NemaContext: Generative Modeling of Complete Embryo Development

Progress Report: Data Integration & Contact Graph Prediction

Progress Report

January 25, 2026

Outline

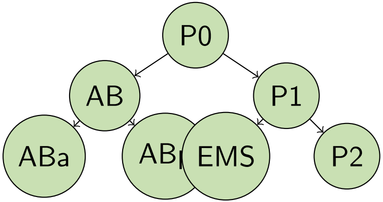
1. Project Vision
2. Data Integration Progress
3. The Contact Graph Problem
4. Scientific Significance
5. Technical Implementation
6. Next Steps

The Grand Challenge: Digital Embryogenesis

Goal: Generate complete embryo states from a single zygote

Input → Output

- **Input**: Zygote state at $t = 0$ (1 cell)
- **Output**: Complete embryo at any time t
  - Every cell's transcriptome
  - Every cell's 3D position
  - Cell-cell neighbor relationships
  - Lineage tree structure



...

959 cells

Why *C. elegans*?

Unique Properties

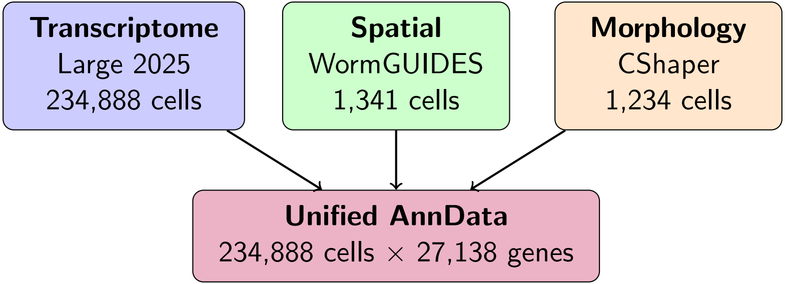
- **Invariant lineage**: 100% deterministic cell divisions
- **Complete spatial tracking**: WormGUIDES 4D atlas
- **Lineage-resolved transcriptomics**: Large et al. 2025
- **Small cell count**: 959 cells (tractable)
- **Extensive ground truth**: Decades of research

Cell as Token Paradigm

- Each cell → A token
- Development → Tree generation
- Cell division → Token splits into two
- Zygote → Adult: 1 → 959 tokens

C. elegans is the **only** *multicellular organism where complete generative modeling is feasible.*

Trimodal Data Integration



Transcriptome
Large 2025
234,888 cells

Spatial
WormGUIDES
1,341 cells

Morphology
CShaper
1,234 cells

Unified AnnData
234,888 cells × 27,138 genes

CShaper Integration: Key Achievement

What is CShaper?

4D morphological atlas of *C. elegans* embryo (Cao et al. 2020)

- **Cell-cell contact matrices**: Physical neighbor relationships
- **Cell morphology**: Volume, surface area, sphericity
- **Standardized coordinates**: Averaged across 46 embryos

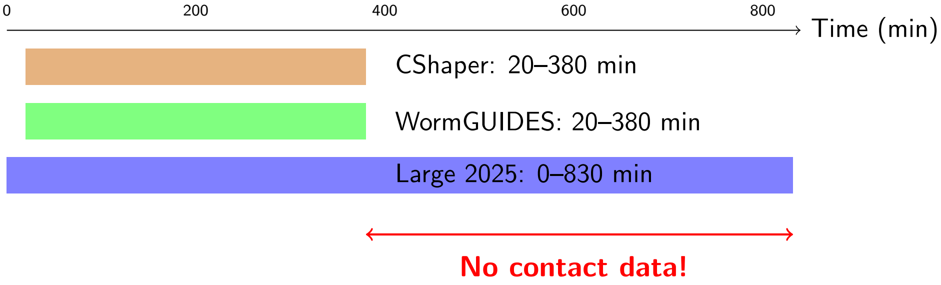
Integration Results

Metric	Value
Cells with morphology	94,005 (40%)
Direct matches	25,264
Ancestor-mapped	1,390
Contact edges	1,854,781

Matching Strategies

1. **Direct**: Cell exists in CShaper
2. **Fuzzy**: Handle ‘x’ and ‘/’ in lineage
3. **Ancestor**: Map to closest ancestor
4. **Expression**: Validate with gene expression

The Temporal Coverage Gap



Problem: We have transcriptomes for late-stage cells (380–830 min), but **no spatial/contact information**!

Question: Can we **predict** contact relationships for late-stage cells?

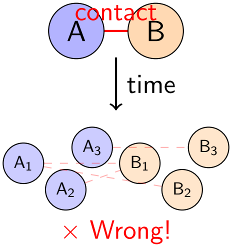
Initial Approach: Ancestor Mapping (Incorrect)

Naïve Approach:

- Map late-stage cells to their early ancestors
- If ancestors A and B contacted → all descendants of A contact all descendants of B
- Result: 50 million edges!

Biologically incorrect!

