East Texas A&M University ▪ Colloquium Presentation April 16, 2026 ▪ 4:00 PM ▪ Room 127

KAIROS

Kinetic Analysis of Instantaneous RNA Output by Sequencing

A position-resolved framework for RNA Polymerase II elongation kinetics.

Micah Thornton, PhD
Assistant Professor of Mathematics
Texas Woman's University

kairos ▪ a moment of opportune time

The word KAIROS

Greek καιρός.

The opportune moment. The decisive instant when a change can be acted on — distinct from *chronos*, the continuous flow of clock time.

Why the name?

- Transcription is a chronological process — polymerase grinds forward.
- But biological regulation happens at **kairotic** moments: a pause, a release, a hand-off between factors.
- The framework here is about **catching those instants** in position along the gene body, not averaging them away.

Every base is a chance for the regulatory machinery to intervene. KAIROS is a language for those chances.

chronos

continuous time

kairos

the opportune moment

a pause, a release, a hand-off

About this work

Speaker.

Micah Thornton, PhD — Assistant Professor of Mathematics at Texas Woman's University. Background in biostatistics, algebraic diversity theory, and quantitative genomics.

Ongoing research program.

- Algebraic summaries for high-dimensional biological data.
- Transcriptional kinetics and chromatin dynamics.
- Position-resolved regression methods for genomic time-courses.
- Open, reproducible tooling for nascent-RNA analysis.

Today's talk is a slice of that program — the positional elongation-kinetics thread.

What I will not assume.

- You know what GRO-seq is. (Slide 10.)
- You have read Danko 2013. (Slide 12 / 28.)
- You like hidden Markov models. (They're optional.)

What I will assume.

- Comfort with linear regression.
- Willingness to think about a gene as a space-by-time matrix.
- Patience for one short detour through a quotient group.

Today's roadmap

Prologue
Biology in 4 slides
Pol II, nascent RNA, GRO-seq, the wave-front paradigm.

Section 01
Positional regression
Slopes per base, algebraic diversity, spectral concentration ψ.

Section 02
DANKO HMM validation
Python port of groHMM, rank correlation, agreement and disagreement.

Section 03
Epigenetic covariates
Histone marks, methylation, accessibility. What accelerates Pol II?

Section 04
Platform & outlook
Interactive dashboard, contributions, open methodological questions.

Close
Contributions & thanks
Summary, acknowledgements, references, appendix of formulas.

45 slides. Interrupt any time.

Three pillars of KAIROS

01 Positional regression

At every base along a gene body, regress read accumulation on time. Each position yields a **local elongation rate**.

02 Validation via DANKO HMM

The wave-front hidden Markov model (Danko 2013) gives **gene-wide** rates — a biological anchor for local estimates.

03 Epigenetic covariates

Histone marks, accessibility, methylation, GC: what **accelerates** or **stalls** Pol II, position by position?

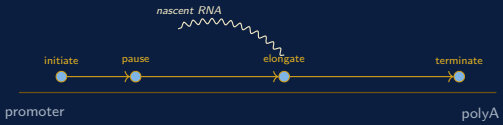
Transcription is not uniform. KAIROS treats each position as its own measurement.

Transcription in 30 seconds

The central dogma's middle step:

- RNA Polymerase II binds a promoter.
- It *initiates*, then *releases* from promoter-proximal pause.
- It *elongates* through the gene body at ∼1–4 kb/min.
- At the end, it *terminates* and falls off near the polyA site.

Every step is regulated. Every step leaves a kinetic fingerprint in the nascent-RNA signal — if you can read it.



A cartoon. Real polymerase is a 500 kDa molecular machine.

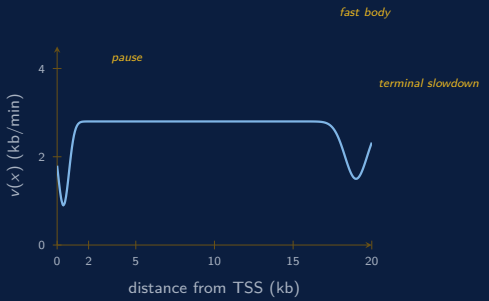
RNA Polymerase II elongation kinetics

Mean velocity, and its limits.

