https://doi.org/10.1038/s41592-026-03055-4

Towards predictive virtual embryos with genomics and AI

Natalie Cao, Yifan Lu & Xiaojie Qiu

Predictive virtual embryo systems that integrate single-cell and spatial data with artificial intelligence (AI) techniques offer a promising avenue for modeling mammalian embryogenesis across scales and could advance our fundamental understanding of development and congenital disease.

Even with decades of progress in developmental biology, building an integrated and predictive model of embryogenesis remains an unsolved problem — one that predictive virtual embryo systems aim to address. Bridging molecular regulation, cellular dynamics and tissue-level organization requires a framework that accounts for developmental phenomena across spatial and temporal scales. This complexity is inherent to embryogenesis itself, which unfolds as a tightly spatiotemporally choreographed process, starting from a single fertilized egg and proceeding through cell divisions, migrations and differentiation events to yield trillions of highly specialized cells. These processes are orchestrated with exquisite precision to form the various tissues and organs of a functional organism in 3D space. The successful emergence of a functional embryo relies on tightly regulated genetic programs, dynamic cell behaviors (migration, growth, division or death) and spatially coordinated signaling interactions between various spatially patterned cell types.

Why we need a virtual embryo

Genome-wide association studies and animal models have shown that thousands of genetic, metabolic and cellular perturbations can disrupt gene expression, cell-fate commitment and organ morphogenesis. The resulting developmental disorders affect millions of newborns globally, including the roughly 1 in 33 babies in the United States who are born with a congenital disability, a leading cause of infant mortality. Understanding how embryos develop and why these processes go awry remains an urgent challenge in biology and medicine. Real-time monitoring and predictive modeling of embryogenesis would not only deepen fundamental understanding of developmental biology by revealing the underlying organizing principles that govern embryogenesis, particularly the spatiotemporal coordination of cell fate decisions and the conserved molecular-cellular mechanisms that drive its tightly regulated progression, but also support wide applications, from drug toxicity screening to the design of targeted genetic or cellular interventions and the guidance of regenerative medicine strategies such as tissue engineering and stem-cell therapies. To this end, we envision a dynamic 'virtual embryo' framework that integrates genomics data and machine learning technologies to model normal development and predict perturbations.

The vision of a virtual embryo

We define the virtual embryo as a fully data-backed digital twin of embryogenesis, capable of capturing multiscale dynamics from the molecular regulation of single cells to tissue patterning and organ formation[1–3]. At its core, the virtual embryo bridges high-dimensional, spatiotemporal single-cell atlases of mammalian development with advanced machine learning techniques, including graph neural network models and foundation models. This framework enables the simulation and prediction of embryogenesis under normal or genetically and environmentally perturbed conditions, thus allowing the identification of molecular mechanism for developmental disorders and the mitigation of them by compensating for these perturbations[3]. Importantly, by incorporating parental and embryonic genetic information measured through sequencing of non-invasively obtained cell-free DNA, fetal circulating trophoblasts or nucleated red blood cells, we can personalize the virtual embryo for each couple and each future baby.

Unlike earlier biophysical simulations constrained by fixed, predefined equations, the virtual embryo leverages semi- and self-supervised learning approaches to extract fundamental developmental rules directly from data. This shift from hand-crafted modeling to data-driven inference unlocks a broad spectrum of applications, ranging from simulating normal embryogenesis and predicting variant effects to assessing drug toxicity and exploring microenvironmental influences on morphogenesis.

While the Virtual Cell initiative focuses on intracellular regulatory complexity, the Virtual Embryo extends this limited view to spatial, temporal and multimodal dimensions at the systems level. By unifying molecular programs, biomechanical dynamics and morphological transformations into a predictive framework, we will approach the long-sought 'holy grail' of developmental biology. Beyond offering a mechanistic explanation, we believe that Virtual Embryo holds transformative potential to uncover the fundamental organizing principles of life by enabling systematic investigation of emergent multicellular coordination and developmental robustness, key aspects of basic developmental biology that underpin tissue assembly and functional multicellularity — ultimately delivering a robust framework for diagnosis, prognosis and therapeutic intervention in congenital and developmental disorders.

