The Virtual Embryo Challenge: Generative Modeling of Embryogenesis Across Space, Scale and Time

Jiayuan Ding[†] Yifan Lu[†] Qingquan Zhang[¶] Zehua Zeng[†] Weize Xu[†]
Nic Fishman[§] You He[†] Siyu He[†] Le Song[‡] Eric Xing[‡]
James Zou[†] Emily Fox[†] Marinka Zitnik[§] Neil Chi[¶] Xiaojie Qiu[†*]

[†]Stanford University [‡]GenBio [§]Harvard University [¶]UC San Diego

[*] xiaojie@stanford.edu

Abstract

Embryogenesis is one of the most fundamental yet least understood processes in biology. Recent advances in single-cell and spatial genomics enable molecular profiling of developing embryos at unprecedented spatial and temporal resolution, creating new opportunities for predictive AI systems that model development across space and time. However, there is currently no standardized benchmark for evaluating whether machine learning systems can accurately predict developmental dynamics, spatial organization, or perturbation outcomes during embryogenesis. We present **The Virtual Embryo Challenge, the first large-scale benchmark competition for predictive embryogenesis modeling**. The competition is built on a multimodal mouse embryo dataset containing approximately one million cells across 11 developmental stages, integrating single-cell multi-omics, spatial transcriptomics, and genetic perturbations. Participants will address three tasks: **temporal gene-expression prediction, spatial-temporal prediction, and mutant perturbation prediction**. Participants will have access to curated datasets, preprocessing pipelines, evaluation scripts, starter notebooks, and reproducible baselines. Submissions will be evaluated using biologically grounded metrics for gene-expression accuracy, cell-type composition, and spatial organization. The competition also includes a novel **Agent Team track**, where autonomous coding agents or recursive LLM-based systems iteratively improve predictive models. The Virtual Embryo Challenge provides a reproducible benchmark framework for predictive developmental biology and aims to accelerate progress toward virtual embryo, virtual organ, and eventually digital human modeling.

Keywords Generative Modeling, Virtual Embryo, AI Agent, Single Cell and Spatial Transcriptomics, Congenital Diseases.

1 Competition description

1.1 Background and impact

Embryogenesis is one of the most fundamental yet least understood processes in biology [1, 2, 3, 4]. A single fertilized cell gives rise to a complete organism through hierarchical spatiotemporal orchestration of gene regulation, cell fate transitions, tissue morphogenesis, and organ formation. When this process is disrupted, the consequences can be severe: congenital defects affect 1 in 33 newborns and remain a leading cause of infant mortality worldwide [5]. Despite decades of

Accepted by the competition track at NeurIPS 2026.

progress, we still lack predictive and generative AI frameworks that can model when, where, and how embryonic development goes off course.

Recent advances in single-cell and spatial genomics have created unprecedented opportunities to model development at the whole-embryo scale [3], turning embryogenesis into a compelling frontier for the NeurIPS community. The problem naturally connects to core areas of machine learning, including trajectory inference [6, 7, 8, 9], neural dynamical systems [10], and deep generative models [11]. Large single-cell and spatial transcriptomics-based embryonic atlases across species provide rich molecular snapshots across developmental time, but they do not by themselves reveal how cell states transition, how local molecular changes propagate to tissue- and organ-level phenotypes, or how development responds to perturbations. This creates a challenging benchmark for models that must jointly learn spatial context[12], temporal dynamics [7, 8, 9], biological hierarchy [13], and perturbation effects [14].

The Virtual Embryo Challenge addresses this gap by establishing a standardized benchmark for predictive embryogenesis modeling. The competition will advance generative, causal, and agentic AI systems that model embryonic development across space, scale, and time under genetic perturbations. Grounded in multimodal whole-embryo and early cardiac development datasets spanning eleven developmental stages, the benchmark will evaluate models on spatial context, multiscale reasoning, and temporal dynamics. By releasing curated datasets, baselines, and evaluation tools, this competition will catalyze open innovation toward robust, interpretable, and generalizable virtual embryo models.

1.2 Novelty

The Virtual Embryo Challenge is a new competition on predictive virtual organ modeling. This is an entirely new competition and does not reuse tasks from any previous NeurIPS competition [15, 16, 17] or prior benchmark series [18, 19]. To our knowledge, it is the first competition focused on building a virtual embryo, and more broadly, one of the first community benchmarks for virtual organ modeling. Unlike existing virtual cell challenges[20, 21], which primarily evaluate cells in isolation or at steady state, this challenge requires models to predict the collective and emergent properties of biological systems—the embryo—and how they evolve across space, scale, and time under genetic perturbation.

The competition introduces a new multimodal benchmark for spatial, temporal, multiscale, and perturbation-aware embryogenesis modeling. We have constructed a 4D mouse embryo development atlas spanning whole-embryo and heart-focused datasets, multiple developmental stages, spatial molecular profiles, cell-type annotations, and genetic perturbation conditions. The benchmark is designed around held-out developmental stages, perturbation conditions, and combinations of these factors, requiring models to learn generalizable developmental dynamics rather than memorize observed examples. The associated tasks and metrics evaluate capabilities that are not jointly tested by existing benchmarks: spatial context modeling, future-stage prediction, perturbation response prediction, multiscale organization, and biological plausibility of generated developmental states.

The challenge will drive generative modeling for biology beyond static atlas analysis. Embryogenesis provides a demanding setting for modern ML methods, including diffusion and flow-based generative models [22, 23, 24, 25], neural dynamical systems [10], causal representation learning [26, 27], graph and geometric deep learning [28, 29], multimodal and multi-scale foundation models [30, 31], and agentic scientific workflows [32]. Participants may build models that generate future molecular states, infer cell fate transitions, predict tissue-level spatial organization, or use agents to reason over embryo atlases, perturbation metadata, and evaluation feedback. By bringing these methods into a shared competition framework, the Virtual Embryo Challenge will help clarify which modeling paradigms are most effective for predictive developmental biology.

The benchmark establishes an evaluation foundation for AI virtual embryo and virtual organ models. A key expected outcome is not only improved leaderboard performance, but also a clearer understanding of how to evaluate virtual organ models: what should be predicted, which generalization settings are biologically meaningful, and which metrics best capture spatial, temporal, perturbational, and multiscale fidelity. The released datasets, baselines, tasks, and evaluation tools will provide a reusable foundation for future benchmarks in, as well as pointing to the right direction for virtual embryo modeling, congenital disease modeling, and AI-driven experimental design.

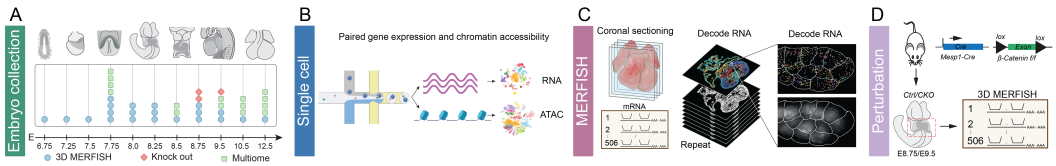

Figure 1: **Overview of the multimodal spatiotemporally-resolved mouse embryo data resource.** The benchmark integrates embryo collections from E6.75 to E12.5 (**A**), single-cell multiome measurements of RNA and chromatin accessibility (**B**), spatial transcriptomics from 3D MERFISH (**C**), and conditional knockout perturbation data around at E8.75/E9.5 (**D**).