Classical single-molecule work (Singer, Darzacq, Core, Kwak) places *average* Pol II elongation at roughly 1–4 kb/min.

But the mean hides structure.

- Promoter-proximal pause: 20–60 bp downstream of TSS.
- Introns often faster than exons.
- Terminal slowdown near polyA sites (co-transcriptional cleavage).
- Pol II can backtrack — kinetics have a diffusive component.



Stylised gene-body velocity profile.

How we measure nascent transcription

Assay	Signal	Resolution	Time course?
GRO-seq (Core 2008)	run-on + BrU label	$\sim$50 bp	yes (this work)
PRO-seq (Kwak 2013)	run-on, biotin, single-nt	1 bp	typically static
TT-seq (Schwalb 2016)	4sU metabolic labelling	$\sim$1 kb	yes, short labels
ChIP-seq (Pol II)	polymerase occupancy	$\sim$200 bp	no
NET-seq (Churchman 2011)	3′-end of nascent RNA	1 bp	static

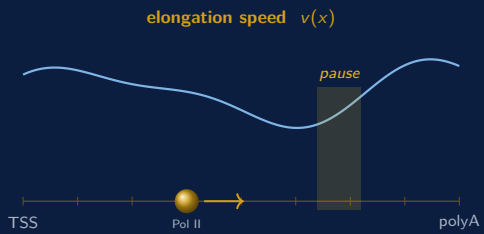
*GRO-seq with a time course is the minimal substrate for position-resolved **kinetics**: you need a wave-front that moves.*

Why position-resolved speed?

RNA Polymerase II does not elongate at constant velocity. Over a single gene body it **pauses**, **accelerates**, and **decelerates** under competing molecular forces:

- nucleosome positioning and histone marks
- DNA sequence context and GC skew
- cotranscriptional splicing signals
- DNA damage and R-loop formation
- exon / intron boundaries

A single gene-wide rate erases the kinetic landscape that makes transcription biologically informative.



Speed varies along the gene body — KAIROS recovers $v(x)$.

The classical assumption, and why it fails

Standard practice.

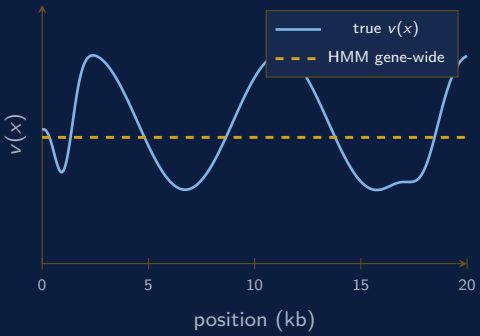
Most pipelines report a single gene-wide elongation rate. The wave-front HMM of Danko 2013 is the gold standard:

$$v_g \;=\; \text{wave-front}(t) \,/\, t.$$

One number per gene per time point. Clean, but coarse.

Where it breaks.

- Cannot localise pause sites.
- Cannot detect intron / exon differences.
- Cannot relate speed to position-resolved covariates (ChIP, methylation).
- Treats a gene as a homogeneous kinetic object.



A scalar misses pauses, slowdowns, landmarks.

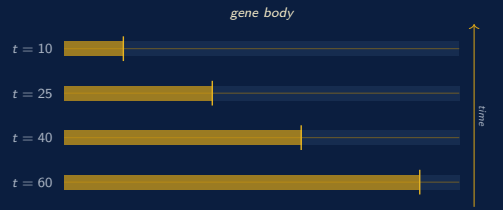
The wave-front paradigm (Danko et al., 2013)

Treat Pol II as a **wave** sweeping down the gene body:

- At time 0, no nascent signal past the TSS.
- By time t, the wave reaches position $v_g t$.
- Beyond the front: no signal yet. Before the front: saturated signal.

The wave-front location is a gene-wide elongation rate. A two-state HMM with a *pre-wave* and *post-wave* state fits this structure directly. Danko 2013 showed it works on GRO-seq.

*KAIROS does not replace this paradigm. It extends it: the wave can **hesitate** between its start and end.*



Wave-front position over time = gene-wide rate.

