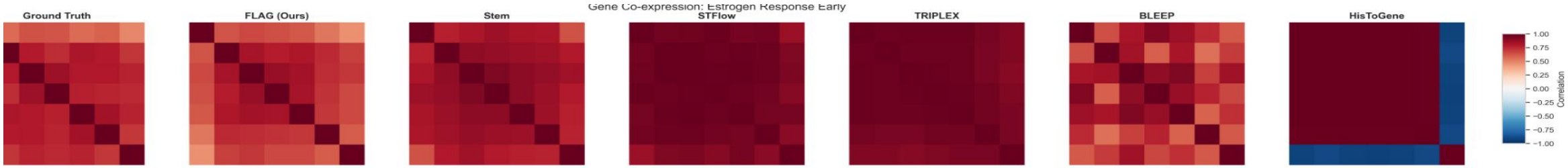


GFM prior gives FLAG a warm start → fastest, highest convergence.

RESULTS

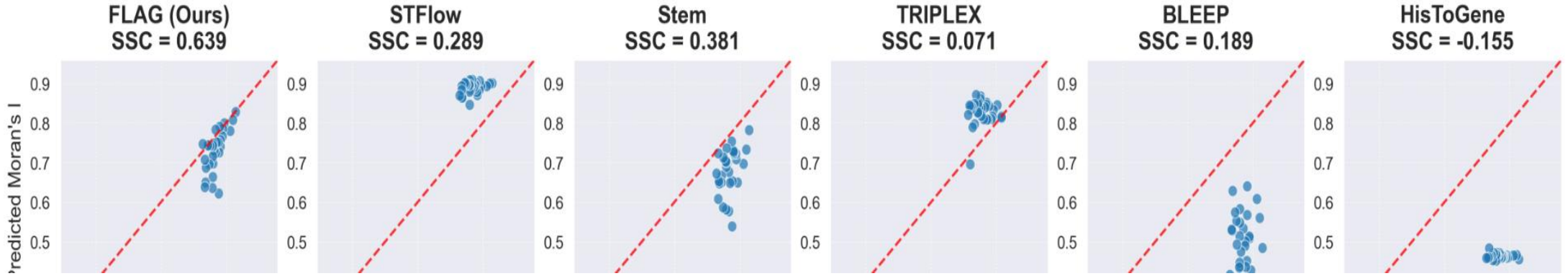
Seeing the structure FLAG recovers

Gene regulatory networks *(GSC — Estrogen-Response-Early co-expression)*



FLAG restores crisp block-diagonal cliques; baselines blur or break them.

Spatial autocorrelation *(SSC — predicted vs. true Moran's I; closer to the diagonal is better)*



FLAG (SSC 0.639) hugs the diagonal; generative baselines over- or under-estimate spatial continuity.

RESULTS

Structural fidelity → real biological discovery

On HER2ST, FLAG's structure-aware predictions translate directly into downstream clinical utility.

0.845
Spatial domain ARI
best baseline 0.674

0.914
Spatial domain NMI
best baseline 0.787

0.500
DEG overlap @Top-50
best baseline 0.469

0.394
DEG overlap @Top-20
best baseline 0.329

Recovers functional biomarkers, not plausible hallucinations. FLAG identifies true marker genes and tissue domains far better than all baselines — and the gains hold on the expert-annotated DLPFC cohort.

CONCLUSION

FLAG, in one line

A structure-aware latent-diffusion framework that predicts spatial gene expression with both pointwise accuracy and biological fidelity.

Diagnosed the Gene Dimension Curse

Theory + experiments on why joint node–edge diffusion collapses (Ω(G) penalty).

Built FLAG

Spatial graph encoder for spot–spot context + GFM alignment for gene–gene fidelity.

New structural metrics

GSC & SSC make biological structure measurable — beyond PCC / MSE.

Thank you!