Methodological foundations

Recent technological breakthroughs provide the foundation for building a virtual embryo framework. Its foundation rests on complementary data pillars: 3D whole-embryo-level spatial transcriptomics, time-resolved single-cell multi-omics, live imaging and single-cell lineage tracing (Fig. 1a), and engineered embryonic models such as blastoids, gastruloids and synthetic embryos (Fig. 1b). Together, these technologies have generated multidimensional datasets that capture mammalian embryogenesis with unprecedented spatiotemporal resolution (Fig. 1c).

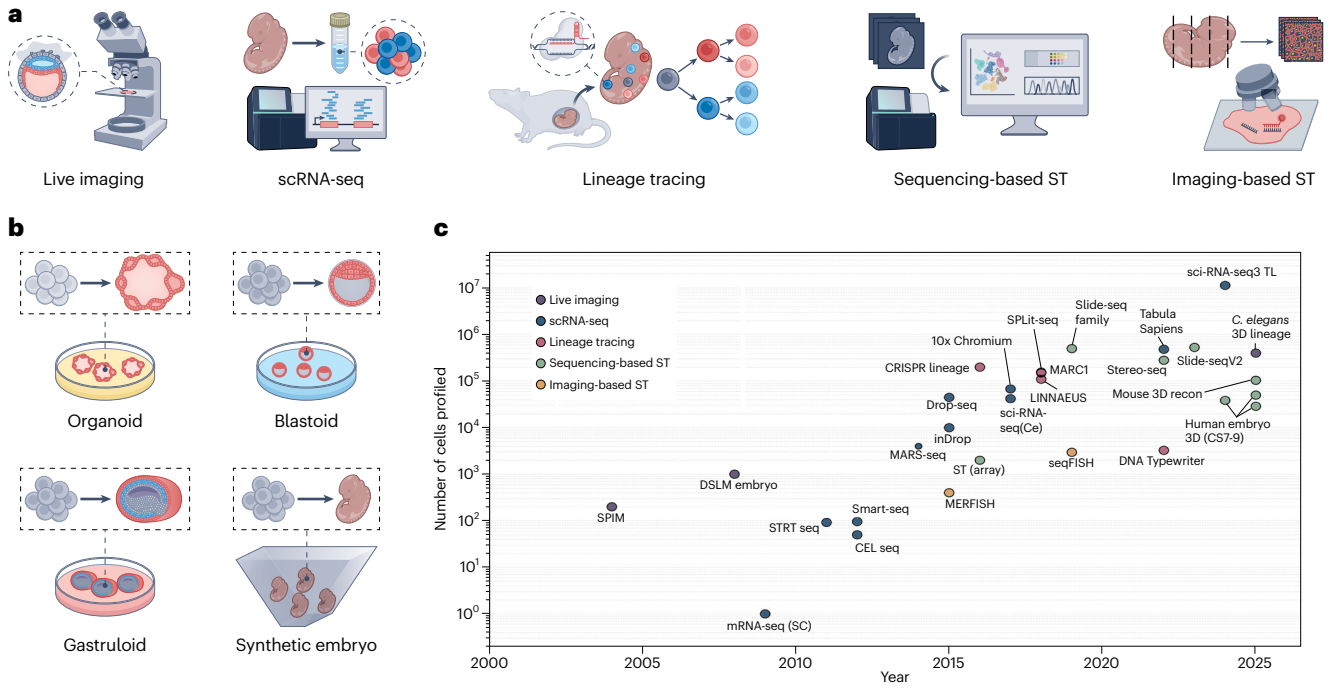

Fig. 1 | Existing technologies and embryonic models lay the foundation for the construction of virtual embryo models. a, Five core technology categories underpinning embryo modeling: live imaging, scRNA-seq, lineage tracing, sequencing-based spatial transcriptomics (ST) and imaging-based spatial transcriptomics. **b**, Engineered embryo-like systems — including organoids, blastoids, gastruloids and synthetic embryos — as in vitro platforms for perturbation and validation. **c**, Timeline of landmark studies, plotted by publication year (*x* axis) and number of cells profiled (*y* axis, log scale). Dots are color-coded by the categories in **a**, with circle size reflecting approximate relative scale; closely related approaches are grouped.

Data foundations. The accumulation of large-scale single-cell atlases across species has transformed developmental biology into a data-rich discipline. The CZI CELLXGENE portal currently hosts more than 26 million human prenatal cells, alongside comparable atlases for other species, such as mouse, *Drosophila*, zebrafish and *Caenorhabditis elegans*. A landmark study recently profiled 12.4 million single cells from 83 embryos from embryonic day (E) 8 to birth, yielding one of the most comprehensive temporal maps of mammalian organogenesis to date[4].

To complement these temporal atlases, spatial genomics has facilitated direct 3D visualization of the embryo. Stereo-seq[2] and Slide-seq enable spatial transcriptomic analysis at, respectively, subcellular and near-single-cell resolution to reconstruct entire mouse embryos and reveal spatially resolved gene-expression gradients and tissue compartmentalization. Stereo-seq's