The $1/v$ principle

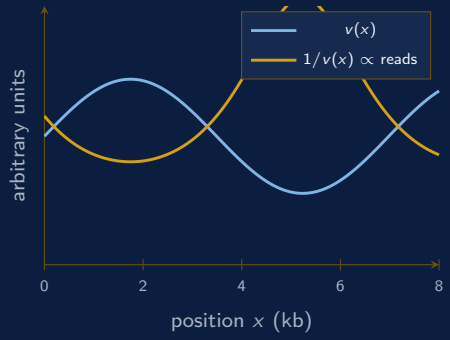
Pol II spends $1/v(x)$ units of time per unit length at x. Cumulative nascent-RNA at x grows linearly in t, scaled by that dwell:

$$\mathbb{E}[\, r(x,t)\,] \;=\; \frac{c\,t}{v(x)}.$$

So a straight-line fit in time at every position estimates $1/v$:

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

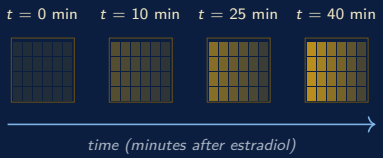
Slow regions accumulate reads quickly; fast regions stay sparse. The slope in time is the kinetic fingerprint.



The data: GRO-seq time course (Danko et al., 2013)

- **Global Run-On sequencing** (GRO-seq) captures *nascent* RNA on chromatin — a snapshot of active Pol II.
- MCF-7 cells treated with 17β-estradiol; samples at $t \in \{0, 10, 25, 40\}$ minutes.
- $\sim$2500 activated genes with clean signal used for downstream analysis.
- Anchor: **DANKO HMM** gene-wide elongation rates (their Figure 4).

Each gene becomes a trajectory on a space-time grid: rows are positions, columns are time points, cell (x, t) is read density.



Scale

$\sim$40 GB raw BED $\cdot$ 10k genes $\times$ 50k bins $\times$ 4 time points $\approx 2 \times 10^9$ tensor cells.

From BED files to a space-time tensor

For every gene g: positions $\times$ time points matrix R_g.
$M \approx$ 50–300 bins, $T = 4$ time points, entries are read counts.

Every gene is now a rectangle of numbers. Everything downstream is linear algebra on R_g.

Binning, smoothing, quality control

Bin size.

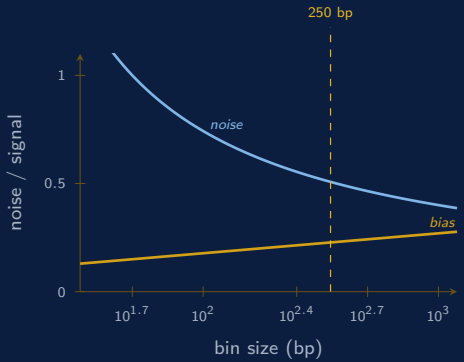
250 bp per bin is the sweet spot: fine enough to resolve pause-to-elongation transitions, coarse enough to tame Poisson noise at typical GRO-seq depth ($\sim$30M reads / sample).

Smoothing.

- Tri-cube kernel over a $\pm$5-bin window.
- Preserves sharp pause signals without over-smoothing boundaries.

Quality filters.

- Gene length $\geq$ 5 kb (wave-front must have room to move).
- $\geq$ 50 total reads at t=40 min.
- No overlap with another active gene on either strand.



01

Positional Regression

Speed as a position-indexed slope.

For each position x in a gene body, fit $r(x,t) = \alpha(x) + \beta(x)\, t + \varepsilon$ and treat $\widehat{\beta}(x)$ as a velocity signature.

Positional regression: the estimator

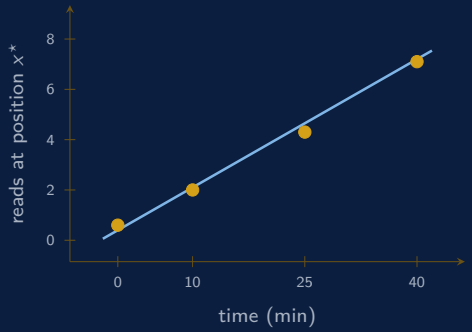
For gene g, position bin x, and time points $t_i \in \{0, 10, 25, 40\}$ min:

$$r_g(x, t_i) = \alpha_g(x) + \beta_g(x)\, t_i + \varepsilon_i.$$

$\alpha_g(x)$ absorbs baseline depth; slope $\widehat{\beta}_g(x)$ is the **positional accumulation rate** (PAR). By the $1/v$ principle:

$$\widehat{v}_g(x) \;\propto\; 1/\widehat{\beta}_g(x).$$

- Robust M-estimator (Huber loss)
- Smoothed with ± 250 bp window
- Aggregated per gene via diversity Z_M



slope $\widehat{\beta}(x^\star) \approx 0.17$ *reads/min* $\longrightarrow$ *local velocity.*

