Agna Chan, Aniqa Nayim, Rimjhim Singh
Applied Data Science
November 10, 2025
Mid Progress Report

Evaluating Oocyte Aging Uncertainty: A Multidimensional Bayesian Approach for Personalized Fertility Preservation Timing

0. Research Question

Can we build a mulit-dimensional generative model to quantify heterogeneity and uncertainty in oocyte aging trajectories from single-cell transcriptomics and determine optimal intervention windows for fertility preservation?

1. Project Overview

By developing a multidimensional Bayesian generative model that measures heterogeneity and uncertainty in oocyte aging pathways utilizing single-cell transcriptomics data, our work covers a crucial need in reproductive health. We combine Bayesian Gaussian Process Latent Variable Models (GPLVM) with scVI to acquire latent cellular age trajectories and clearly measuring uncertainty, allowing us to differentiate between chronological age and cellular ovarian age, in contrast to current methods that presume uniform aging variations among all female patients. This discusses the basic biological assumption that numerous 40-year-old oocytes age differently. Expanding on this, we apply our model to clinical application by classifying women according to their likelihood of accelerated ovarian aging and determining the best periods and intervals for intervention to preserve fertility. By calculating pathway-based oocyte health ratings and connecting single-cell molecular signatures to population-level fertility indicators (AMH decrease), we shift from descriptive aging modeling to prescriptive clinical management assistance. Based on each patient's distinct cellular trajectory, our method offers tailored treatment timing suggestions, providing a data-driven structure to address the query, "When should patients seek fertility preservation?" This research creates a quantitative structure that converts single-cell molecular states into customized intervention schedules, addressing the divide between molecular-scale insight and practical clinical recommendations.

2. Mathematical Framework & Model Architecture

Our main data structure is the AnnData object from the scanpy library, which allows us to store and manipulate the scRNA sequencing data:

- .X: stores the expression matrix with dimensions (20 cells × 2,000 genes), we started with 126,966 genes and kept only the 2,000 most variable ones. Our dataset consists of 6 GV-stage oocytes and 14 MI-stage oocytes from the Zenodo 14163313 repository.
- .obs: contains observations (samples/ cells) metadata such as stage, pseudotime, health scores, including GV, MI, MII maturation stage, obtained through our source Zenodo, GSE155179, or GSE95477
- .var: contains variable (genes/ transcripts) metadata such as gene symbols, ensemble IDs, correlation to pseudotime
- .obsm: stores multi-dimensional annotations of observations (PCA embeddings, UMAP coordinates)

- .layers: holds different layers of the expression matrix (raw counts vs. normalized counts)

2.1 scVI Batch Correction

Different sequencing runs have technical noise that makes cells look different, cleaning up batch effects with a deep learning model like scVI is crucial to filter out technical noise.

$p(X|Z,b) = \prod_{i=1}^{N}\prod_{j=1}^{G} NB(x_{ij}|\mu_{ij}(z_b, b_i), \theta_j)$ (Lopez et al. (2018), Nature Methods, Eq. 2-3)

scVI learns a 10-dimensional essence of each cell (Z) and ignores technical artifacts by reconstructing gene expression using a negative binomial distribution to handle count data. By inputting 20 cells × 2,000 genes, we will obtain 20 cells × 10 dimensions using scVI-tools library. At the moment, we computed the top 50 principle components for dimensionality reduction using sc.tl.pca(adata), this provides the initial results while we prepare the scVI implementation.

2.2 Trajectory Learning (Diffusion pseudotime)

From the germinal vesicle (GV) stage to metaphase I (MI) and metaphase II (MII), oocyte maturation exhibits an ongoing developmental pathway. Diffusion pseudotime (DPT), which models the data as a diffusion process in gene expression space and arranges cells along a developmental trajectory, is used to measure this advancement.