1.3 Data

The Virtual Embryo Challenge will be built on a multimodal mouse embryo development resource containing approximately one million cells across 11 developmental time points between Embryonic day 6.75 or E6.75 to E12.5 (Figure 1). The data cover both whole-embryo and heart-focused datasets, profiled either with single cell multi-omics [33, 34], 3D MERFISH [35] with or without genetic knockout (Mab12l2, B-catenin, Gata4), spanning early gastrulation, cardiac progenitor emergence, heart tube formation, looping, and later heart morphogenesis. For each cell or spatial unit, the released data will include developmental stage, embryo or sample identifier, measured gene-expression features, cell-type annotations, and, when available, 3D spatial coordinates, anatomical-region labels, tissue-domain labels, and morphology-derived features.

The benchmark will use staged train, validation, and hidden-test splits. Public training data will contain earlier developmental stages and selected perturbation conditions. Public validation targets will support method development and leaderboard feedback. Final rankings will be computed on hidden test targets whose ground-truth expression, cell-type composition, spatial organization, and perturbation outcomes will be withheld until the end of the competition. Splits will be defined by developmental stage, perturbation condition, or combinations of these factors so that participants must learn generalizable developmental dynamics rather than memorize public validation examples.

All competition data will be derived from non-human mouse embryo experiments. The benchmark, therefore, does not involve identifiable human-subject data. The organizers will release processed matrices, metadata, coordinate conventions, cell-type vocabularies, baseline outputs, and data-loading utilities under terms that permit research use during the competition and open release where possible.

1.4 Tasks and application scenarios

This competition introduces three predictive embryogenesis tasks in Figure 2 that progress from molecular forecasting to spatial-temporal modeling and perturbation prediction. Together, these tasks evaluate whether virtual embryo models can extrapolate across developmental time, recover multiscale tissue organization, and predict how genetic perturbations alter embryonic and cardiac development.

Task 1: Temporal gene-expression distribution prediction. This task evaluates future molecular-state prediction from single-cell data. Given training stages before the target time point, participants predict the gene-expression distribution at future developmental stages. For the single cell embryonic multi-omics dataset, we covered the following states: E8.5, E9.5, E10.5, and E12.5. E10.5 will be used as the public validation target and E12.5 as the hidden test target. The primary output is the predicted gene-expression distribution for the target stage. This task tests temporal extrapolation: models must learn developmental trends rather than only interpolate between nearby samples.

Task 2: Spatial-temporal multiscale prediction. This task evaluates whether models can predict development jointly across molecular, cellular, and spatial scales using the 4D MERFISH dataset. Prediction targets include gene expression together with 3D spatial location. We have to different settings: one for the whole embryo and the other for the heart. For the embryo setting, we will train datasets from E6.75, E7.25, and E8.0, while validate on E7.5 and test on E7.5 stage to test the model's interpolation capability. For the heart setting, we will test both interpolation and extrapolation. We will use datasets from E8.25 and E9.5 as for the training while (1) validate on E8.5 and test on E8.75

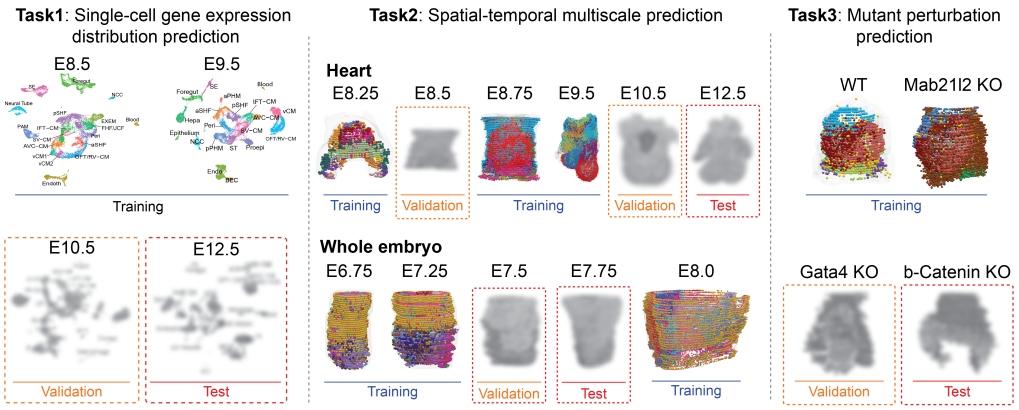

Figure 2: **Overview of the three prediction tasks. Task 1** predicts future single-cell gene-expression distributions from earlier developmental stages. **Task 2** predicts temporal changes of spatial transcriptomic states and spatial organization from 3D spatial data. **Task 3** predicts gene-expression and spatial outcomes for held-out genetic perturbations, with separate knockout conditions used for validation and testing.

for interpolation benchmark, and (2) validate on E10.5 and test on 12.5 for extrapolation benchmark. This task is designed to distinguish models that match global expression profiles from models that also recover where cell states are located and how tissue-scale structure changes.

Task 3: Mutant perturbation prediction. This task evaluates prediction under genetic perturbation with 3D MERFISH data. At E8.75, the dataset contains two perturbation conditions (B-catenin, Gata4). Additionally, we have another perturbation condiation at E9.5 for Mab21l2 mutant embryo. The benchmark will release one perturbation condition for participants for training (Mab21l2, because the mutant phenotype is most obvious and the gene is very specific) and hold out one for validation (Gata4, also very specific) and one for hidden testing (B-catenin, more broadly expressed). Participants will predict mutant gene expression distribution together with 3D spatial location. This setting tests whether virtual embryo models can generalize from wild-type development and observed perturbations to unseen genetic contexts across collection time points.

Two participation tracks. The competition runs in two parallel tracks (Figure 3). The **Human Team track** evaluates methods designed and supervised by human participants, following the conventional competition workflow of algorithm/model design, submission, and evaluation. The **Agent Team track** evaluates methods produced by coding agents or LLM-based evolutionary algorithms that iteratively select, mutate, and evaluate candidate solutions without human supervision. Both tracks address the same three tasks and are scored with the same metrics and hidden test sets; prizes are awarded separately for each track, so the leaderboards directly contrast human-designed and agent-designed approaches to predictive embryogenesis modeling.

1.5 Metrics

Submissions are evaluated by a fixed panel of nine metrics computed on held-out embryos after validation of the submission schema, gene order, coordinate convention, normalization convention, and missing-value rules. The panel is organized into groups, each corresponding to one aspect of the prediction, and each metric is normalized against two empirically measured anchors before aggregation, so that quantities with different units and orientations are placed on a common scale. The complete evaluation code, including the reference implementations of all metrics below, is released with the starter kit.

