KAIROS: math companion.

Speaker notes for the colloquium equations. Not for the audience.

A short walkthrough of every formula that appears in the deck, with the intuition, the derivation, and the pitfalls I want to be ready to defend if asked.

1. Set-up: the position-by-time tensor

For each gene g we build a matrix

$$R_g \in \mathbb{R}^{M \times T}, \qquad R_g[x,t] = \text{read count in bin } x \text{ at time } t.$$

M is the number of 250-bp position bins along the gene body (typically 50–300, depending on gene length). T is the number of sampled time points (here $T = 4$: 0, 10, 25, 40 minutes after estradiol).

The matrix is the full measurement. Everything downstream compresses it into a summary statistic — either position-by-position or gene-wide.

2. The $1/v$ principle (the kinetic identity)

The core physical intuition: at position x, a polymerase moving at speed $v(x)$ spends time $1/v(x)$ per unit length. In a GRO-seq run-on snapshot, the number of nascent RNA reads accumulated at that bin by time t grows proportionally to that dwell time, scaled by the initiation rate.

$$\mathbb{E}[R_g[x,t]] = \frac{c\,t}{v(x)} \qquad (c \text{ absorbs initiation rate and depth})$$

Consequence: fit a straight line across the time course at each position:

$$R_g[x,t] = \alpha_g(x) + \beta_g(x)\,t + \varepsilon.$$

The slope is the reciprocal of speed, up to the constant c:

$$\beta_g(x) = \frac{c}{v(x)} \qquad \Longrightarrow \qquad \widehat{v}_g(x) \propto \frac{1}{\widehat{\beta}_g(x)}.$$

Pitfall to flag.

The proportionality constant c is not recovered by the per-position regression. All $\widehat{v}(x)$ values are velocities *up to a gene-specific global scale*. This is why we cannot compare absolute speeds across genes from $\widehat{\beta}$ alone. We report *relative* velocity profiles within a gene, and use ψ or DANKO for cross-gene magnitude.

Pitfall to flag.

At early times and distal positions, $R_g[x,t] = 0$ because the wave hasn't arrived. Fitting a slope through an all-zero column is meaningless. We classify each position as pre-wave, transition, or elongated (three-state logistic on cumulative counts) and only fit slopes on transition-plus-elongated positions.

3. Robust regression

Because GRO-seq counts are heavy-tailed and PCR duplicates create occasional spikes, we replace ordinary least squares with Huber's M-estimator:

$$L_c(u) \;=\; \begin{cases} \tfrac{1}{2}\,u^2, & \text{if } |u| \le c, \\ c\,|u| - \tfrac{1}{2}\,c^2, & \text{if } |u| > c, \end{cases}$$

and solve $\widehat{\beta}(x) = \arg\min_\beta \sum_i L_c(R[x,t_i] - \alpha - \beta t_i)$. Tuning $c = 1.345\,\widehat{\sigma}$ gives 95% efficiency at Gaussian noise while capping the leverage of a single outlier at c.

Why not Poisson GLM?

For $T = 4$, the Poisson likelihood is dominated by the small-count regime where maximum-likelihood estimates are unstable; the Huber linear model is more forgiving and the slope retains its read-per-minute interpretation.

4. Gene-level summary I: algebraic diversity Z_M

The per-position slope vector $\widehat{\boldsymbol{\beta}}_g = (\widehat{\beta}_g(x_1), \ldots, \widehat{\beta}_g(x_M))$ lives in $\mathbb{R}^M$. We want one scalar per gene.

The diversity summary follows the algebraic-diversity framework of M. A. Thornton (arXiv:2604.03634, 2026; arXiv:2604.03725, 2026), which establishes that a group-averaged estimator from a single observation recovers the spectral structure of a covariance or density operator. Specialized to the kinetic setting, we treat $\{\widehat{\beta}_g(x)\}$ as elements of the abelian group $(\mathbb{R}, +)$ and count effectively distinct group elements, weighted by a continuous resolution function f:

$$Z_M \;=\; \frac{1}{M-1} \sum_{x=1}^{M} \mathbf{1}[\widehat{\beta}(x) \neq 0]\; f(\widehat{\beta}(x)).$$

In practice we use $f(\beta) = |\beta|$ (magnitude) or $f(\beta) = |\beta - \bar{\beta}|$ (spread around the mean). The estimator is monotone in the number of positions with non-trivial signal and the variation among them.

Why not the mean slope?

The mean is dominated by pause spikes, ignores concentration, and cannot distinguish one long slow patch from many brief ones. Z_M separates these.