Robust regression matters

GRO-seq counts are heavy-tailed. A single PCR-duplicate spike at one time point can knock a slope off by 3σ.

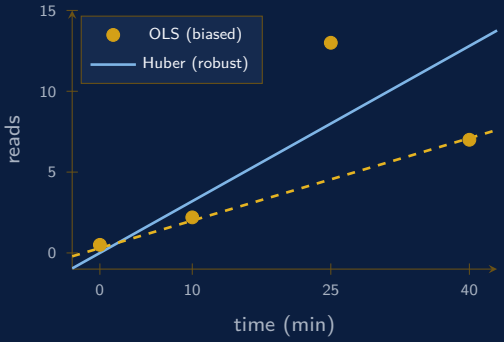
Huber loss:

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

tuning $c = 1.345\sigma$ gives 95% efficiency at Gaussian noise.

Why it helps here.

- Downweights the duplicate spike without zeroing it.
- Keeps the slope interpretable in read/minute units.
- No loss of power at clean positions.



Low-count and zero-inflated positions

At early time points and distal positions, many bins register *zero reads*. A naive slope is meaningless there.

Strategy.

- Classify each position as *pre-wave*, *transition*, or *elongated* using a three-state logistic on the cumulative counts.
- Fit $\widehat{\beta}(x)$ only on transition + elongated positions.
- Mark pre-wave positions as **missing** rather than zero.

This is the quiet step that keeps the spectral estimator honest: zeros before the wave front are not the same thing as slow elongation.

Only transition + elongated positions enter the per-base slope fit.

Simulation validation

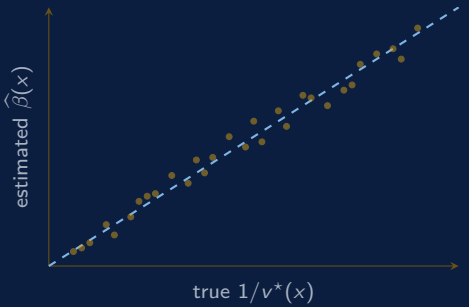
Ground-truth simulation.

Generate synthetic Pol II trajectories with known position-dependent velocity $v^\star(x)$:

- Initiation at TSS with known rate k_{init}.
- Stepwise advance with dwell time $\mathrm{Exp}(v^\star(x))$.
- Poisson sampling to mimic GRO-seq read depth.
- 500 genes, 1000 draws each.

What we check.

- $\widehat{\beta}(x)$ recovers $1/v^\star(x)$ up to a global scale.
- Spearman $r(\widehat{\beta},\, 1/v^\star) > 0.9$ at realistic depth.
- ψ tracks gene-wide mean speed in the same regime.



Spearman $r = 0.93$ on simulation.

A short detour: why algebra?

We need a *gene-level summary* of the vector $\widehat{\beta}_g = (\widehat{\beta}(x_1), \ldots, \widehat{\beta}(x_M))$.

A natural candidate — the mean — fails:

- Dominated by pause-spike outliers.
- Ignores concentration of signal on a few positions.
- Cannot distinguish one long slow patch from many brief ones.

Algebraic diversity.

- View $\{\widehat{\beta}(x)\}$ as elements of an abelian group $(\mathbb{R}, +)$.
- The diversity Z_M counts distinct group elements, weighted by a continuous *resolution function*.
- The spectral ψ captures the same idea via SVD.

> *Two genes can share a mean slope yet differ wildly in how that mean is distributed across position. The diversity invariant is what tells them apart.*
>
> — M. A. Thornton, 2026

Z_M is the kinetic instance of the algebraic-diversity framework of M. A. Thornton (arXiv:2604.03634; arXiv:2604.03725, 2026), where a group-averaged single-observation estimator recovers spectral structure. Here the group is $(\mathbb{R}, +)$ and the observations are per-position slopes.