Notation. Let $\widehat{X}$, X, and X^{ref} denote the predicted, observed, and reference log-normalized expression matrices over a common ordered gene panel $\mathcal{G}$ with $|\mathcal{G}| = G$, where the reference is the preceding developmental stage for Tasks 1 and 2 and the stage-matched wild type for Task 3. We write $\overline{\mathrm{pred}}$, $\overline{\mathrm{true}}$, and $\overline{\mathrm{ref}} \in \mathbb{R}^G$ for the corresponding pseudobulk (cell-averaged) profiles, and $\widehat{C}, C$

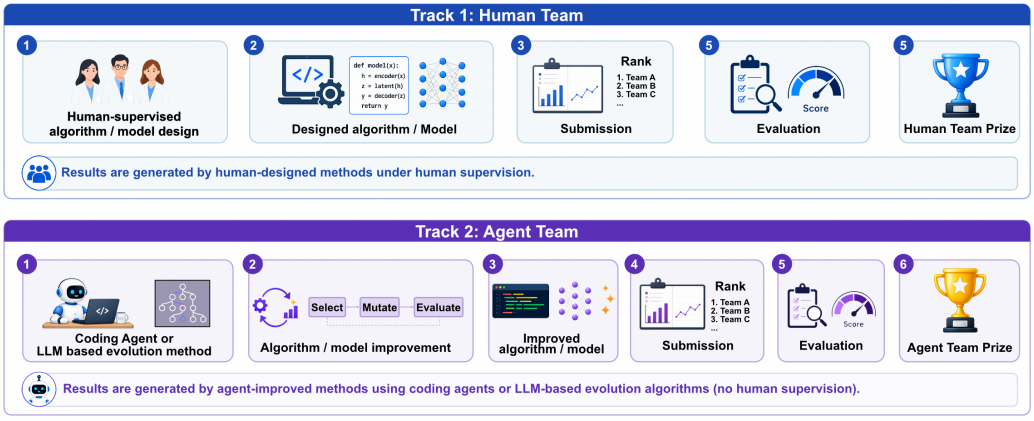

Figure 3: The competition runs in two tracks. **Track 1 (Human Team)**: results are generated by human-designed methods under human supervision. **Track 2 (Agent Team)**: results are generated by agent-improved methods using coding agents or LLM-based evolution algorithms, without human supervision. Both tracks share the same tasks, metrics, and hidden test sets, and are awarded prizes separately.

for predicted and observed spatial coordinates in Tasks 2 and 3. Two properties of a submission are treated as free and are absorbed by the metric definitions rather than constrained by submission rules: the coordinate frame, which is removed by the canonicalization step described below, and the number of predicted cells, since a model of growth and proliferation may legitimately predict a cell count different from that of the target. Comparisons that would otherwise be confounded by cell count subsample both populations to a common size.

Differential expression recovery. Differentially expressed genes are determined from the observed data by a two-sided Mann–Whitney U test of X against X^{ref}, with Benjamini–Hochberg control of the false discovery rate at $\alpha = 0.05$ and a minimum effect size of 0.25 in mean log-expression, yielding up- and down-regulated sets U_{true} and D_{true}. The prediction reports exactly $n_u = |U_{\mathrm{true}}|$ up-regulated and $n_d = |D_{\mathrm{true}}|$ down-regulated genes, ranked by its own log fold-change $\Delta_p = \overline{\mathrm{pred}} - \overline{\mathrm{ref}}$, which fixes the size of the reported sets and removes it as a free parameter. Writing U_p and D_p for the resulting sets and $\mathrm{ov} = (|U_p \cap U_{\mathrm{true}}| + |D_p \cap D_{\mathrm{true}}|)/(n_u + n_d)$ for the signed overlap,

$$\texttt{de\_score} = \frac{\mathrm{ov} - \mathrm{chance}}{1 - \mathrm{chance}}, \qquad \mathrm{chance} = \max\big(\mathrm{ov}[\Delta_p = +\overline{\mathrm{ref}}],\ \mathrm{ov}[\Delta_p = -\overline{\mathrm{ref}}]\big). \tag{1}$$

The chance correction is necessary because highly expressed genes have greater statistical power to be declared differentially expressed at a given effect size, so an uncorrected overlap rewards predictions that encode only baseline expression level. Subtracting the empirical null obtained from such a prediction places the origin of the scale at that level of information: zero indicates no improvement over baseline expression level alone, one indicates exact recovery of the observed differential expression call, and negative values indicate systematic assignment of the wrong genes or the wrong direction.

Directional agreement and response magnitude. Agreement in the direction of change is assessed genome-wide and independently of magnitude by a rank-based partial correlation that controls for baseline expression level. Let r_p, r_t, and r_r denote the rank transforms of Δ_p, $\Delta_t = \overline{\mathrm{true}} - \overline{\mathrm{ref}}$, and $\overline{\mathrm{ref}}$, and let C be their 3×3 Pearson correlation matrix. For Task 3, let β denote the zero-intercept ordinary least-squares slope of predicted on observed log fold-change restricted to the observed differentially expressed set, with coefficient of determination R^2. Then

$$\texttt{de\_direction} = \frac{C_{pt} - C_{pr}C_{tr}}{\sqrt{(1 - C_{pr}^2)(1 - C_{tr}^2)}}, \quad \texttt{severity\_slope} = \begin{cases} \log \beta, & \beta > 0,\ R^2 > 0.01, \\ \text{undef.}, & \text{otherwise.} \end{cases} \tag{2}$$

Partial correlation is required because developmental and perturbation responses are themselves correlated with baseline expression level, so an unadjusted correlation between predicted and observed log fold-changes credits a submission for reproducing this shared dependence in the absence of predictive signal. The quantity `severity_slope`, defined for Task 3 only, assesses whether the magnitude of the predicted response matches the observed one: zero indicates that predicted and observed effect sizes agree on average gene by gene, positive values indicate overestimation, and negative values underestimation. A regression slope is preferred to a ratio of norms, which is insensitive to direction and is therefore attained by any response of correct magnitude and arbitrary orientation; the variance gate reports the slope as undefined rather than assigning a value when the predicted response is unrelated to the observed one.

Distributional fidelity. Agreement between the predicted and observed cell populations is assessed by the maximum mean discrepancy. Both matrices are subsampled and projected onto a 30-component principal subspace estimated from the target alone, with radial basis function kernels $k_s(x, y) = \exp(-\gamma_s \|x-y\|^2)$ at bandwidths $\gamma_s = s\gamma_0$, $s \in \{0.25, 0.5, 1, 2, 4\}$, where γ_0 is the median-heuristic bandwidth of the target. Denoting by $\widehat{\mathrm{MMD}}_s^2$ the unbiased estimator, which omits the kernel diagonal,

$$\texttt{mmd\_u} = \tfrac{1}{5} \textstyle\sum_s \widehat{\mathrm{MMD}}_s^2. \tag{3}$$

Estimating both the projection and the bandwidths from the target alone ensures that the evaluation space is independent of the submission. The unbiased estimator is used because the biased form assigns a nonzero value of order $2/n$ to an exact prediction, which would otherwise be taken for the attainable upper bound during normalization.

Gene–gene covariance structure. The preceding metrics are either per-gene marginal statistics or are computed in a low-dimensional projection, and therefore do not constrain the joint structure across genes: a prediction may reproduce every marginal distribution while failing to preserve co-regulation between genes. This is assessed by a variogram statistic evaluated on 20,000 randomly sampled gene pairs (i, j), with $v_X(i, j) = \widehat{\mathbb{E}}_{\mathrm{cells}} |x_i - x_j|^{0.5}$:

$$\texttt{variogram} = \widehat{\mathbb{E}}_{(i,j)} \big[(v_{\mathrm{pred}}(i, j) - v_{\mathrm{true}}(i, j))^2 \big]. \tag{4}$$

Morphology and growth. Point clouds are compared after canonicalization, $Z = (C - \bar{C})V^\top / \mathrm{rms}$, where V is the principal-axis rotation of the cloud and $\mathrm{rms} = (\widehat{\mathbb{E}}_i \|c_i - \bar{C}\|^2)^{1/2}$, which removes translation, rotation, and scale; overall size is then assessed by a separate term. Since principal component analysis determines each axis up to sign, registered terms are optimized over the four sign assignments f with determinant $+1$; consequently these terms are invariant to reflection and do not resolve left–right inversion, which we state as a known limitation of the current panel.