high-field-of-view DNA nanoball array technology has been recently used[5–8] for 3D molecular reconstructions of human embryos at Carnegie stages 7–9 that characterize body-plan patterning at early stages, dual hindbrain trajectories and germ-cell localization[5–7], providing insights into human organogenesis and morphogenesis[5–8]. For mouse embryonic stages, the Mouse Organogenesis Spatiotemporal Transcriptomic Atlas (MOSTA) spans E9.5–E16.5 embryonic development, though with only a few sections per time point[2]. Finally, our group and others have built to our knowledge the largest whole-embryo spatial transcriptome to date[1], encompassing 8 million cells at both E9.5 and E11.5. With this massive dataset, we 3D reconstructed a molecular hologram of later mouse embryos, modeled cross-scale gene-expression gradients and intercellular interactions, and predicted cell morphogenesis across time[1].

Beyond static snapshots, new imaging and lineage-tracing analyses allow us to comprehensively and dynamically follow morphogenesis. Light-sheet microscopy sequentially tracks the course of every individual cell during postimplantation development[9], while CRISPR–Cas9-based time-course lineage tracing prospectively records cell divisions and fate decisions in vivo. Together with single-cell RNA sequencing (scRNA-seq), these methods recover the manifold of developmental trajectories describing cell-state transitions and lineage bifurcations over time[9]. By integrating multimodal datasets, including spatial transcriptomes, live imaging and lineage barcodes, we can achieve a comprehensive, quantitative characterization of embryogenesis.

Engineered embryo-like models such as blastoids, embryoid bodies, gastruloids and synthetic embryos have also been used as in vitro experimental systems. These models enable controlled perturbations to study morphogen gradients, test causal hypotheses and model congenital disorders that would be ethically or practically inaccessible in human embryos. Collectively, all these available data and the new data to be generated promise to build us a multimodal 'Google Earth-like' atlas of developmental biology at single-cell to organismal levels, powering up computational frameworks to create the eventual virtual embryo model.