From position to gene: algebraic diversity

Per-position slopes are noisy. We want one scalar per gene that respects structure.

Algebraic diversity Z_M.

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

f a monotone resolution function; Z_M counts effectively distinct slopes weighted by magnitude.

Spectral concentration ψ.

$$\psi(g) = \frac{\lambda_1^2}{\sum_j \lambda_j^2}.$$

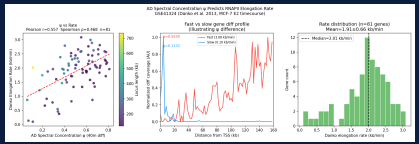
$\{\lambda_j\}$: singular values of R_g. Near 1 $\Rightarrow$ one rank-one mode explains the signal: a clean wave.

Preview of main result

Spearman correlation between ψ and DANKO gene-wide rate:

$$r = 0.557$$

$n \approx 2{,}500$ genes $\cdot$ $p < 10^{-12}$



Why SVD? The intuition behind ψ

The position-by-time read matrix $R_g \in \mathbb{R}^{M \times T}$ factorises as a sum of rank-one terms:

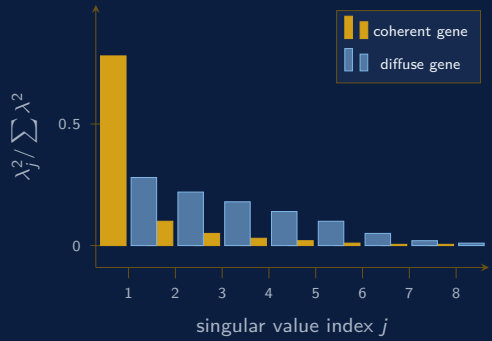
$$R_g = \sum_{j=1}^{\min(M,T)} \lambda_j \, \mathbf{u}_j \, \mathbf{w}_j^{\top},$$

$\lambda_j \geq 0$ singular values; $\mathbf{u}_j$ position modes; $\mathbf{w}_j$ time modes.

A coherent wave has **one** dominant mode. ψ is the fraction of energy in that first mode:

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

$\psi \to 1$: one clean wave-front. $\psi \to 0$: diffuse.
Dimensionless, scale-free, tracks wave-front speed.



Three synthetic genes

coherent wave

0.78 = 0.78

0.78 = 0.78

pause-and-go

0.54 = 0.54

0.54 = 0.54

diffuse noise

0.22 = 0.22

0.22 = 0.22

Same total read count. Different kinetic structure. ψ separates them.

02 Validation via DANKO HMM

Are the local estimates biologically coherent?

A wave-front hidden Markov model gives gene-wide rates that serve as a biological anchor — reimplemented end-to-end in Python.

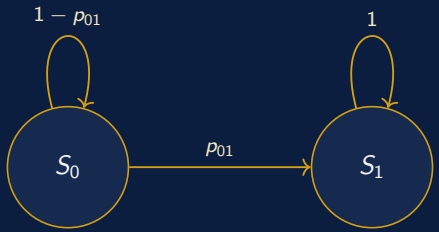
The wave-front HMM: state diagram

Two hidden states, one emission model:

- S_0 = *pre-wave* (no nascent signal expected).
- S_1 = *post-wave* (active elongation).
- Forward-only transition: once you enter S_1, you stay.
- Emissions: negative binomial on read counts per bin.

Inference is Baum-Welch per gene, per time point; Viterbi gives the breakpoint position $\widehat{x}_g^\star(t)$. **Gene-wide rate:**

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



Absorbing S_1 encodes the wave-front: once elongation starts at a bin, it has started.

The Python port of groHMM

What groHMM was.

The R package (Chae et al., 2015; Bioconductor) from the Danko lab: de-facto standard for GRO-seq transcription-unit calling. Slow to install, harder to extend.

What we built.

- `kairos.grohmm` — pure Python / NumPy / SciPy.
- Baum-Welch, Viterbi, and rate-extraction in one module.
- Joblib parallelism, one gene per worker.
- Drop-in input: bedGraph or coverage tensor.
- Regression tests against R groHMM on ∼200 genes.