$$\texttt{d2\_shape} = W_1\big(D_2(\mathrm{pred}), D_2(\mathrm{true})\big), \qquad D_2(\cdot) = \{\|x_i - x_j\| / \mathrm{rms}\}, \tag{5}$$

$$\texttt{occupancy\_dice} = \max_f \frac{2\big|\mathrm{occ}(f \odot Z_{\mathrm{pred}}) \cap \mathrm{occ}(Z_{\mathrm{true}})\big|}{\big|\mathrm{occ}(f \odot Z_{\mathrm{pred}})\big| + \big|\mathrm{occ}(Z_{\mathrm{true}})\big|}, \tag{6}$$

$$\texttt{scale\_log\_ratio} = \log \frac{\mathrm{rms}(\widehat{C})}{\mathrm{rms}(C)}, \tag{7}$$

where D_2 is estimated from 2×10^5 within-cloud point pairs and $\mathrm{occ}(\cdot)$ denotes occupancy on a shared 16^3 voxel grid spanning $[-3, 3]$ RMS radii. The statistic `d2_shape` compares distributions of within-cloud pairwise distances and therefore requires no registration, while `occupancy_dice` assesses whether the predicted tissue occupies the correct region of a shared rotation-invariant frame, and `scale_log_ratio` assesses growth in physical scale. The RMS radius is used in place of a bounding-box extent because it is invariant to rotation.

Local spatial organization. Joint fidelity of expression and position is assessed at the level of local neighborhoods. Let $\mathrm{nb}_i(X, C) = \frac{1}{15} \sum_{j \in \mathrm{kNN}_{15}(i;C)} X_j$ denote the mean expression of the 15 nearest spatial neighbors of cell i. Then

$$\texttt{neighborhood\_mmd} = \texttt{mmd\_u}\big(\{\mathrm{nb}_i(\widehat{X}, \widehat{C})\}, \{\mathrm{nb}_i(X, C)\}\big) \tag{8}$$

compares the distributions of neighborhood-averaged profiles and therefore requires no correspondence between predicted and observed cells. This statistic assesses expression and position jointly, rather than evaluating the two as independent quantities.

Score normalization. Each metric is normalized against a reference baseline m_{base} and an attainable upper bound m_{ub}. The baseline is `copy_last`, which reproduces the preceding developmental stage, for Tasks 1 and 2, and `wt_identity`, which reproduces the stage-matched wild type, for Task 3. The upper bound is estimated by scoring one half of the target against the other half, rather than by resubmission of the target, which is not attainable in the presence of sampling noise. Writing $d(m)$ for the distance of a metric from the upper bound, oriented so that larger values indicate worse performance, and $d_{\text{base}} = |m_{\text{ub}} - m_{\text{base}}|$,

$$\text{skill}(m) = \min\Big(\frac{d_{\text{base}}}{d_{\text{base}} + d(m)},\, 1\Big) \in (0, 1]. \tag{9}$$

This transformation maps the upper bound to one and the baseline to one half, and approaches zero asymptotically as performance degrades, so that scores are bounded without truncation. Metrics whose optimum is zero rather than an extremum, namely `severity_slope` and `scale_log_ratio`, are normalized on $|m|$. A metric that is undefined for a given submission is assigned $\text{skill} = 0$ rather than omitted, so that all submissions are evaluated on the same set of criteria.

Aggregation and ranking. Metrics are aggregated by group, with within-group weights $w_{m|g}$ and group weights w_g given in Table 1. Grouping ensures that the number of statistics available for a given aspect does not act as an implicit weight. The task and overall scores are

$$S_t = 100 \times \sum_g w_g \Big(\sum_{m \in g} w_{m|g}\, \text{skill}(m)\Big), \qquad S_{\text{overall}} = \frac{1}{3}\sum_{t=1}^{3} S_t, \tag{10}$$

with $\sum_g w_g = 1$ and $\sum_{m \in g} w_{m|g} = 1$. On this scale $S_t = 50$ corresponds to baseline performance, values above 50 indicate improvement over the baseline, and values below 50 indicate degradation.

Table 1: Metric groups and weights by task.

Task	Group	w_g	Metrics (within-group weight)
T1	Differential expression recovery	0.25	`de_score` (1.0)
	Directional agreement	0.25	`de_direction` (1.0)
	Distributional fidelity	0.30	`mmd_u` (1.0)
	Covariance structure	0.20	`variogram` (1.0)
T2	Expression change	0.25	`de_score` (0.5), `de_direction` (0.5)
	Cell-state distribution	0.25	`mmd_u` (0.6), `variogram` (0.4)
	Morphology and growth	0.25	`d2_shape`, `occupancy_dice`, `scale_log_ratio` ($\frac{1}{3}$ each)
	Local spatial organization	0.25	`neighborhood_mmd` (1.0)
T3	Perturbation response, identity	0.30	`de_score` (1.0)
	Perturbation response, direction	0.25	`de_direction` (1.0)
	Perturbation response, magnitude	0.25	|`severity_slope`| (1.0)
	Mutant-state distribution	0.20	`mmd_u` (0.6), `variogram` (0.4)

For Task 3, morphological and spatial terms are not scored, since they would compare the absolute post-perturbation morphology, whereas the biologically meaningful quantity is the change in morphology attributable to the perturbation; a formulation based on the predicted shift relative to wild type is planned. A wild-type-relative distributional term was likewise implemented and evaluated, but scored all methods below the `wt_identity` baseline, as prediction noise superimposed on a subtle perturbation effect renders the predicted residual distribution less similar to the observed residual distribution than a residual containing no shift; the distributional component for Task 3 is therefore computed on the mutant state directly.

In addition to the scored panel, every submission is checked against the released normalization and population conventions, since a submission that does not follow them is not comparable to the target; such submissions are flagged during validation. Both participation tracks are evaluated with the same panel, baselines, upper bounds, and hidden test sets. Confidence intervals are estimated by bootstrap resampling over held-out embryos, regions, cell types, and genes, and differences between leading submissions are assessed using paired bootstrap intervals on the same evaluation units. Ties are resolved by hidden-test performance, then by the differential expression group, then

by the distributional group. All weights, baseline and upper-bound values, smoothing constants, and minimum-cell thresholds are fixed in the public evaluation code before the leaderboard opens, and the public evaluator reproduces the final scoring interface exactly.

1.6 Baselines, code, and material provided

The baseline suite already includes two reference methods, shared across all three tasks, which are also the `copy_last`/`wt_identity` floor used for skill normalization above. `copy_last` reproduces the reference condition verbatim,

$$\widehat{X} = X^{\mathrm{ref}}, \tag{11}$$

and `pseudobulk_shift` extrapolates a per-cell-type constant velocity, estimated from the two most recent training stages, onto the real cells of the most recent stage, which preserves within-type heterogeneity rather than replacing each cell with a fitted mean. Writing $X^{(S-1)} = X^{\mathrm{ref}}$ for the most recent training stage and $X^{(S-2)}$ for the one preceding it, with cell-type label y_i for cell i,

$$\Delta_k = \overline{X}_k^{(S-1)} - \overline{X}_k^{(S-2)}, \qquad \widehat{x}_i = \max\bigl(0,\ x_i^{(S-1)} + \Delta_{y_i}\bigr), \tag{12}$$

with $\Delta_k = 0$ for a cell type k absent from stage $S-2$, so that cell types with no observed transition are held fixed rather than extrapolated from none. For **Task 1**, we will additionally provide temporal baselines such as pseudotime and OT-coupled flow-matching extrapolation. For **Task 2**, we will provide spatial baselines including optimal-transport alignment and integration. For **Task 3**, we will provide perturbation baselines that transfer wild-type-to-mutant expression shifts from observed perturbations, along with conditional generative baselines for mutant expression and spatial organization. These baselines are intended to establish transparent reference points rather than competitive upper bounds.

