NemaContext: The Organism as Context

Flow Matching Transformers for Digital Embryogenesis

Senquan Gao

January 26, 2026

Outline

1. Core Philosophy: The Organism as Context
2. The Temporal Coverage Gap
3. Why Not GNNs? The Bitter Lesson
4. Flow Matching Transformers
5. Training Strategy
6. Validation Strategy
7. Implementation & Resources
8. Summary

The Central Thesis

"A cell is not an island.
Its identity, position, and fate are defined not by intrinsic properties alone,
but by its place within the developing whole."

The Name: NemaContext

Nema(tode) + **Context**

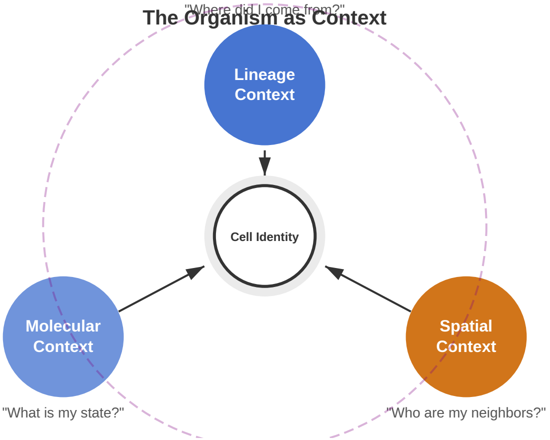
The organism *is* the context that gives meaning to each cell.

The Computational Principle

$$\text{Cell}_i = f(\text{Cell}_i, \text{All Other Cells})$$

Each cell's representation is computed as a function of the **entire embryo**.

Three Levels of Developmental Context



1. Lineage Context

Temporal: Where did this cell come from?

2. Molecular Context

State: What genes is this cell expressing?

3. Spatial Context

Relational: Who are this cell's neighbors?

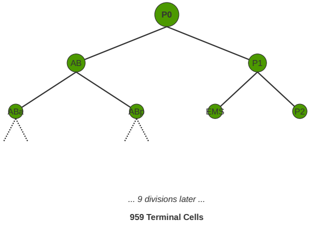
Why *C. elegans*?

The Only Tractable System

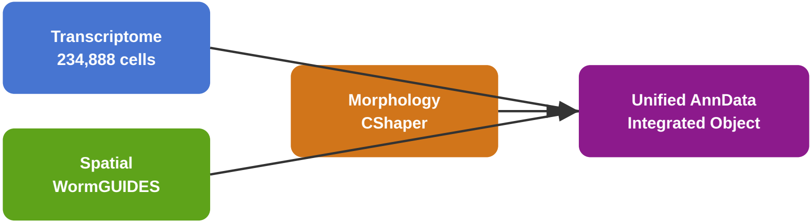
- **Invariant lineage**: 100% deterministic divisions
- **Complete connectome**: Adult wiring known
- **Lineage-resolved transcriptomics**: 234K cells
- **4D morphological atlas**: CShaper, WormGUIDES
- **Small cell count**: 959 terminal cells

Digital Embryogenesis is Feasible

C. elegans is the **only** organism where we can attempt to generate complete embryo trajectories.



Data Landscape: Trimodal Integration



Transcriptome
234,888 cells

Spatial
WormGUIDES

Morphology
CShaper

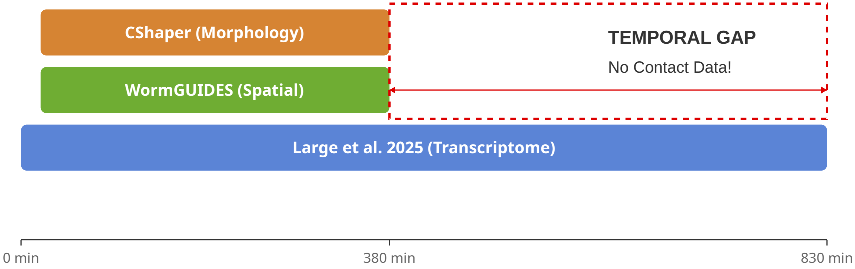
Unified AnnData
Integrated Object

Transcriptome
Large et al. 2025
234,888 cells
0–830 min

Spatial
WormGUIDES
1,341 cells
20–380 min

Morphology
CShaper
1,234 cells
20–380 min

The Temporal Coverage Gap



Problem: For cells in the 380–830 min window, we have:

- ✓ Full transcriptome data (gene expression)
- ✓ Lineage identity (from Sulston tree)
- × No spatial coordinates
- × No contact graph

Our Question:

Can we **infer** spatial context from lineage + molecular context?

The Contact Graph Problem

Why Contact Graphs Matter

- **Notch signaling**: Requires direct cell-cell contact (GLP-1/APX-1, LIN-12/LAG-2)
- **Inductive fate decisions**: Neighbors determine cell identity
- **Tissue organization**: Contacts define morphogenesis

The Inverse Problem

Given: Lineage + Transcriptome (what we know)

Infer: Contact Graph (what we need)

Hypothesis: The organism provides sufficient context. If we know a cell's developmental history and current molecular state, we can predict its spatial relationships.