Computational advances. Parallel computational breakthroughs now support embryo-scale modeling (Fig. 2). We have developed 3D reconstruction pipelines that apply Bayesian mixture modeling, non-rigid registration, global refinement and mesh correction to mitigate tissue

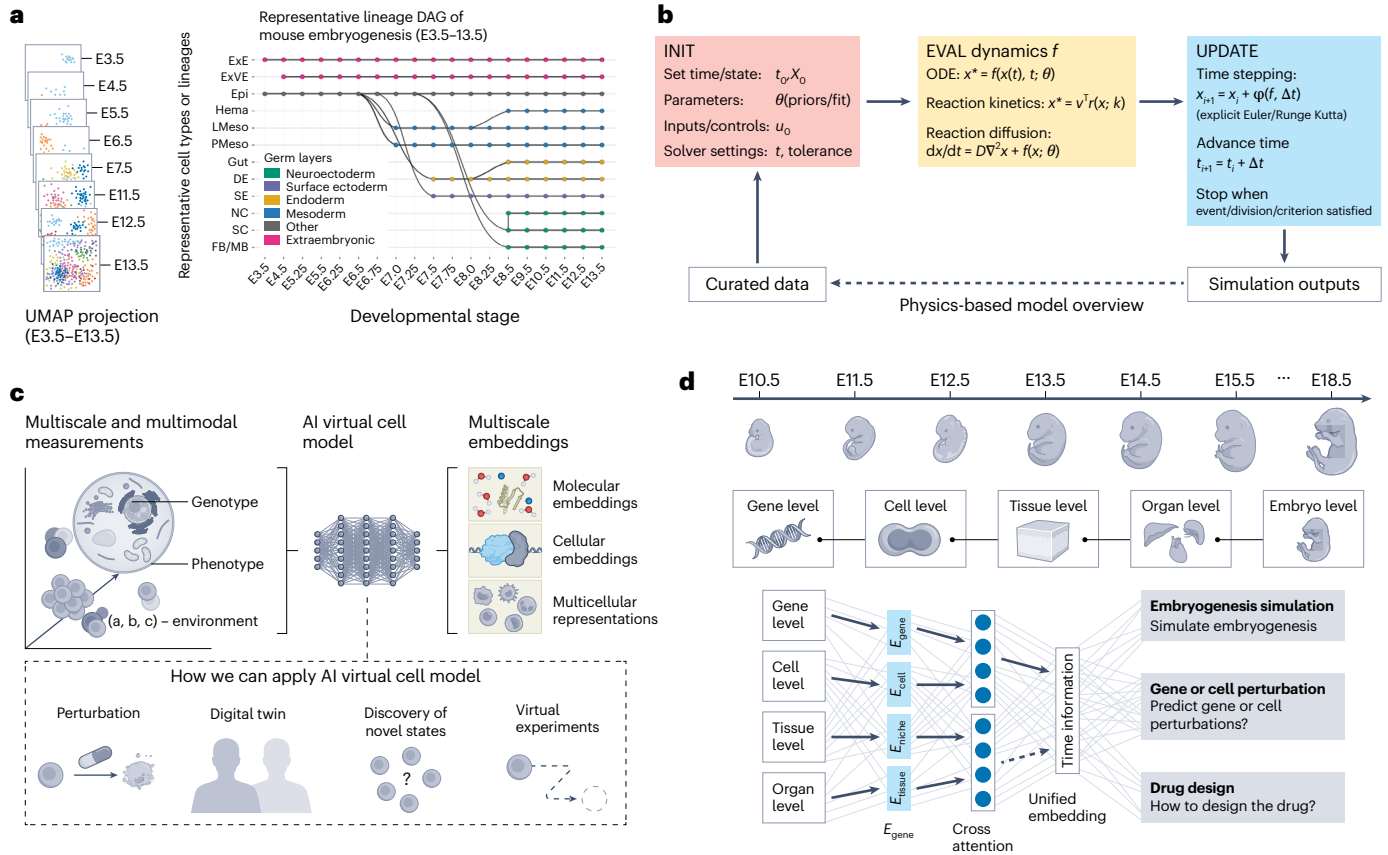

Fig. 2 | Three types of computational approaches that move us toward the virtual embryo. a, Descriptive approaches integrating statistically the embryonic datasets illustrated by UMAP projection of mouse embryonic cells (left) and representative developmental trajectories across germ layers (right). DAG, directed acyclic graph. **b**, Rule-based or physics-based approaches modeling the dynamics of biological systems using mechanical equations[13,14], illustrated by schematic modeling workflows. ODE, ordinary differential equation. **c**, Deep learning-based approaches integrating multimodal measurements into foundational AI Virtual Cell models for cellular representation learning[15]. **d**, Virtual Embryo Platform: a multiscale systems-level modeling paradigm spanning gene, cell, tissue, organ and whole-embryo levels across developmental stages, incorporating single-cell and spatial genomics data into a spatiotemporal foundation model of mammalian embryogenesis and coupled with an AI-agent interface for predictive simulation of developmental dynamics across scales.

deformation and create coherent volumetric maps of whole embryos[1]. This framework, together with TOME[10] and related systems, provides essential descriptive analyses of embryogenesis over time and space (Fig. 2a). However, these models remain largely statistical, capable of describing the dynamics of developmental processes but not of predicting developmental outcomes, unlike prior biophysical simulations such as the Whole Cell model (Fig. 2b).

Given the lack of known kinetic equations in mammalian morphogenesis and the incomprehensibility of multicellular systems to mechanism-based modeling, there is an increasing trend toward data-intensive deep learning-based methods. Large-scale single-cell foundation models such as Geneformer[11] and scGPT[12], trained on tens of millions of profiles, have started to provide a unifying computational basis for embryo-scale simulation (Fig. 2c). These transformer-family models learn low-dimensional representations of genetic dependencies and context-dependent cellular embeddings to predict cell type, gene regulatory networks and responses to perturbations. Simultaneously, flow-matching algorithms, inspired by optimal transport theory, learn vector fields that link cell states over developmental time, enabling interpolation of intermediate states and estimation of continuous migration paths.