We will release the starting kit two weeks before the official competition launch. It will include: 1) Code to run and evaluate baseline models on the competition tasks; 2) Tools for data loading and formatting; 3) Documentation to support setup, experimentation, and reproducibility. All materials will be hosted on a public GitHub repository and announced on the competition website and mailing list.

1.7 Website, tutorial and documentation

The competition website will be hosted at `virtualembryo.ai`, with a dedicated challenge page at `virtualembryo.ai/challenge` that presents the background, the three tasks, the evaluation metrics, the full timeline, and the prize structure as a single self-contained page.

A distinguishing feature is that the competition data can be explored interactively before download. The website is built on the Virtual Embryo Atlas, a browser-based viewer for the same multi-modal mouse embryo resource the benchmark is drawn from. Participants can inspect 3D spatial-transcriptomics specimens, the single-cell developmental time-lapse UMAP embeddings, gene-expression coloring directly in the browser, organized both by developmental stage and by data modality. This lets participants assess modality coverage, stage spacing, and cell-type structure early, lowering the barrier to entry and supporting well-grounded modeling decisions before they commit to an approach.

The website features a prominent **FAQ/Tutorial section** covering registration, the submission file format, the data schema (gene order, cell-type vocabulary, coordinate conventions, missing-value rules), and worked examples for each evaluation metric, kept up to date throughout the competition. A **dedicated email address**, `virtual.embryo.moonshot@gmail.com`, is reachable from the home page for organizer contact, alongside the starter-kit GitHub repository for technical questions and bug reports. The website will be online and populated with the starter kit, tutorials, and reproducible baseline walkthroughs well before the official launch — comfortably ahead of the two-week post-acceptance requirement — so participants have ample time to prepare.

2 Organizational aspects

2.1 Protocol

The competition consists of three phases (Figure 4). The first is the **test phase**, which focuses on validating and refining the competition infrastructure and protocol. During this phase, participants

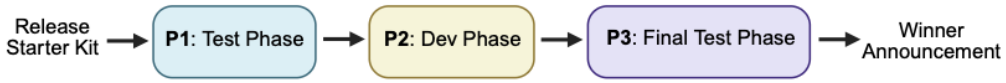

Figure 4: The competition is divided into three phases: testing, development, and final evaluations.

can access the starter kit, test the submission workflow, run the provided evaluation code, and submit trial prediction files to ensure that the platform, format, and evaluation pipeline work correctly. The second is the **development phase**, during which we release a validation dataset without ground-truth labels. Participants develop and iterate on their algorithms using the provided starter kit and submit prediction results in the required format through the competition platform. Submissions are evaluated automatically by our backend, and scores are displayed on a public leaderboard. The final phase is the **final test phase**, lasting approximately one week. We release a new held-out and more challenging test dataset without ground-truth labels. Participants submit predictions in the same required format, and final rankings and awards are determined based on performance on this held-out dataset.

Join the competition. Participants will create an account on the competition platform, download the starter kit and released datasets, develop their methods, and upload prediction files through the online submission interface. The starter kit will include baseline implementations, data loading utilities, evaluation scripts, and example submission files.

Preventing cheating and overfitting. To ensure fair evaluation, the development and final test datasets are separate, and all ground-truth labels remain hidden. Participants may iterate on the validation dataset during the development phase, but final rankings are determined only on the held-out final test set. The final phase may also limit the number of submissions to reduce leaderboard probing. Each person can join only one team and each team can work on one track. For the agent track, the agent system and complete evolutionary trace will need to be shared before the prize will be decided.

2.2 Rules and Engagement

Draft rules for The Virtual Embryo Challenge:

- **Submission:** Teams may submit prediction files during the development and final phases, with submission limits enforced to reduce leaderboard probing. Submissions must follow the required format, and participants may not infer hidden labels, manipulate evaluation, or use unauthorized test data.
- **Evaluation:** Submissions will be evaluated automatically using predefined metrics. Development scores will appear on a public leaderboard, while final rankings will be based solely on performance on the hidden held-out test set.
- **Eligibility for prizes:** Prize eligibility requires reproducible results and sufficient method documentation. Top teams may be asked to share code, outputs, and a short technical report. Violations such as duplicate registration, unauthorized data use, code sharing, or falsified results may lead to disqualification.

These rules promote fairness, transparency, and broad participation. Submission limits reduce leaderboard probing, hidden labels ensure evaluation generalization, and reproducibility requirements support validation and reuse of top-performing methods.

Communication plan.

Participants may contact organizers through the competition email, GitHub issues for technical questions, and the public discussion forum for general discussion. Announcements, rule updates, deadlines, dataset changes, and leaderboard information will be shared through the competition website and these channels.

2.3 Schedule and readiness

- **Readiness:** Evaluation data generation, evaluation code, data validation, baseline models.
- **06/30/2026:** Launch NeurIPS competition website, submission portal, and evaluation platform.
- **07/20/2026:** Release starter kit and open website to participants.
- **07/30/2026:** Begin beta test phase to validate the submission infrastructure (P1 in Figure 4).

- **08/15/2026:** Start the development phase and release the validation dataset (P2 in Figure 4).
- **10/25/2026:** Begin the final test phase with a new held-outdataset (P3 in Figure 4).
- **11/02/2026:** Final submissions due; start official evaluation.
- **11/18/2026:** Announce top-ranked teams and award recipients.

2.4 Competition promotion and incentives.

We will promote the competition through Stanford and Laude Institute networks and AI-for-science, machine learning, computational biology, single-cell/spatial genomics, developmental biology, and biomedical AI communities, using mailing lists, social media, invited talks, conference announcements, workshops, and direct outreach. The website will provide participation instructions, timelines, FAQs, tutorials, starter kits, baselines, and communication channels.

Incentives will include monetary prizes, travel awards for early-career researchers, poster and podium presentations at the NeurIPS competition workshop, and opportunities for technical reports or short papers. Top teams will present their methods, and reproducible submissions may be highlighted in post-competition summaries or publications. Authorship and acknowledgment policies for any organizer-led manuscript will be announced before launch with transparent contribution criteria.

We will broaden participation by advertising travel awards, providing accessible tutorials and starter materials, supporting both human-led and agentic tracks, and reducing barriers through standardized data loaders, baselines, evaluation scripts, and clear submission formats.

2.5 Competition track workshop and dissemination.

We will use the NeurIPS competition track workshop to present the competition goals, datasets, tasks, metrics, baselines, and final results. The workshop will include organizer and invited talks, posters, winner presentations across both tracks, and moderated discussions among participants, organizers, data providers, and domain experts. We will emphasize leaderboard performance as well as biological interpretability, robustness, reproducibility, and lessons from different modeling strategies.

Results will be shared through the competition website, public leaderboards, post-competition reports, workshop presentations, and, where appropriate, an organizer-led summary manuscript or white paper. Participants will be encouraged to release technical reports, model cards, and code when possible, subject to data-use and competition rules. The organizing team will maintain tutorials, documentation, communication channels, and post-competition engagement to support reuse of the datasets, evaluation framework, and baselines.

3 Resources

The competition is supported by the Laude Institute Moonshots Seed Grant and organizer-provided computational infrastructure. Submissions will be evaluated on Stanford Sherlock and GenBio GPU clusters, including NVIDIA H100 and H200 80GB GPUs, enabling scalable evaluation across both tracks, reproducibility checks, and final scoring on held-out embryogenesis datasets.