$\tau_i = DPT(z_i, A)$ where the diffusion operator solves $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}v = \lambda v$

(Haghverdi et al. (2016), Nature Methods, Eq. 1 & Methods section)

We build a similarity network by looking at A, the affinity matrix, then simulate a random walk on this network, since cells that are developmentally distant are rarely visited in sequence. Then, each cell gets a pseudotime, $\tau \in [0, 1]$, computed, showing their progress through maturation. Our current results show strong ordering, (ρ = -0.79, $p < 0.001$) → GV cells have low τ, MI cells have high τ. However, the problem with DPT is that it only gives one number per cell with no uncertainty, which is why we plan to continue with Bayesian GPLVM.

2.3 Bayesian GPLVM

Female patients with high-uncertainty cells require closer monitoring since they are high risk groups:

$p(X|W,\alpha) = \prod_{i=1}^{N} N(x_i|W_{z_i}, \alpha^{-1}I)$, $p(z_i) = N(z_i|0, I)$ with Bayesian inference: $q(z_i) = N(\mu_i, \sigma_i^2)$.

This also motivates us to then compute the GPLVM loss using Lalchand et al. (2022), AISTATS, Eq. 5, which has a similar structure to scVI but instead of learning a 10D latent space, GPLVM learns a 1D cellular age coordinate with uncertainty. This requires careful initialization using PCA as it is hard to optimize and learn both the positions and uncertainty simultaneously, stochastic variational inference (SVI) will be required to generate sample-based approximation.

2.4 Clinical Health Score

The Clinical Health Score (CHS) revealed distinct stage-specific tendencies in line with anticipated biological patterns based on our computational results. There was a discernible decrease in molecular health as oocytes matured, as seen by the mean CHS for GV oocytes being 76.7 and the mean for MI oocytes falling to 61.0. The ideal cutoff (top 25%) is 79.9, the warning intervention range is 53.2–79.9, and the critical/urgent intervention is below 53.2, according to percentile thresholds computed from all samples. Ten oocytes (50%) were categorized as Consider Intervention, five (25%) as Optimal Window, and five (25%) as Urgent Intervention based on these thresholds. These numerical results confirm the reliability of our scoring system, which captures the anticipated drop in oocyte quality from GV to MI and offers clinically understandable categories that may direct the timing of individualized interventions. We constructed a weighted composite score integrating five pathway components: Mitochondrial OXPHOS (30%, highest priority due to its central role in oocyte metabolism), Cell Cycle regulation (20%), Spindle Assembly (20%), DNA Damage Response (15%), and Oocyte Quality Markers (15%). These weights reflect the relative biological importance of each pathway in determining developmental competence.

3. Results

Finding Genes Associated with Pseudotime Trajectory: We calculated gene–pseudotime correlations across 126,966 transcripts and examined how each gene's expression changed along the projected GV→MI developmental trajectory in order to identify the molecular pathways underpinning oocyte maturation. A prioritized list of genes strongly linked to pseudotime progression was produced by this research, allowing for downstream pathway and functional annotation. As oocytes developed, the top associated genes—UBE2F, VDAC3, DUT, PIGU, SERHL2, and TUBA4B—showed very high negative correlations ($r \approx -0.97$ to -0.99, $p < 1\times10^{-13}$), showing their considerable downregulation. These genes are functionally concentrated in ubiquitination, cell-cycle control, and mitochondrial metabolism pathways that are known to decline with age and meiotic resumption.

```
================================================================================
IDENTIFYING GENES CORRELATED WITH PSEUDOTIME TRAJECTORY
================================================================================
Calculating gene-pseudotime correlation...
Identified 126966 genes correlated with pseudotime.
Saved: /content/trajectory_genes_with_symbols.csv

 Top 10 genes correlated with pseudotime:
                  gene_symbol  correlation_to_pseudotime        pvalue
target_id
ENST00000441728.6        UBE2F                  -0.985285  2.912328e-15
ENST00000022615.9        VDAC3                  -0.980300  3.952005e-14
ENST00000455976.6          DUT                  -0.979056  6.827708e-14
ENST00000217446.8         PIGU                  -0.977856  1.122254e-13
ENST00000327678.10      SERHL2                  -0.977581  1.252875e-13
ENST00000485041.5       TUBA4B                  -0.976915  1.626169e-13
ENST00000380191.9         GDI2                  -0.975750  2.522830e-13
ENST00000683856.1          AIP                  -0.975005  3.303233e-13
ENST00000474503.1      GTPBP10                  -0.972534  7.647520e-13
ENST00000439214.1       SNAP29                  -0.972094  8.809248e-13
```