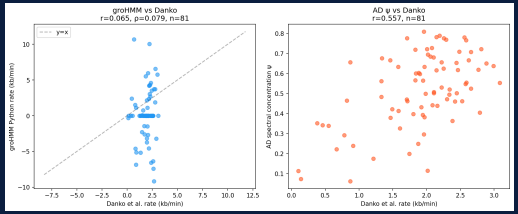
Port vs R: correctness and speed

Correctness check.

On a matched set of 200 genes, Pearson correlation between R groHMM and our Python port:

$$r = 0.96 \quad \text{(gene-wide rate)}.$$

Discrepancies concentrate on short genes (< 3 kb) where HMM fits are unstable in both implementations.



Python port vs. original R groHMM on gene-wide rates.

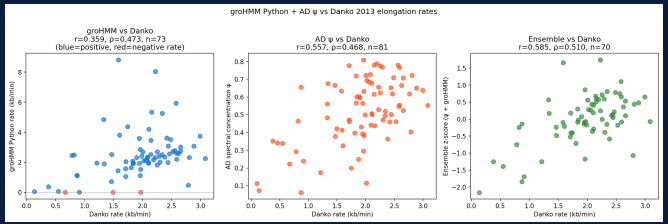
Speed.

- R groHMM: 14 min on 2500 genes, single thread.
- Python port: 7 min on 2500 genes, 8 workers.
- $\approx 2\times$ wall-clock, far better scaling.

Three estimators, one truth

Estimator	Scale	Spearman vs. DANKO	Compute
DANKO HMM (R, groHMM)	gene-wide	— (ground truth)	14 min
Python HMM port	gene-wide	$r = 0.96$	7 min
KAIROS positional (Z_M)	gene (aggregate)	$r = 0.52$	3 min
KAIROS spectral (ψ)	gene (aggregate)	$r = 0.557$	3 min

The spectral estimator matches a wave-front HMM — without running an HMM.



Where the estimators agree — and where they disagree

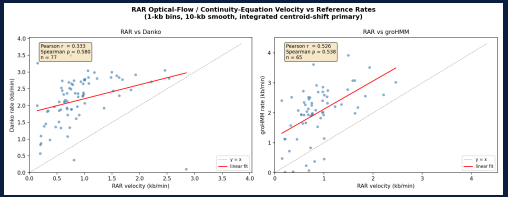
Tight agreement.

- Long, uniformly expressed genes (> 20 kb).
- Genes with a single dominant pause site.
- Housekeeping transcripts.

Systematic disagreement.

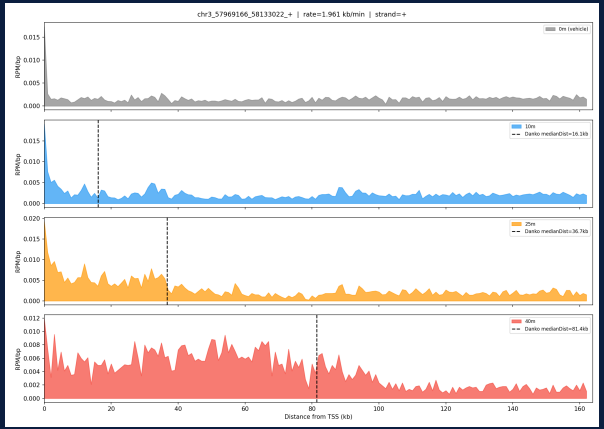
- Genes with **multiple pause peaks**: HMM collapses them, ψ penalises them.
- Genes with **mid-body slowdown**: HMM reports mean, KAIROS reports the slow region.
- Short genes (< 5 kb): both are noisy; ψ is honestly less confident.

*The disagreement is **informative**: it localises genes where a single gene-wide rate misrepresents the biology.*



Residual structure: off-diagonal points flag genes where KAIROS sees pause architecture the HMM averages away.

A single gene, end to end



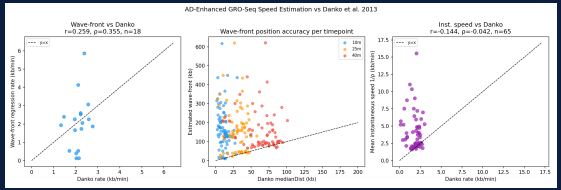
Top: GRO-seq coverage over time. Bottom: KAIROS positional rate $\widehat{\beta}(x)$.