The organizing team and technical staff will maintain the evaluation pipeline, leaderboard, documentation, communication channels, and post-competition analysis to ensure rigorous, accessible, and fair participation.

The Laude Institute Moonshots Seed Grant will support three areas:

- **Winner Prizes ($54,000):** Awards for top-performing teams across both tracks. For each human or agent track, there will be one first prize ($8,000), two second prizes ($5,000 each), and three third prizes ($3,000 each).
- **Travel Awards ($30,000):** Support for 15–20 early-career researchers to attend the NeurIPS workshop, prioritizing trainees and participants from groups or institutions with limited travel support.
- **Outreach and Education ($20,000):** Support for the website, tutorials, starter kits, reproducible walkthroughs, demos, baseline documentation, and participant communication channels.

3.1 Organizing team

The organizing team spans computer science, statistics, biomedical AI, computational biology, single-cell/spatial genomics, developmental biology, and embryology, ensuring technical rigor, biological relevance, and accessibility. The team includes coordinators, data providers, platform administrators, baseline developers, beta testers, and evaluators for competition execution, infrastructure, dry runs, final scoring, and interpretation.

The team will oversee rule enforcement, participant support, fair ranking, and reproducibility checks, while emphasizing diversity across disciplines, career stages, institutions, and expertise.

3.2 Resources provided by organizers

Organizers will provide the website, documentation, FAQs, tutorials, starter kits, baseline code, data loaders, evaluation scripts, leaderboard infrastructure, hidden test-set evaluation, and technical support. Computational resources will include Stanford Sherlock and GenBio GPU clusters with NVIDIA H100 and H200 80GB GPUs. The Laude Institute Moonshots Seed Grant will support prizes, travel awards, outreach, education, tutorials, starter kits, and participant engagement. Scientific and technical support will be provided through office hours, online channels, and workshop sessions.

3.3 Support requested

We request a conference room with display support for invited talks, winner presentations, and poster sessions.

References

[1] Natalie Cao, Yifan Lu, and Xiaojie Qiu. Towards predictive virtual embryos with genomics and ai. Nature Methods, 2026.

[2] Vivien Marx. How to build a virtual embryo. Nature Methods, 20:1838–1843, 2023.

[3] Xiaojie Qiu, Daniel Y Zhu, Yifan Lu, Jiajun Yao, Zehua Jing, Kyung Hoi Min, Mengnan Cheng, Hailin Pan, Lulu Zuo, Samuel King, et al. Spatiotemporal modeling of molecular holograms. Cell, 187(26):7351–7373, 2024.

[4] Chengxiang Qiu, Beth K Martin, Ian C Welsh, Riza M Daza, Truc-Mai Le, Xingfan Huang, Eva K Nichols, Megan L Taylor, Olivia Fulton, Diana R O'Day, et al. A single-cell time-lapse of mouse prenatal development from gastrula to birth. Nature, 626(8001):1084–1093, 2024.

[5] World Health Organization. Every journey matters: World birth defects day. `https://www.paho.org/en/news/1-3-2024-every-journey-matters-world-birth-defects-day`, 2024. Accessed: 2026-05-15.

[6] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. Nature biotechnology, 37(5):547–554, 2019.

[7] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. Nature methods, 14(10):979–982, 2017.

[8] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. Nature Methods, 14(10):979–982, 2017.

[9] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745):496–502, 2019.

[10] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in neural information processing systems, 31, 2018.

[11] Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. Advances in neural information processing systems, 34:21325–21337, 2021.

[12] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. Nature methods, 19(5):534–546, 2022.

[13] Huasen Jiang, Xiaoyu Huang, Xiangpeng Bi, Wenjian Ma, Haibo Ni, Zhiqiang Wei, Pin Sun, Henggui Zhang, and Shugang Zhang. Artificial intelligence-enabled multi-scale virtual cell: perspective, challenges, and opportunities. Briefings in Bioinformatics, 27(2):bbag104, 2026.

[14] Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Machine learning for perturbational single-cell omics. Cell Systems, 12(6):522–537, 2021.

[15] Christopher Lance, Malte D Luecken, Daniel B Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, et al. Multimodal single cell data integration challenge: results and lessons learned. BioRxiv, pages 2022–04, 2022.

[16] Daniel Burkhardt, Jonathan Bloom, Robrecht Cannoodt, Malte D Luecken, Smita Krishnaswamy, Christopher Lance, Angela O Pisco, and Fabian J Theis. Multimodal single-cell integration across time, individuals, and batches. NeurIPS Competitions, page 12, 2022.

[17] Open Problems in Single-Cell Analysis. Neurips 2023: Single-cell perturbation prediction. `https://openproblems.bio/events/2023-08_neurips`, 2023. Generalizing experimental interventions to unseen contexts, a NeurIPS 2023 competition.

[18] Jiayuan Ding, Renming Liu, Hongzhi Wen, Wenzhuo Tang, Zhaoheng Li, Julian Venegas, Runze Su, Dylan Molho, Wei Jin, Yixin Wang, et al. Dance: a deep learning library and benchmark platform for single-cell analysis. Genome Biology, 25(1):72, 2024.

[19] Jiayuan Ding, Zhongyu Xing, Yixin Wang, Renming Liu, Sheng Liu, Zhi Huang, Wenzhuo Tang, Yuying Xie, James Zou, Xiaojie Qiu, et al. Dance 2.0: Transforming single-cell analysis from black box to transparent workflow. bioRxiv, pages 2025–07, 2025.

[20] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell, 187(25):7045–7063, 2024.

[21] Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual cell challenge: Toward a turing test for the virtual cell. Cell, 188(13):3370–3374, June 2025.

[22] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys, 56(4):1–39, 2023.

[23] Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. Bioinformatics, 40(9):btae518, 2024.

[24] Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, et al. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. Nature methods, 23(1):65–77, 2026.

[25] Dongyi Wang, Yuanwei Jiang, Zhenyi Zhang, Xiang Gu, Peijie Zhou, and Jian Sun. Joint velocity-growth flow matching for single-cell dynamics modeling. Advances in Neural Information Processing Systems, 38:160552–160588, 2026.

[26] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634, 2021.

[27] Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. Nature Genetics, 57(4):797–808, 2025.

[28] Dylan Molho, Jiayuan Ding, Wenzhuo Tang, Zhaoheng Li, Hongzhi Wen, Yixin Wang, Julian Venegas, Wei Jin, Renming Liu, Runze Su, et al. Deep learning in single-cell analysis. ACM Transactions on Intelligent Systems and Technology, 15(3):1–62, 2024.

[29] Hongzhi Wen, Jiayuan Ding, Wei Jin, Yiqi Wang, Yuying Xie, and Jiliang Tang. Graph neural networks for multimodal single-cell data integration. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pages 4153–4163, 2022.

[30] Zhaoyu Fang, Ziyang Miao, Jianhui Lin, Yuying Xie, Jiliang Tang, Jiayuan Ding, and Min Li. sclinguist: A pre-trained hyena-based foundation model for cross-modality translation in single-cell multi-omics. bioRxiv, pages 2025–09, 2025.

[31] Sumeer Ahmad Khan, Xabier Martínez-de Morentin, Abdel Rahman Alsabbagh, Alberto Maillo, Vincenzo Lagani, David Gomez-Cabrero, Robert Lehmann, and Jesper Tegner. Multimodal foundation transformer models for multiscale genomics. Nature Methods, 23(2):299–311, 2026.