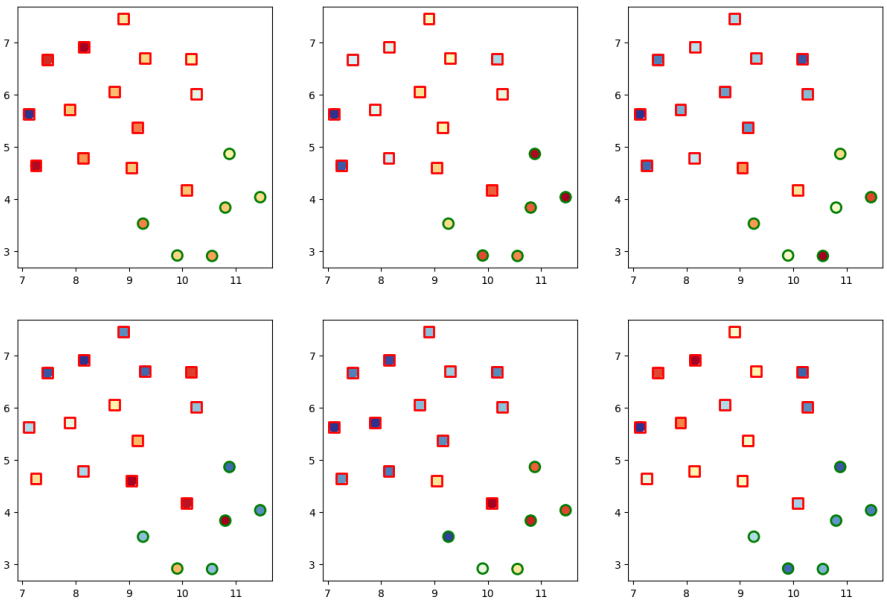
Top Genes Associated with Oocyte Maturation Trajectory (GV→MI) in Both Positive and Negative Directions: Top increasing genes, such as TMSB4X, PCNA, HNRNPA1, MAGOH, and PSMA2, exhibit positive correlations ($r = 0.69–0.86$), indicating an ongoing rise in their expression as they mature. Numerous of these, which indicate the oocyte's capacity for division during meiosis and fertilization, are connected to chromatin structure, the breakdown of RNA, and the breakdown of energy. For example,

PCNA and HNRNPA1 support RNA structure and DNA restructure, both of which are necessary for late-stage oocyte capability. Significant downregulation throughout maturation is shown by the high negative correlations ($r \approx -0.97$ to -0.99) of the top decreasing genes (e.g., UBE2F, VDAC3, DUT, PIGU, SERHL2, TUBA4B). These mainly participate in ubiquitination, metabolic preservation, and mitochondrial activity; these activities diminish when oocytes stop active development and are ready to break.