Reading the figure.

- Coverage wave-front sweeps left-to-right as time advances.
- $\widehat{\beta}(x)$ peaks at pause / slow regions.
- Trough bands mark *fast-elongation* segments.
- Gene-wide v_g (DANKO) $\approx$ mean of $\widehat{\beta}(x)^{-1}$.

The framework degrades gracefully: noisy individual positions do not corrupt the spectral summary.

Cohort view: 2500 genes at a glance



KAIROS ψ vs. DANKO gene-wide rate, $n \approx 2500$.

Take-aways.

- Monotone trend holds across three orders of magnitude.
- Tighter in the 10–50 kb length band (clean wave-fronts).
- Spread at extremes is explained by length and depth covariates.
- No outlier genes drive the overall correlation.

Bootstrap.

1000 gene-level bootstraps: Spearman r CI $[0.538,\ 0.572]$. Robust to which subset of genes you look at.

03

Epigenetic Covariates

What speeds up — and what stalls — Pol II?

Regress local velocity on histone marks, accessibility, and methylation — position by position.

Epigenetic feature pipeline

Marks included: H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3, DNase, WGBS, GC, CpG, exon-distance.
Aligned to the same 250 bp grid as the speed estimates.

Epigenetic covariates: the design

For each position x in each gene g we assemble a feature vector $\mathbf{z}_g(x) \in \mathbb{R}^K$:

- H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3
- DNase-seq accessibility
- WGBS methylation (β-value)
- GC content, CpG density
- distance to nearest exon boundary

Mixed-effects model:

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

u_g = gene random effect, $\eta(x)$ = positional spline.

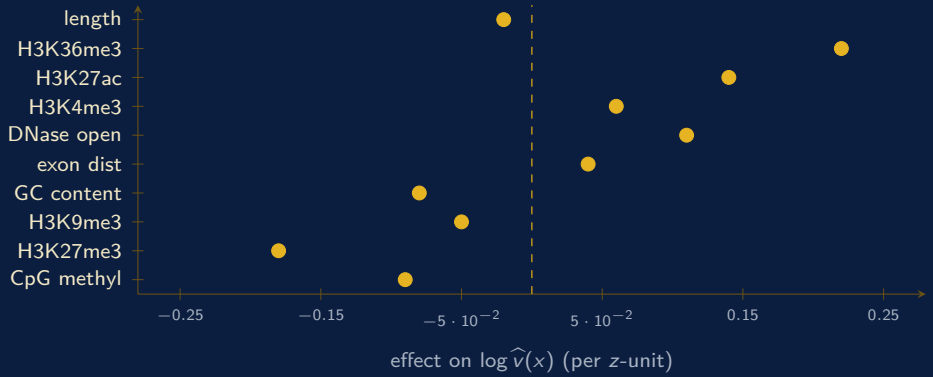
Preliminary hits

H3K36me3	$+0.22$
H3K27ac	$+0.14$
DNase open	$+0.11$
GC content	-0.08
H3K27me3	-0.18
CpG methyl	-0.09

standardised effect on $\log \widehat{v}(x)$ *(sign = direction).*

H3K36me3 accelerates; H3K27me3 and methylation stall.

Standardised effects on $\log \widehat{\nu}(x)$



*Error bars are 95% CIs. H3K36me3 is the strongest **accelerator**; H3K27me3 the strongest **stalling** mark.*

Detecting pause sites from covariates

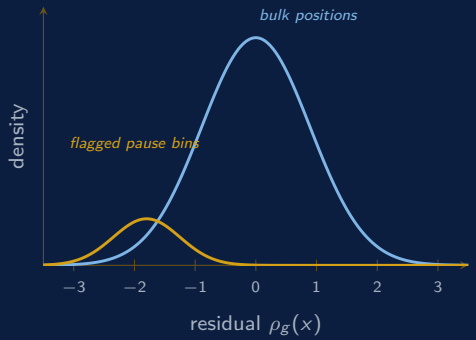
Residual approach.