[32] Weize Xu, Erwin Poussi, Quan Zhong, Zehua Zeng, Christopher Zou, Xuehai Wang, Yifan Lu, Miao Cui, Daiji Okamura, Cinlong Huang, et al. Pantheonos: An evolvable multi-agent framework for automatic genomics discovery. bioRxiv, pages 2026–02, 2026.

[33] Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. Nature Reviews Molecular Cell Biology, 24(10):695–713, 2023.

[34] Lia Chappell, Andrew JC Russell, and Thierry Voet. Single-cell (multi) omics technologies. Annual review of genomics and human genetics, 19(1):15–41, 2018.

[35] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. Science, 348(6233):aaa6090, 2015.

A Biography of all team members

[Platform administration, Beta tester, Evaluator, Coordination] **Jiayuan Ding** is a Staff Research Scientist and Tech Lead on the Retrieval-Augmented Generation (RAG) Team at Hippocratic AI, a leading healthcare large language model startup that reached a $3.5 billion valuation in less than three years. At the same time, he is a Visiting Researcher in the Department of Genetics at Stanford University, advised by Dr. Xiaojie Qiu, where he works on foundation models and agentic workflows for single-cell biology. He received his Ph.D. in Computer Science from Michigan State University. His research focuses on graph neural networks, LLM-based RAG systems, and AI for Science, with an emphasis on foundation models for single-cell biology and AI-driven scientific discovery. He is the founder of OmicsML, a non-profit dedicated to developing open-source computational tools leveraging LLMs and generative AI for scientific discovery. His recent work includes Tabula, a foundation model for single-cell transcriptomics, and DANCE, a benchmarking platform for single-cell analysis. He has published extensively in top-tier venues such as NeurIPS, ICLR, WWW, ACL, EMNLP, KDD, ICDE, CIKM, ACM TIST, and Genome Biology, and actively serves the research community as Area Chair or Program Committee Member for NeurIPS, ICLR, ICML, WWW, ACL, EMNLP, KDD, and ICDE. He has received multiple honors, including two MSU Engineering Distinguished Scholar Awards, first place in the NeurIPS 2021 Single-Cell Competition, and a Kaggle Silver Medal (Top 2%) in the NeurIPS 2022 Multimodal Challenge. His tools and models have influenced both academia and industry and have been featured by Forbes, Yahoo Finance, and the Associated Press.

[Task design, Metric design, Baseline method provider] **Yifan Lu** is a visiting researcher in Dr. Xiaojie Qiu's laboratory at Stanford University. His work focuses on computational methods for single-cell and spatial genomics, with interests in developmental biology, generative modeling, and predictive modeling of cellular dynamics. He contributes to the design of biologically meaningful benchmark tasks for virtual embryo modeling, including future-stage prediction, perturbation-response prediction, and spatial-temporal developmental modeling. In the Virtual Embryo Challenge, he leads task definition, metric design, and baseline development, helping translate biological questions in embryogenesis into rigorous machine learning benchmarks and evaluation protocols.

[Task design, Data generation] **Qingquan Zhang** is an Assistant Project Scientist at the University of California, San Diego. His research focuses on heart development and the gene regulatory networks underlying cardiogenesis, with an emphasis on how cellular interactions and transcriptional programs coordinate cardiac morphogenesis. He leads human and mouse 3D/4D spatial transcriptomic heart atlas projects, generating high-resolution spatial maps of developing hearts to reconstruct cellular organization and developmental trajectories across species. He integrates spatial transcriptomics, single-cell multiomics, and computational approaches to study heart field specification and tissue patterning during embryogenesis. He received his PhD from Tongji University in 2016 and has over 15 years of research experience in heart development. He completed postdoctoral training in the laboratory of Neil Chi at the University of California, San Diego starting in 2019.

[Protocol design, Platform development] **Zehua Zeng** is a Graduate Visiting Researcher in the Department of Genetics at Stanford University and a Ph.D. student at the University of Science and Technology Beijing. His research focuses on deep learning methods for integrative spatial multi-omics analysis, especially across single-cell transcriptomics, spatial transcriptomics, and epigenomics. He developed OmicVerse, an open-source Python framework for transcriptomics and multi-omics analysis that integrates bulk RNA-seq, single-cell RNA-seq, spatial transcriptomics, visualization, and AI-assisted workflows. His work on OmicVerse was published in Nature Communications and introduced a unified framework for bridging bulk and single-cell sequencing analysis. More recently, he has extended this ecosystem toward agent-enabled biological data analysis through OmicClaw and the J.A.R.V.I.S. runtime. In the Virtual Embryo Challenge, he contributes to protocol design and platform development, with a focus on reproducible omics workflows and agent-enabled execution infrastructure.

[Agent track lead, Platform development] **Weize Xu** is a postdoctoral researcher in the Department of Genetics at Stanford University, working in Dr. Xiaojie Qiu's laboratory on computational tools for genomics and single-cell biology. During his Ph.D. he developed computational methods and pipelines for spatial transcriptomics and single-cell Hi-C, with particular expertise in bioimage-

processing software and web-based analysis platforms. He created U-FISH, a deep-learning method for detecting signal points in fluorescence in situ hybridization (FISH) images that achieves robust performance across diverse data types through a carefully curated training dataset, and integrated it with large language models to provide an intuitive interface for non-expert users. More recently he developed PantheonOS, an evolvable, privacy-preserving multi-agent framework for automatic genomics discovery: it enables agentic code evolution — driving batch-correction and reinforcement-learning-augmented gene-panel-selection algorithms to super-human performance — and has powered biological discoveries spanning early mouse-embryo 3D data, multi-omic integration of the developing heart, and adaptive selection of virtual cell models for cardiac perturbation prediction. He has also contributed to web development for the Bioimage.IO deep-learning platform at SciLifeLab. He has published as first or co-first author in Nature Biomedical Engineering, Nature Communications, Genome Biology, and BMC Bioinformatics. In the Virtual Embryo Challenge, he leads the Agent track and contributes to platform development, drawing on his multi-agent systems and bioimaging expertise.

[Baseline development, Evaluator] **Nic Fishman** is a Ph.D. student in Statistics at Harvard University, where he works in Dr. Marinka Zitnik's lab. His research focuses on building large-scale machine learning systems at the intersection of generative AI, multiscale modeling, and causal inference, with applications in biology, healthcare, and social science. He develops statistical and machine learning methods for reasoning over large, complex datasets, including applications in genomics, causal discovery, network analysis, and generative modeling. He previously completed graduate study in statistics and machine learning at the University of Oxford as a Rhodes Scholar and received his undergraduate degree from Stanford University. In the Virtual Embryo Challenge, he contributes expertise in generative modeling, causal inference, multiscale learning, and evaluation of predictive biological systems.

[Baseline development, Evaluator] **You He** is a Ph.D. student in the Department of Bioengineering at Stanford University. She received her bachelor's degree in Theoretical and Applied Mechanics from Tsinghua University. Her research interests lie at the intersection of quantitative biology, biophysics, and machine learning.

[Baseline development, Evaluator] **Siyu He** is a Katharine McCormick postdoctoral fellow in the Department of Biomedical Data Science at Stanford University, where she is co-advised by Dr. James Zou and Dr. Stephen Quake. Her research lies at the intersection of AI and biomedicine, with a focus on developing generative AI models to learn and predict the spatiotemporal dynamics of cells in healthy and diseased states. She has developed computational methods for single-cell and spatial omics, including Starfysh, CORAL, Squidiff, and Devo, spanning spatial deconvolution, integrated spatial multi-omics, diffusion-based perturbation prediction, and developmental modeling. Her work has been published in venues including Nature Biotechnology and Nature Methods. In the Virtual Embryo Challenge, she contributes expertise in generative modeling, spatial transcriptomics, cell atlas modeling, and evaluation of spatiotemporal developmental predictions.