The GNN Approach (and Why We Reject It)

Standard GNN Paradigm

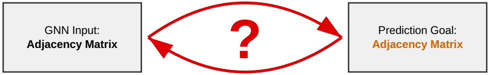
- Assume graph structure is **given**
- Message passing along edges
- Learn node representations

Fundamental Problem

GNNs require the graph as **input**.

But the contact graph is exactly what we're trying to **predict**!

Chicken-and-egg problem.



The Chicken-and-Egg Paradox

The Bitter Lesson

Rich Sutton (2019)

"The biggest lesson that can be read from 70 years of AI research is that **general methods that leverage computation** *are ultimately the most effective, and by a large margin."*

GNN Approach

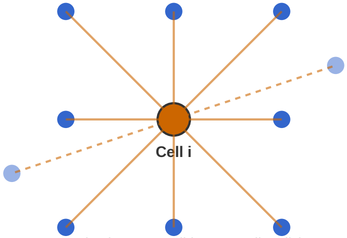
- – Encodes **human knowledge** into architecture
- – Hand-crafted graph topology
- – Message passing = limited context
- – Over-smoothing with depth
- – Poor GPU utilization (sparse ops)

Transformer Approach

- + **Learns from data**
- + No assumed topology
- + Full attention = organism as context
- + Scales with depth
- + Excellent GPU utilization (dense ops)

Transformers Embody “Organism as Context”

Transformer Self-Attention: "Seeing" the Whole Embryo



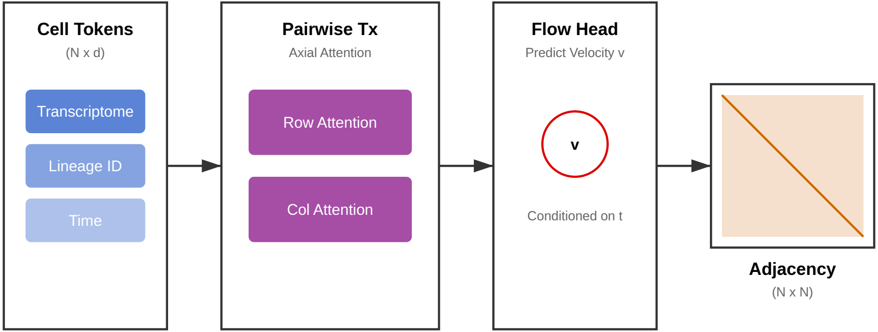
Attention(Query=Cell i, Keys=All Cells)

Context is Global, not Local

The Mathematical Formalization

$$\text{Cell}_i^{\text{repr}} = \sum_{j \in \text{Embryo}} \text{Attention}(Q_i, K_j) \cdot V_j$$

Architecture Overview



Input
Cell tokens
(Transcriptome + Lineage + Time)

Encoder
Pairwise Transformer
(Axial Attention)

Output
Contact Graph
(Generated via Flow Matching)

Cell Tokenization

Each Cell = One Token

$$\text{Token}_i = [\underbrace{\mathbf{e}_i^{\text{expr}}}_{\text{scGPT}} \parallel \underbrace{\mathbf{e}_i^{\text{lin}}}_{\text{Binary Path}} \parallel \underbrace{\mathbf{e}_i^{\text{time}}}_{\text{Sinusoidal}} \parallel \underbrace{\mathbf{e}_i^{\text{morph}}}_{\text{CShaper}}]$$

Transcriptome Embedding

- scGPT foundation model (768-dim)
- Or: PCA + MLP (lightweight)
- Captures molecular state

Temporal Encoding

- Sinusoidal (Transformer-style)
- Developmental time: 0–830 min
- Captures temporal position

Lineage Encoding

- Binary path from zygote
- Example: "ABplp" → [0,1,0,1,0,...]
- Encodes developmental history

Morphology (when available)

- Volume, surface area, sphericity
- From CShaper (early embryo)
- Imputed for late embryo

Flow Matching: Generative Graph Modeling

Why Flow Matching?

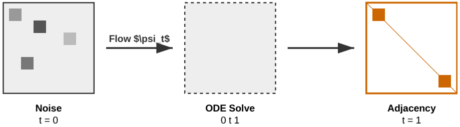
- **Generative**: Produces graphs, not just scores
- **Deterministic sampling**: Faster than diffusion
- **Stable training**: No score matching issues
- **Structured outputs**: Natural for adjacency matrices

The Formulation

Transform noise $\mathbf{Z} \sim \mathcal{N}(0, 1)$ to adjacency $\mathbf{A}$:

$$\mathbf{Z} \xrightarrow{\text{Flow } \psi_t} \mathbf{A}$$

Conditioned on: (transcriptome, lineage, time)



Training Objective: OT-CFM

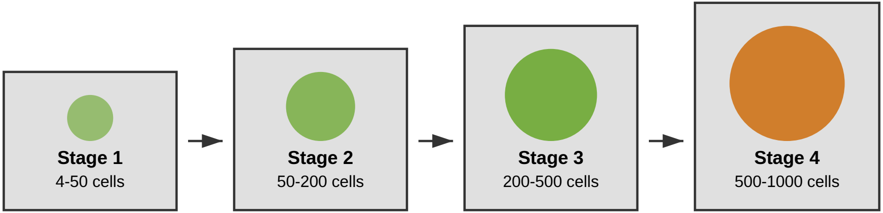
Optimal Transport Conditional Flow Matching