- Ancestor contact $\neq$ descendant contact
- Cells migrate, tissues reorganize
- Late-stage spatial arrangement differs from early



Correct Approach: Two Distinct Graphs

contact_graph (Ground Truth)

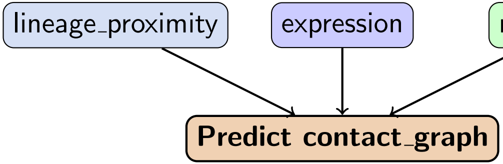
- **Direct matches only**
- True physical contacts
- 1.85M edges, 4,050 cells
- Supervision signal for training

lineage_proximity (Prior)

- Based on ancestor relationships
- Decays with lineage distance:

$$\text{proximity} = \frac{\text{ancestor\_contact}}{d_i + d_j + 1}$$

- 28.5M edges, 19,939 cells
- Feature for prediction



lineage_proximity

expression

morphology

Predict contact_graph

Why Predict Contact Graphs?

1. Cell Fate Depends on Neighbors

- **Notch signaling**: Requires direct cell-cell contact
- **Induction**: Classic experiments show neighbors influence fate
- **Lateral inhibition**: Adjacent cells adopt different fates

2. Enables Complete Spatial GNN

- Spatial GNN requires neighbor graph
- Early cells: Use true `contact_graph`
- Late cells: Use predicted contacts
- ⇒ **Complete spatial modeling** from zygote to adult

3. Scientific Hypothesis

Developmental history (ancestor contact) + current state (expression, morphology) can predict current spatial neighbors.

Link Prediction as Machine Learning Task

Task Formulation:

Given:

- Node features: expression + morphology + lineage
- Edge prior: `lineage_proximity`
- Labels: `contact_graph` (where available)

Learn:

$$P(\text{contact}_{ij} | \text{proximity}_{ij}, \mathbf{x}_i, \mathbf{x}_j)$$

Predict: Contacts for late-stage cells

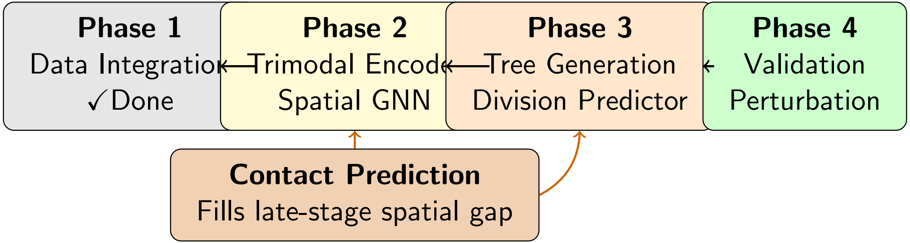
Data Split

Set	Cells
Training	3,790
Prediction	16,149
No lineage	214,949

Training cells: Have both true contacts AND lineage proximity

Prediction cells: Have lineage proximity but no true contacts

Broader Impact: Complete Developmental Modeling



Phase 1
Data Integration
✓Done

Phase 2
Trimodal Encoder
Spatial GNN

Phase 3
Tree Generation
Division Predictor

Phase 4
Validation
Perturbation

Contact Prediction
Fills late-stage spatial gap

Key Contribution: Contact prediction enables complete spatial modeling across all developmental stages, not just the CShaper coverage window.

Implementation Highlights

GPU Acceleration

- PyTorch for expression similarity
- 234K × 231 correlation: 6 seconds
- A100 GPU utilized for matrix ops

Optimized Data Structures

- Sparse matrices for graphs
- Cached consensus morphology
- Vectorized operations

AnnData Structure

`adata.X` – Expression

```
adata.obsm['X_spatial']
adata.obsm['X_lineage_binary']
adata.obsp['contact_binary']
adata.obsp['lineage_proximity']
adata.obs['has_true_contact']
adata.obs['has_lineage_proximity']
```

Processing Time

Full pipeline (234,888 cells): ∼**9 minutes** on A100 GPU

Immediate Next Steps

1. **Implement Link Prediction Model**
    - GNN-based edge predictor
    - Input: node features + lineage proximity prior
    - Output: contact probability
2. **Validate Predictions**
    - Cross-validation on early-stage cells
    - Literature validation (known tissue neighbors)
    - Connectome consistency (neurons that synapse should be neighbors)
3. **Integrate with Spatial GNN**
    - Use predicted contacts as edges
    - Aggregate neighbor information for cell state prediction

Long-Term Vision

"From a single zygote, generate the complete developmental trajectory of an organism — every cell's state, position, and neighbor relationships, from fertilization to adulthood."

Contact graph prediction is a critical step toward this vision:

- Bridges the temporal gap in spatial data
- Enables complete cell-cell communication modeling
- Tests the hypothesis that spatial organization is predictable from developmental history

Key Question

Is the spatial organization of C. elegans as deterministic as its cell lineage?

Summary

Completed

- ✓ CShaper data integration
- ✓ 40% morphology coverage
- ✓ True contact graph (direct matches)
- ✓ Lineage proximity prior
- ✓ Training/prediction split
- ✓ GPU-accelerated pipeline

Key Insight

- Ancestor contact $\neq$ descendant contact
- But ancestor contact → **prior** for prediction
- Link prediction: Learn the mapping

Scientific Contribution

Testing whether spatial organization follows from developmental history + cell state

Thank You

Questions?