[Organization, Dissemination] **Le Song** is a co-founder and Chief Scientist of GenBio AI, a company focused on building large-scale foundation models for genomics and drug discovery. He was previously a Professor in the College of Computing at Georgia Institute of Technology and has held positions at Carnegie Mellon University and Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). He received his Ph.D. from the University of Sydney. His research spans machine learning theory and methodology, including kernel methods, probabilistic graphical models, deep learning, and reinforcement learning, with increasing focus on biological foundation models and AI-driven scientific discovery. He has published extensively at NeurIPS, ICML, and ICLR, and has received numerous honors including the NSF CAREER Award and the Kolon Faculty Fellowship. His current work at GenBio AI applies large-scale generative models to single-cell genomics and therapeutic design, directly relevant to the biological modeling challenges at the core of this competition.

[Organization, Dissemination] **Eric Xing** is the President of Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) and a Professor at Carnegie Mellon University, with appointments in the Machine Learning Department, Language Technology Institute, and Computer Science Department. He is a co-founder of GenBio AI. He received his Ph.D. from the University of California, Berkeley. His research focuses on the development of machine learning and statistical methodology,

and large-scale computational systems for automated learning, reasoning, and decision-making in high-dimensional, multimodal, and dynamic settings. His recent work includes AIDO (AI-Driven Digital Organism), a multiscale biological foundation model system, and PAN-World, an action-conditioned world model for reasoning and planning — both directly relevant to the generative modeling challenges at the core of this competition. He served as Program Committee Chair for ICML 2014 and General Chair for ICML 2019, and is a Board Member of the International Machine Learning Society. He serves as Action or Associate Editor for JASA, AOAS, JMLR, MLJ, and PAMI, and has mentored numerous students who now hold faculty and leadership positions at major universities and technology companies worldwide.

[Organization, Dissemination] **James Zou** is an Associate Professor of Biomedical Data Science at Stanford University, with courtesy appointments in Computer Science and Electrical Engineering, and Director of the Stanford AI for Science Lab. He received his Ph.D. from Harvard University, was a Gates Scholar at Cambridge University, a Simons Fellow at UC Berkeley, and a researcher at Microsoft Research before joining Stanford in 2016. His research focuses on machine learning for biomedicine and health, AI for scientific discovery, data valuation and quality assessment, fairness and interpretability in AI systems, and genomics and precision medicine. He has received numerous honors, including the Overton Prize (2025), Sloan Research Fellowship, NSF CAREER Award, two Chan-Zuckerberg Investigator awards, and faculty awards from Google, Amazon, Genentech, and Apple. His algorithms have received FDA approval and are used by millions of developers globally, and his work has been recognized with multiple best paper awards at top venues. He is a two-time Chan-Zuckerberg Investigator and serves as Faculty Director at the Stanford Data Science Institute.

[Organization, Dissemination] **Emily Fox** is a Professor of Statistics and, by courtesy, of Computer Science at Stanford University. She received her Ph.D. in Electrical Engineering and Computer Science from MIT and was previously an Associate Professor at the University of Washington. Her research focuses on Bayesian nonparametric methods, time series modeling, sequential data analysis, and scalable machine learning, with applications in neuroscience, genomics, and healthcare. She is known for foundational contributions to Bayesian nonparametric models for dynamical systems, including hidden Markov models and state space models that are widely used in computational biology and clinical data analysis. She has received numerous honors including the Sloan Research Fellowship and the NSF CAREER Award, and has served in leadership roles at major machine learning venues. Her expertise in probabilistic modeling of complex biological time series is directly relevant to the spatiotemporal generative modeling challenges at the heart of this competition.

[Organization, Dissemination] **Marinka Zitnik** is an Associate Professor of Biomedical Informatics at Harvard Medical School, at the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University, and at the Broad Institute of MIT and Harvard. Zitnik investigates foundations of AI that contribute to the scientific understanding of medicine and therapeutic design, eventually enabling AI to learn and innovate on its own. Her research won best paper and research awards, including the Kavli Fellowship of the National Academy of Sciences, Kaneb Fellowship award at Harvard Medical School, NSF CAREER Award, awards from the International Society for Computational Biology, International Conference in Machine Learning, Bayer Early Excellence in Science, Amazon Faculty Research, Google Faculty Research, Roche Alliance with Distinguished Scientists, and Sanofi iDEA-iTECH Award. Zitnik founded Therapeutics Data Commons, a global open-science initiative to access and evaluate AI across stages of development and therapeutic modalities, and she served as the faculty lead of the AI4Science initiative.

[Data generation, Dissemination] **Neil Chi** is a Professor of Medicine at the University of California San Diego (UCSD) and Director of the UCSD Cardiac Tissue Harvest and Biorepository Core and the UCSD MD-PhD Medical Scientist Training Program. He holds an MD and a PhD and is a leading expert in cardiovascular development, cardiac regeneration, and congenital heart disease. His research focuses on cardiac lineage specification and development, myocardial reprogramming, gene regulatory networks directing heart formation, hemodynamic influences on cardiac morphogenesis, and single-cell and spatial transcriptomics of developing hearts. He maintains active NIH funding through multiple R01 and U01 grants and has published over 90 peer-reviewed articles in high-impact journals including *Nature*, *Cell*, *PNAS*, and *Development*. For this competition, his laboratory is responsible for generating the virtual embryo datasets — including time-resolved, organism-level 3D

single-cell and spatial transcriptomics data from mouse embryogenesis — that form the biological foundation of the challenge.

[Platform administration, Organization, Dissemination, Coordination] **Xiaojie Qiu** is an Assistant Professor in the Department of Genetics at Stanford University, with a courtesy appointment in Computer Science, and a member of Bio-X, the Cardiovascular Institute, the Maternal & Child Health Research Institute, and the Wu Tsai Neurosciences Institute. He received his Ph.D. in Molecular & Cellular Biology and Genome Sciences from the University of Washington in 2018 and completed his postdoctoral training at MIT's Whitehead Institute. His research bridges single-cell and spatial genomics with machine learning and computational modeling to investigate mammalian cell fate transitions, with a particular focus on cardiac evolution, development, and disease. He is the principal developer of Monocle 2 & 3, the widely adopted framework for pseudotemporal trajectory analysis of single-cell RNA-seq data; Dynamo, a framework for RNA velocity vector field reconstruction and prediction of cell fate published in *Cell*; and Spateo, a toolkit for advanced spatiotemporal modeling of spatial transcriptomics data published in *Cell* — tools that have collectively become foundational infrastructure in the single-cell and spatial genomics community. He is a recipient of the NIH Director's New Innovator Award (2024), the NHGRI Pathway to Independence Award, and the Arc Ignite Award (2023), and was recognized as a Highly Cited Researcher by Clarivate in 2025. His lab was selected as a Laude Institute Moonshots Seed Grant winner for the Virtual Embryo project, chosen from 125 proposals evaluated by over 600 leading researchers. His work has appeared in *Nature*, *Cell*, *Nature Methods*, and *Nature Genetics*. He conceived and defined the scientific vision of the Virtual Embryo Challenge and leads its overall organization and dissemination.