1. Sample random flow time $t \sim U(0,1)$
2. Sample noise $\mathbf{A}_0 \sim \mathcal{N}(0,1)$
3. Interpolate: $\mathbf{A}_t = (1-t)\mathbf{A}_0 + t\mathbf{A}_{\text{target}}$
4. Predict velocity: $\mathbf{v}_\theta(\mathbf{A}_t, t, \text{context})$
5. Loss: $\mathcal{L} = \|\mathbf{v}_\theta - (\mathbf{A}_{\text{target}} - \mathbf{A}_0)\|^2$

Inference: Generate Contact Graph

1. Start from noise: $\mathbf{A}_0 \sim \mathcal{N}(0,1)$
2. Euler integration: $\mathbf{A}_{t+\Delta t} = \mathbf{A}_t + \mathbf{v}_\theta \cdot \Delta t$
3. Binarize: $\mathbf{A}_{\text{final}} = \mathbf{1}[\mathbf{A}_1 > 0.5]$
4. Symmetrize

Curriculum Learning



Stage 1
4–50 cells
Simple topology
Fast convergence

Stage 2
50–200 cells
Gastrulation
Cell migration

Stage 3
200–500 cells
Organogenesis
Tissue formation

Stage 4
500–1000 cells
Differentiation
Complex topology

Progressive training on increasingly complex embryo stages

Temporal Extrapolation

The Key Challenge

- **Training data**: Early embryo (20–380 min) with ground truth contacts
- **Prediction target**: Late embryo (380–830 min) with NO ground truth

Our Hypothesis

The rules of contact formation are **learnable** and **generalizable**:

- Cells with complementary adhesion molecules contact
- Lineage proximity predicts spatial proximity (with caveats)
- Morphological constraints limit possible contacts

These rules apply across developmental time.

Uncertainty Quantification

Generate multiple samples → Compute variance → Report confidence

Validation Without Ground Truth

1. Cross-Validation (Early Embryo)

- Leave-one-stage-out
- Train on stages 1,2,3 → Test on 4
- Metrics: AUC-ROC, Average Precision

2. Connectome Consistency

- Adult synapses require prior contact
- If neurons A-B synapse in adult...
- ...model must predict A-B contact in embryo
- **Peter's Rule**: Contact is necessary for synapse

3. Notch Signaling Logic

- Notch requires direct contact
- Check: Predicted neighbors have L-R pairs?
- GLP-1/APX-1, LIN-12/LAG-2
- Known developmental inductions

4. Cross-Species (C. briggsae)

- Conserved lineage, divergent genome
- Predicted patterns should be similar
- Evolution validates predictions

Computational Requirements

Model Size

Component	Parameters
Cell Encoder	∼10M
Pairwise Transformer	∼50M
Flow Network	∼30M
Total	**∼90M**

Scalability

Stage	Cells	Memory
Early	50–200	∼4 GB
Mid	200–500	∼16 GB
Late	500–1000	∼48 GB

Hardware

- A100 80GB × 1–2 GPUs
- Training: ∼24 hours total
- Flash Attention for efficiency

Inference

- 1000-cell embryo: ∼30s/sample
- 10 samples (uncertainty): ∼5 min
- Batched across time points

Current Progress

Completed

- ✓ Trimodal data integration
- ✓ CShaper contact extraction
- ✓ 40% morphology coverage (94K cells)
- ✓ Lineage proximity prior (28.5M edges)
- ✓ GPU-accelerated pipeline
- ✓ Unified AnnData structure

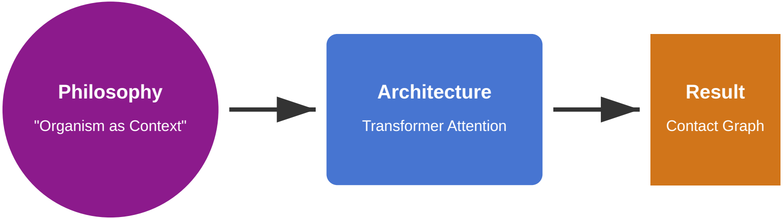
Next Steps

1. Implement Flow Matching Transformer
2. scGPT embedding integration
3. Curriculum training pipeline
4. Validation framework
5. Late-stage prediction

Key Metrics

234,888 cells — 27,138 genes — 1.85M contact edges — 28.5M proximity edges

The Synthesis



Key Insight

The Bitter Lesson provides the **technical** justification.
“Organism as Context” provides the **biological** justification.

Transformers are not an arbitrary choice —
they are the computational formalization of developmental biology.

The Question We Answer

*"Given everything we know about a cell's history (lineage)
and current state (transcriptome),
can we infer its spatial relationships (contacts)
by understanding its place in the developing organism?"*

Our hypothesis: Yes.

Because the organism is the context.

Thank You

Questions?

`github.com/maxwell-gao/NemaContext`