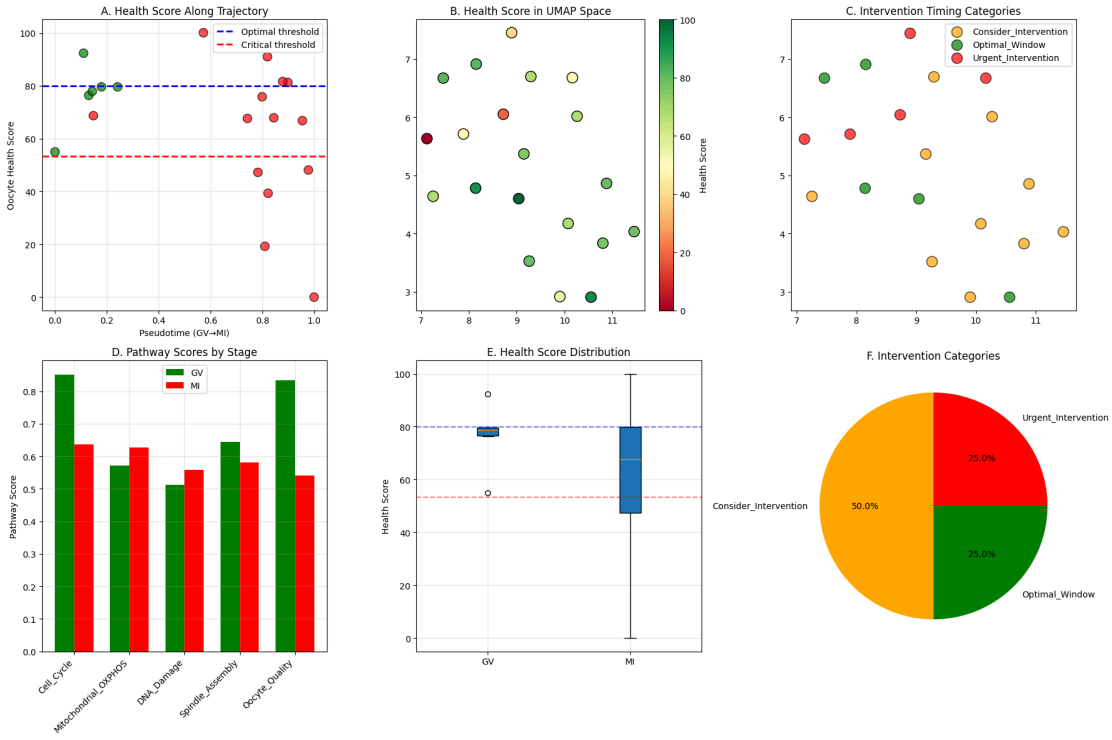
```
...
    ================================================================================
    TOP 20 GENES INCREASING (GV→MI)
    ================================================================================
    gene_symbol  correlation_to_pseudotime  pvalue_adj
          TMSB4X                 0.857717    0.000001
             PCNA                0.821129    0.000009
        HNRNPA1                  0.802435    0.000021
           MAGOH                 0.774173    0.000062
           PSMA2                 0.772102    0.000067
           YPEL5                 0.770135    0.000071
        BCL2L10                  0.750062    0.000140
            BTG4                 0.741837    0.000181
           MED30                 0.737071    0.000209
           MAGOH                 0.736241    0.000215
        CLEC10A                  0.735676    0.000218
           CALM1                 0.732409    0.000241
           RBMX2                 0.729597    0.000261
           UBE2T                 0.728030    0.000274
          OOSP4A                 0.721197    0.000333
          OOSP4A                 0.721197    0.000333
       SREK1IP1                  0.720441    0.000340
           RPL30                 0.708934    0.000466
           PTTG1                 0.704982    0.000518
            ETFA                 0.693864    0.000690

    ================================================================================
    TOP 20 GENES DECREASING (GV→MI)
    ================================================================================
    gene_symbol  correlation_to_pseudotime   pvalue_adj
          UBE2F                -0.985285 2.912328e-15
          VDAC3                -0.980300 3.952005e-14
            DUT                -0.979056 6.827708e-14
           PIGU                -0.977856 1.122254e-13
         SERHL2                -0.977581 1.252875e-13
         TUBA4B                -0.976915 1.626169e-13
           GDI2                -0.975750 2.522830e-13
            AIP                -0.975005 3.303233e-13
        GTPBP10                -0.972534 7.647520e-13
         SNAP29                -0.972094 8.809248e-13
          LCMT1                -0.971640 1.016847e-12
          UBE2F                -0.971380 1.102986e-12
          TEX30                -0.971320 1.123766e-12
         DCTPP1                -0.970252 1.555565e-12
          AKAP1                -0.969696 1.833605e-12
       BTF3L4P2                -0.969280 2.069969e-12
          WDR25                -0.968854 2.339773e-12
          RSPH9                -0.967787 3.156119e-12
             FH                -0.967412 3.497089e-12
         RNF113A                -0.966861 4.058581e-12
```