5. Gene-level summary II: spectral concentration ψ

The more powerful summary comes from the singular-value decomposition of the full tensor slice R_g:

$$R_g \;=\; \sum_{j=1}^{\min(M,T)} \lambda_j \, \mathbf{u}_j \, \mathbf{w}_j^{\top}, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq 0.$$

- $\mathbf{u}_j \in \mathbb{R}^M$ is the j-th *position mode*.
- $\mathbf{w}_j \in \mathbb{R}^T$ is the j-th *time mode*.
- λ_j^2 is the *energy* of that mode in the total Frobenius norm.

The spectral concentration is the fraction of total energy in the first mode:

$$\psi(g) \;=\; \frac{\lambda_1^2}{\sum_j \lambda_j^2} \;\in\; (0, 1].$$

Why this tracks wave-front speed.

A clean elongation wave *is* a rank-one matrix up to noise: the position profile $\mathbf{u}_1$ is the shape of the wave; the time profile $\mathbf{w}_1$ is its growth through the time course. A gene with a single crisp wave has $\psi \approx 1$. A gene where the wave is interrupted by multiple pauses, or where read accumulation is diffuse across position and time, has ψ near the uniform value $1/\min(M, T)$.

Interpretation of the main result.

On 2,500 genes,

$$\text{Spearman}(\psi,\; v^{\text{DANKO}}) \;=\; 0.557, \qquad p < 10^{-12}.$$

ψ is not a substitute for the HMM's estimate of magnitude — it does not give you kilobases per minute. It gives you a *ranking* of genes by how coherently they elongate, and that ranking tracks a standard wave-front rate remarkably well, for a fraction of the compute.

6. DANKO wave-front HMM (the anchor we validate against)

For each gene, at each time point t, fit a two-state hidden Markov model along position:

- S_0 = *pre-wave*, low emission mean μ_0 (negative binomial).
- S_1 = *post-wave*, high emission mean $\mu_1 \gg \mu_0$.
- Transition $S_0 \to S_1$ at probability p_{01}.
- S_1 is *absorbing*: once elongation has started at a bin, it stays started.

Baum-Welch gives the maximum likelihood parameters; Viterbi gives the most likely breakpoint position $\widehat{x}_g^\star(t)$. The gene-wide elongation rate is simply

$$\widehat{v}_g^{\text{DANKO}} \;=\; \widehat{x}_g^\star(t)/t,$$

measured at $t = 40$ minutes.

Why we reimplemented it in Python.

R groHMM is 14 minutes single-threaded on 2,500 genes; our port is 7 minutes on 8 workers with identical (Pearson 0.96) output, and it exposes the internals (forward-backward, emission model, Viterbi decoding) for extension.

7. Epigenetic covariates: the mixed-effects model

For each position x in each gene g, assemble a feature vector $\mathbf{z}_g(x) \in \mathbb{R}^K$ with $K = 10$ chromatin and sequence covariates. The model:

$$\log \widehat{v}_g(x) \;=\; \mathbf{z}_g(x)^\top \boldsymbol{\gamma} \;+\; u_g \;+\; \eta(x) \;+\; \varepsilon.$$

- $\boldsymbol{\gamma} \in \mathbb{R}^{10}$: fixed effects of the K covariates on log velocity.
- $u_g \sim \mathcal{N}(0, \sigma_g^2)$: gene-level random intercept.
- $\eta(x)$: positional spline soaking up TSS / polyA structure that is not covariate-driven.
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$: residual noise.

The log-link is important: it makes effects multiplicative and prevents negative velocity estimates. Positive entries of $\widehat{\boldsymbol{\gamma}}$ are accelerators; negative entries are stallers.

Top hits.

H3K36me3 (+0.22 per z), H3K27ac (+0.14), DNase accessibility (+0.11) accelerate; H3K27me3 (-0.18), CpG methylation (-0.09), GC content (-0.08) stall.

8. Residual pause detection

Pause candidates are positions where the observed velocity is much lower than what the chromatin model predicts:

$$\rho_g(x) \;=\; \log \widehat{v}_g(x) \;-\; \mathbf{z}_g(x)^\top \widehat{\boldsymbol{\gamma}}.$$

Large-magnitude negative $\rho_g(x)$ flags a slow position that is unexplained by local chromatin, i.e. a candidate regulated pause. We call bins with $\rho < -1.8$ at FDR < 0.05.

9. The recurring pattern

Every KAIROS estimator follows the same recipe:

1. Build the space-time tensor R_g for each gene.
2. Produce a position-indexed quantity from it: $\widehat{\beta}(x)$ (slope), singular values λ_j, or HMM breakpoint.
3. Compress to a scalar via linear algebra (Z_M, ψ, or v/t).
4. Validate against an external anchor (DANKO HMM, simulation, chromatin covariates).

The philosophy: replace hidden-state inference with spectral / algebraic summaries wherever possible. Cheaper, more transparent, and in our dataset, nearly as good.