Our spatiotemporal modeling framework, Spateo[1], exemplifies this revolution: it simulates the migration behavior of cells through the reconstruction of high-resolution vector fields representing cellular motion and measures differential motion geometries such as curvature, divergence and torsion of tissue deformation. This is then correlated with differential morphogen gradients and transcriptional programs. By linking physical tissue deformation to underlying molecular programs, Spateo connects the spatial and temporal realms of tissue development by predicting morphogenetic trajectories and collective cell behaviors through vector-field-based dynamical modeling of underlying regulatory and signaling networks.

Together, these advances in data generation and computational approaches establish the methodological basis for the virtual embryo – a digital twin capable of simulating organogenesis with sufficient resolution to capture gene expression dynamics and physical migration patterns ranging from single-cell to organ-scale (Fig. 2d). Such dynamic, predictive and generative models will enable in silico

perturbation experiments and developmental forecasting, moving the field beyond descriptive mapping toward true predictive biology.

The broader landscape: how Virtual Embryo complements and extends Virtual Cell projects

The development of Virtual Embryo is closely intertwined with parallel efforts in large-scale organismal datasets and Virtual Cell initiatives. Expanding single-cell and spatial genomics resources have fueled the emergence of foundation models for single-cell biology, such as Geneformer[11], scGPT[12] and others. These models learn universal gene and cell embeddings that capture molecular programs, cellular identities and state transitions with high fidelity. By distilling complex regulatory landscapes from transferable representations for numerous cell types across species, they provide a unifying computational backbone for the renaissance of virtual cells. Complementary community efforts, such as the Virtual Cell Challenge (https://virtualcellchallenge.org/), have further set stringent benchmarks for generative virtual cell modeling, including tasks like cross-condition forecasting and perturbation simulation, laying the groundwork for cellular-scale modeling with unprecedented resolution.