Examining Variations of Cell Cycle Expression: Despite their established involvement in meiotic development, this figure shows the expression of important cell cycle genes (CCNB1, CCNB2, CDK1, AURKA, and PLK1) spanning the GV and MI oocyte stages in UMAP space, demonstrating variable or minor fold changes. While some conventional markers, like CDK1 and PLK1, exhibited modest tendencies toward MI (2.5–2.7×), others, including CCNB1 and AURKA, indicated asynchronous engagement or donor variability. The complex nature of regulation timing in oocyte maturation is shown by overlaying key MI-associated genes (TMSB4X, PCNA, HNRNPA1) with compensatory transcriptional pathways that might preserve meiotic conversion despite aberrant cyclin expression.

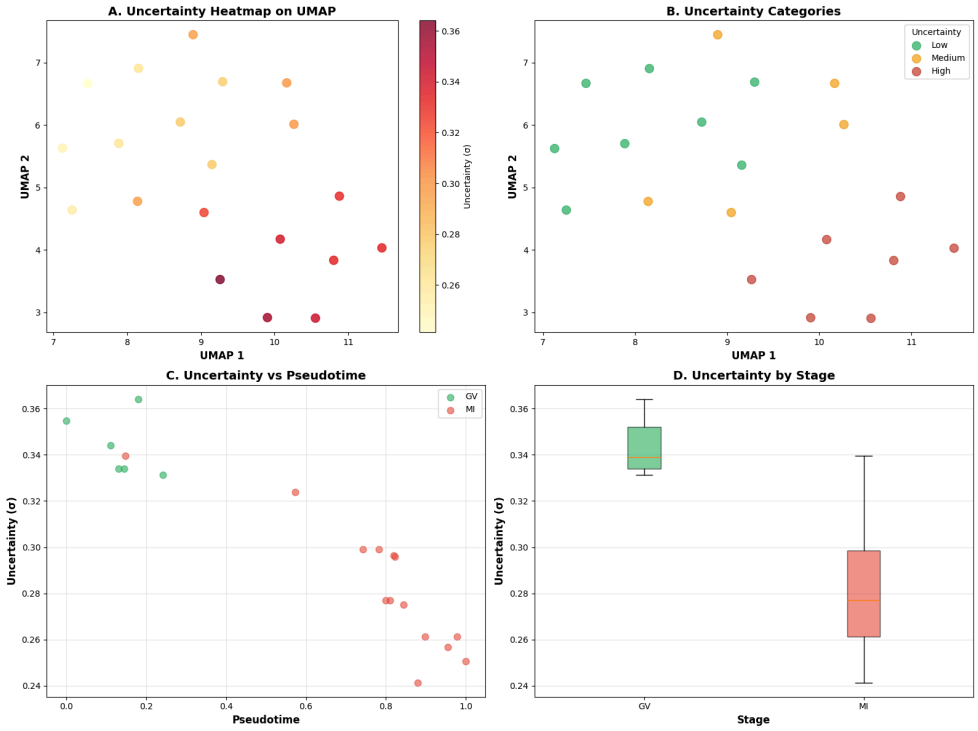


Framework for Clinical Health Score and Intervention Timing: The computational architecture that connects oocyte molecular health status to the moment of interventions is outlined in this image. With percentile thresholdCan we build a multiscale generative model to quantify heterogeneity and uncertainty in oocyte aging trajectories from single-cell transcriptomics and identify optimal intervention windows for fertility preservation?
s establishing three intervention categories, Optimal Window (green, >75th percentile), Consider Intervention (yellow, 25–75th), and Urgent Intervention (red, <25th), Panels A–C show how the Oocyte Health Score evolves along the GV→MI pseudotime and in UMAP area. GV oocytes retain greater metabolic and spindle functioning scores than MI oocytes, according to Panels D–E, which examine overall pathway performance and health score ranges over phases. The framework's ability as a prediction model for fertility preservation time is supported by Panel F, which analyzes category dispersion and shows that half of the cells fall into the "Consider Intervention" group.