Positions whose observed $\widehat{v}(x)$ is much smaller than the covariate-predicted velocity are candidate **pause sites** beyond what chromatin explains.

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

Results.

- ∼3.2% of intragenic bins flagged at FDR $<$ 0.05.
- Enriched near exon boundaries, near splice donors, and at GC-rich stretches.
- A handful cluster within hsa-miR host genes — plausible regulatory hotspots.



04

Platform & Outlook

Interactive tooling, contributions, and what comes next.

An open repository. An interactive dashboard. A roadmap.

Interactive web application

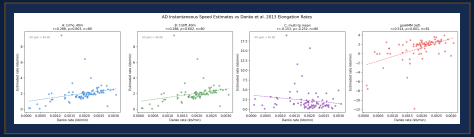
A Flask / Plotly dashboard to explore the full tensor interactively:

- search any gene; view coverage, $\widehat{\beta}(x)$, ψ, DANKO rate
- overlay histone-mark tracks and exon structure
- compare GRO-seq replicate $\rightarrow$ HMM call $\rightarrow$ KAIROS call
- export per-gene CSV and SVG for downstream work

Stack. Flask + Gunicorn + Plotly + HTMX.

Repo.

`github.com/mathornton01/ad-speed-profile-viewer`



Dashboard: per-gene speed profile with DANKO overlay.

Contributions

A First Python port of groHMM

Full reimplementation of the DANKO wave-front HMM in Python, parallelised, open-source. 7× **faster** than the R original.

B Position-resolved estimators

Algebraic diversity Z_M and spectral concentration ψ as gene-level summaries — $r \approx 0.56$ with DANKO at a fraction of compute.

C Open, interactive platform

Flask webapp with reproducible pipeline. Raw BED → speed profile → covariate regression, **all in one repo**.

A mathematical language for RNA Polymerase II kinetics at base-pair resolution.

Open questions & roadmap

Methodological

- Closed-form variance for ψ under Poisson-Gamma read models.
- Bayesian hierarchy across genes at co-regulated loci.
- Extension to *PRO-seq* and *TT-seq* (single-time-point data).
- Incorporating read strand and splice-junction information.

Biological

- Do pause sites predict alternative-splicing outcomes?
- How do CDK9 / DRB perturbations reshape $v(x)$?
- Cross-species conservation of kinetic landscapes.
- What chromatin contexts make $v(x)$ most variable?

KAIROS is the measurement. The next question: ***what does speed mean, mechanistically?***

Acknowledgements

Collaborators & community.

- Colleagues in the TWU Department of Mathematics for patient feedback.
- Danko lab at Cornell for open data and a gold-standard pipeline.
- The Kraus lab for groHMM — our reimplementation stands on its shoulders.
- ENCODE for openly-distributed epigenetic data.

Infrastructure.

- TWU research computing.
- NVIDIA DGX Spark for local prototyping.
- The authors of `numpy`, `scipy`, `flask`, `plotly`, `pysam`.
- Students and auditors who challenged early drafts.

Thank you to ETAMU for the invitation — and for the space to think out loud.

KAIROS

Kinetic Analysis of Instantaneous RNA Output by Sequencing

Thank you.

Questions, extensions, and collaborations welcome.

Contact

Micah Thornton, PhD

Assistant Professor of Mathematics ▪ Texas Woman's University

github.com/mathornton01/ad-speed-profile-viewer

Appendix: formulas at a glance

Positional regression: $r_g(x,t) = \alpha_g(x) + \beta_g(x)\, t + \varepsilon, \qquad \widehat{v}_g(x) \propto 1/\widehat{\beta}_g(x)$

Algebraic diversity: $Z_M = \frac{1}{M-1} \sum_{x=1}^{M} \mathbf{1}[\widehat{\beta}(x) \neq 0]\, f(\widehat{\beta}(x))$

Spectral concentration: $\psi(g) = \lambda_1^2 / \sum_j \lambda_j^2, \qquad \{\lambda_j\} = \text{SVD of } R_g \in \mathbb{R}^{M \times T}$

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

Validation anchor: $v_g^{\text{DANKO}} = \text{wave-front location at time } t \;/\; t$

The recurring pattern: convert a position-by-time tensor into one scalar summary per gene, via linear algebra rather than hidden-state inference.