While Virtual Embryo builds on these advances, it is conceptually distinct from but simultaneously complementary to Virtual Cell (Fig. 2c,d). Embeddings from Virtual Cells can be transferred or fine-tuned within Virtual Embryo frameworks, enabling the modeling of multicellular interactions, morphogenetic flows and organ-level development. In contrast to Virtual Cell's focus on intracellular regulatory complexity, Virtual Embryo emphasizes collective emergent behaviors: how ensembles of diverse cells coordinate through biochemical, mechanical and geometric coupling to form robust tissues and organs despite stochastic cellular variation.

By modeling embryos across time as a whole, a virtual embryo will allow hypothesis testing at scales previously inaccessible. For example, it will be able to simulate the impact of any genetic mutations and cellular perturbation on cardiac morphogenesis or forecast teratogenic drug responses such as thalidomide-induced limb malformations and other developmental defects. In doing so, Virtual Embryo goes beyond Virtual Cell to bridge microscopic molecular dynamics with systems-level developmental biology, offering a framework for jointly interrogating gene-regulatory networks, cellular mechanics and tissue- and organ-level morphogenesis.

Key challenges and limitations

Despite progress in genomics and machine learning, building the virtual embryo faces hurdles in data generation, computational modeling and validation.

Data integration and comprehensiveness is one of the major challenges. To model a virtual embryo that comprehensively captures embryogenesis, it is essential to collect data across multiple modalities — such as transcriptomics, epigenetics, imaging, and lineage tracing — and across different embryos. Merging these into a cohesive model demands careful standardization and alignment both spatially (to register cells to a standard embryo reference) and temporally (to ensure accurate developmental temporal coupling of cells across time points). Furthermore, current coverage of ideal time-resolved 3D datasets remains narrow, and joint profiling across modalities in 3D at the whole-embryo scale remains technically challenging. Synthetic embryo models offer partial solutions for comprehensively profiling multimodal data over time, but are constrained by regulations, such as the 14-day rule, that limit in vitro modeling of human development. Cross-species approaches leveraging evolutionary conservation in organisms like the mouse and the rhesus macaque remain indispensable.

Another challenge is model scaling and complexity. A simple embryo, even in a mouse model like the one we described, involves hundreds of millions of cells working together at birth. Simulating all these cells in 3D over time, with every cell following complex gene regulatory networks and responding to biochemical and mechanical cues at high temporal resolution, requires both algorithmic innovation and substantial computational resources. Strategies such as multiscale modeling (for example, modeling cells at the tissue or coarse level instead of modeling every single cell in detail) could be necessary for managing complexity (Fig. 2d).

A further challenge is the biological interpretability and validation of the model's predictions. An AI-driven model could make a prediction (for example, that perturbing gene *Tbx5* in a particular time frame causes heart malformation) that can only be comprehended if the model is interpretable and not a 'black box'. Developmental biology is complex, and ensuring the virtual embryo's reasoning aligns with known biology (or highlights interesting deviations) will be critical for practical and clinical applications. Unlike protein structure prediction, where the Critical Assessment of Structure Prediction (CASP) benchmarks drove the field, embryogenesis lacks widely accepted model-assessment criteria. Experimental anchoring is further constrained by the scarcity of systematic in vivo perturbation datasets, though initiatives such as the Mouse Perturbation Atlas and nematode lineage perturbation studies give promising starts. To accelerate progress, we propose annual Virtual Embryo Challenges to promote community standards and rigorous benchmarking.

Future opportunities

We believe the future of Virtual Embryo lies in open, collaborative ecosystems built on multispecies, multiscale developmental datasets. Once the Virtual Embryo model is integrated with AI agents, we can create a Virtual Embryo Platform that would allow us to replay and analyze embryogenesis in silico at any desired speed or resolution. Thus, for example, one could trace the lineage of a given cell type back to the early embryonic stage or identify windows of vulnerability during which specific cell types or genetic networks are particularly sensitive to disruption. Cost-effective, native 3D spatial transcriptomics remains to be developed to a state where it may yield ultra-high-resolution maps of embryogenesis, driving next-generation models of the virtual embryo.