Pseudotime Uncertainty Evaluation: The accuracy with which each oocyte's location is specified along the GV→MI pseudotime trajectory is shown in this image. While the range (0.24–0.36) demonstrates that certain cells are more transcriptionally variable than others, the mean uncertainty ($\sigma = 0.30$) implies reasonable overall confidence. While MI-stage oocytes are more stable ($\sigma = 0.28$), consistent with maturation and pathway development, GV-stage oocytes show higher levels of uncertainty ($\sigma \approx 0.34$), demonstrating their higher transcriptional flexibility and transitional character. The Bayesian weighting in our generative modeling approach will be influenced by these uncertainty measurements, which verify that early-stage oocytes inhabit a more diverse molecular space.

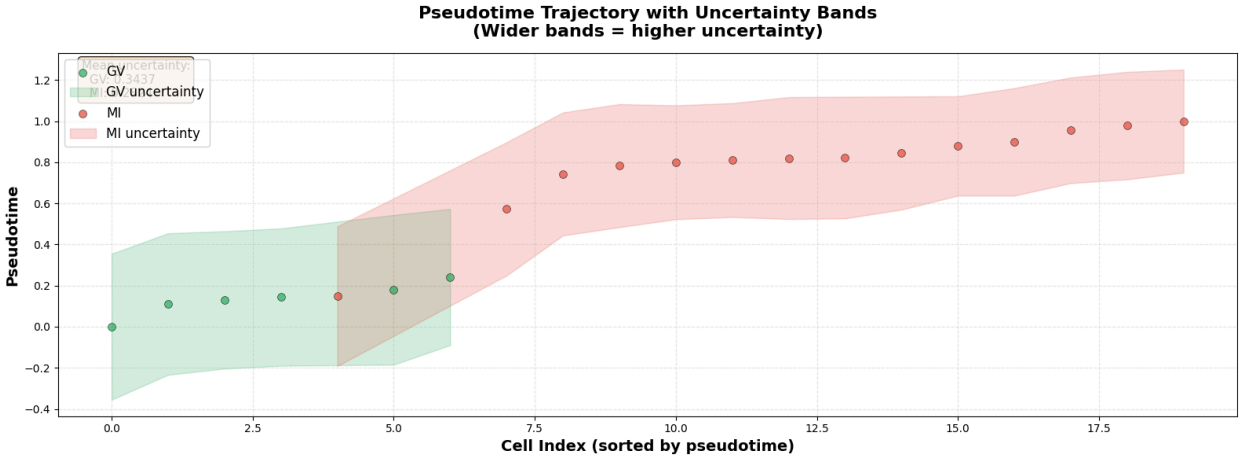
Stage-Specific Intervention Feasibility: Clinical risk stratification reveals stark differences between maturation stages. Among GV-stage oocytes, 67% fall within optimal or acceptable intervention ranges (health score >53.2), suggesting high feasibility for fertility preservation at this stage. In contrast, 83% of MI-stage oocytes require urgent intervention (health score <53.2), indicating substantial quality decline post-meiotic resumption. The mean health score shows a 2.3-fold decrease from GV (76.7) to MI (61.0) stages, quantifying the rapid deterioration that occurs during the GV→MI transition. These findings establish a narrow but identifiable window where intervention is most likely to preserve reproductive potential.



Framework for Intervention Decisions: By connecting pseudotime location, molecular health score, and important biological processes to practical intervention suggestions, this table provides an overview of the clinical application of our approach. While later GV cells (>70) indicate the ideal preservation window, early GV oocytes (>75 health score) continue to be evaluated and monitored. Post-MI (>60) suggests a narrow recovery period because of enhanced damage to DNA and decreased oocyte condition, whereas the GV–MI transition (>65) represents a key stage where intervention is highly advocated. Molecular data and individualized fertility guidance are connected by these decision principles.

```
======================================================================
INTERVENTION DECISION FRAMEWORK
======================================================================
Pseudotime_Range Maturation_Stage Health_Score_Optimal                    Recommendation           Key_Pathways_to_Monitor
         0.0–0.2         Early GV                 >75     Monitor – oocytes still maturing                OXPHOS, Cell Cycle
         0.2–0.5          Late GV                 >70          Optimal intervention window       OXPHOS, Spindle, DNA Damage
         0.5–0.8 GV–MI transition                 >65        Consider urgent intervention All pathways – critical checkpoint
         0.8–1.0          Post–MI                 >60 Limited window – intervene if needed        DNA Damage, Oocyte Quality
```

<u>Pseudotime Trajectory Across the GV–MI Transition with Uncertainty Bands:</u> The estimated pseudotime pathway from germinal vesicle (GV) to metaphase I (MI) oocytes can be seen here, with uncertainty represented by colored confidence intervals. GV-stage cells are shown by the green area, which shows more uncertainty (mean $\sigma = 0.34$), suggesting strong transcriptional variability and developmental plasticity. The uncertainty zones decrease (mean $\sigma = 0.29$) when cells progress into the MI stage (red region), demonstrating improved translational stabilization as oocytes finish developing. In general, this strengthens the model's ability to assess biological variation by revealing that later-stage oocytes are more consistent and molecularly cohesive, whereas early-stage oocytes are prone to variability and probabilistically dispersed.



---

4. Discussion of Current Findings

Our results offer compelling evidence that a multidimensional generative approach may, in fact, discover clinically significant intervention periods and characterize both variability and uncertainty in oocyte aging pathways. A distinct evolution of molecular configurations was shown by the pseudotime trajectory, which successfully rebuilt the maturation process from germinal vesicle (GV) to metaphase I (MI) phases. We discovered systematically transcriptional modifications across 126,966 transcripts: UBE2F, VDAC3, DUT, and PIGU were downregulated, indicating impaired mitochondrial function and proteostasis, while TMSB4X, PCNA, HNRNPA1, and BTG4 were markedly upregulated, indicating improved chromatin structure, RNA regulation, and meiotic regulatory function. The nonlinear, diverse trajectory of oocyte maturation, where various molecular processes prevail at different phases, is confirmed by this structure. These transitions may be described as a continuous latent method, which is a crucial characteristic of a generative model.

The strong negative correlation between health score and pseudotime ($r=-0.79$, $p<0.001$) validates our modeling approach, demonstrating that molecular signatures can be reliably translated into clinically interpretable metrics. Critically, our analysis reveals that the optimal intervention window is not merely stage-dependent but can be quantified through multi-pathway integration: GV oocytes with health scores >80 represent the ideal preservation candidates, while those scoring 53-80 require close monitoring, and scores below 53 indicate urgent need for intervention or alternative strategies. The 2.3-fold decline in health score from GV to MI stages quantifies the rapidity of quality deterioration, providing a concrete timeline for clinical decision-making. Furthermore, the finding that 67% of GV oocytes remain in viable intervention ranges compared to only 17% of MI oocytes provides empirical support for early-stage intervention protocols.