Federated platforms and modeling frameworks will be crucial for integrating these resources, allowing continuous community-based refinement while safeguarding sensitive embryo data. In turn, Virtual Embryo Platform-based perturbation can facilitate experimental prioritization, guide prognosis and liberate early drug efforts from the ethical and logistical constraints of in vivo studies. A well-validated Virtual Embryo Platform will offer a 'sandbox' to test what-if scenarios, such as "What happens if we knock out or overexpress a given gene at a certain stage of embryo growth?", to estimate efficacy or failure at the organ level. Prototype components are already in place, as demonstrated by our efforts at MOSTA[2], Spateo[3] and 4D molecular holograms[3]. The short-term goals should focus on increasing spatiotemporal resolution and simulation accuracy, as well as designing overall validation systems.

Achieving this vision requires close collaboration across disciplines: developmental biologists and geneticists to conceptualize and interpret data, computational scientists to design scalable algorithms,

and ethicists to ensure responsible innovation. With the refinement of dynamic computer-simulated embryos to ever greater accuracy, we envision them increasingly developing into essential tools of scientific experimentation beyond embryogenesis to include the modeling of whole tissue and organ systems in the form of virtual biological entities in developmental biology and human medicine. Ultimately, the Virtual Embryo Platform will accelerate regenerative medicine and enable early modeling of congenital disorders, bringing to the forefront the underlying principles for understanding the origins of life itself.

Natalie Cao[1,2,3], Yifan Lu[1,2,3] & Xiaojie Qiu[1,2,3]
[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [2]Basic Sciences and Engineering Initiative, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA.
e-mail: xiaojie@stanford.edu

Published online: 26 March 2026

References

1. Qiu, X. et al. *Cell* **188**, 1744 (2025).
2. Chen, A. et al. *Cell* **185**, 1777–1792.e21 (2022).
3. Sampath Kumar, A. et al. *Nat. Genet.* **55**, 1176–1185 (2023).
4. Qiu, C. et al. *Nature* **626**, 1084–1093 (2024).
5. Cui, L. et al. *Nat. Cell Biol.* **27**, 360–369 (2025).
6. Xiao, Z. et al. *Cell* **187**, 2855–2874.e19 (2024).
7. Yuan, Y. et al. *Cell Stem Cell* **32**, 1006–1024.e5 (2025).
8. Xie, P. et al. *Cell* **188**, 4754–4772.e18 (2025).
9. McDole, K. et al. *Cell* **175**, 859–876.e33 (2018).
10. Qiu, C. et al. *Nat. Genet.* **54**, 328–341 (2022).
11. Theodoris, C. V. et al. *Nature* **618**, 616–624 (2023).
12. Cui, H. et al. *Nat. Methods* **21**, 1470–1480 (2024).
13. Karr, J. R. et al. *Cell* **150**, 389–401 (2012).
14. Macklin, D. N. et al. *Science* **369**, eaav3751 (2020).
15. Bunne, C. et al. *Cell* **187**, 7045–7063 (2024).

Acknowledgements

X.Q. acknowledges support from the Laude Moonshot Seed Grant, the NIH New Innovator Award (DP2-HG014282-01), the NIH Pathway to Independence Award (R00-HG012887-03) and the Pantas And Ting Sutardja Foundation, the Wu Tsai Neurosciences Institute Big Ideas in Neuroscience Program. Y.L. and X.Q. acknowledge support from the Arc Ignite Award.

Author contributions

N.C.: led the conception, planning, writing, figure design and revision of the paper.
Y.L.: contributed to the design of the multiscale spatiotemporal model of the virtual embryo.
X.Q.: conceived and supervised the Virtual Embryo project, guiding project planning, writing, figure design and manuscript revision. Mentored N.C. and Y.L. throughout the study.

Competing interests

The authors declare no competing interests.
Primary handling editor Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.