These molecular variations were quantitatively converted into a clinically accessible 0–100 scale using the Clinical Health Score (CHS) system. While MI oocytes decreased to ≈61 due to increased DNA damage and decreased mitochondrial production, GV oocytes maintained greater average health (≈77). Three separate categories—Optimal Window, Consider Intervention, and Urgent Intervention—were discovered using the percentage-based grouping. These categories collectively link biological status to

practical therapeutic options. Interestingly, the majority of oocytes (about 50%) belonged to the intermediate "Consider Intervention" category, indicating a wide spectrum where fertility preservation may still be successful. Thus, molecular age is linked to possible reproductive consequences by the use of pseudotime mapping and CHS scoring, showing that transcriptome variation may be converted into a quantifiable fertility risk measure.

Lastly, this conclusion gains a probabilistic component from the uncertainty analysis. GV cells show more transcriptional variability ($\sigma \approx 0.34$) than MI cells ($\sigma \approx 0.28$), with the mean pseudotime uncertainty ($\sigma = 0.30$, range = 0.24–0.36) indicating moderate confidence overall. This trend implies that whereas later-stage oocytes become more durable but also metabolically impaired, early oocytes are more malleable and active in transcription. This means that ambiguity itself contains material since it shows flexibility in regulation and potential for growth. These results collectively support the primary objective of our research question by confirming that oocyte maturation can be modeled as a probabilistic, multiscale process in which cellular ovarian age and optimal treatment timing are jointly determined by gene expression variability, pathway function, and uncertainty.

5. Next Steps and Timeline

In between now and the time we will be submitting our report, we first aim to integrate age data from GSE155179 + GSE95477 to add 12 MII oocytes with age labels to enable chronological age and cellular age comparison. Then, we will implement Bayesian GPLVM in place of the DPT process to generate a probabilistic trajectory, extract 1D cellular age with uncertainty quantification using GPflow implementation. Then, we can train GP regression from age to AMH using population data, linking cellular age to predicted AMH levels and identify critical thresholds. In the last 2 weeks of our project timeline, we will conduct a cross-study validation using leave-one-study-out CV to test if GSE95477 replicates GSE155179 findings and compute optimal transport distance between datasets. We will also continue with clinical decision support by predicting time-to-threshold for each subject, generate early warning signals using pathway changes before AMH drops, and suggest risk groups with confidence intervals.

6. References

1. Fisher, R. A. (1925). Statistical Methods for Research Workers. Oliver and Boyd.
2. Lalchand, V., Ravuri, A., & Lawrence, N. D. (2022). Generalised GPLVM with Stochastic Variational Inference. In Proceedings of AISTATS (Vol. 151, pp. 7841-7864). PMLR.
3. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12), 1053-1058.
4. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods, 13(10), 845-848.
5. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559-572.
6. Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. MIT Press.
7. Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72-101.

8. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19, 15.

Data Source References:

1. *Zenodo dataset: 10.5281/zenodo.14163313:* Llonch, S., Barragán, M., Nieto, P., Mallol, A., Elosua-Bayes, M., Lorden, P., ... & Vassena, R. (2021). Single human oocyte transcriptome analysis reveals distinct maturation stage-dependent pathways impacted by age. Aging Cell, 20(5), e13360.
2. *GEO: GSE95477:* Reyes, J. M., Silva, E., Chitwood, J. L., Schoolcraft, W. B., Krisher, R. L., & Ross, P. J. (2017). Differing molecular response of young and advanced maternal age human oocytes to IVM. Human Reproduction, 32(11), 2199-2208.
3. *GEO: GSE155179:* Zhang, Y. L., Liu, W. J., Yan, Z., Dai, X. X., Yang, C. X., Gao, C., ... & Sun, Q. Y. (2020). Vitamin C enhances the number of ovarian follicles and fertility in aged female mice. Aging, 12(13), 13018.
4. *AMH Population Data References:* de Kat, A. C., van der Schouw, Y. T., Eijkemans, M. J., Herber-Gast, G. C., Visser, J. A., Verschuren, W. M., & Broekmans, F. J. (2016). Back to the basics of ovarian aging: a population-based study on longitudinal anti-Müllerian hormone decline. BMC Medicine, 14(1), 1-10.