Predicting single-cell transcription dynamics even when the central limit theorem fails.

New computational analyses predict spatiotemporal dynamics of single-cell transcription, using minimal assumptions and finite datasets.

Brian Munsky,[1,2,*] Guoliang Li,[3] Zachary Fox,[2]
Douglas P. Shepherd,[4] Gregor Neuert[4,*]

[1]Department of Chemical and Biological Engineering, Colorado State University
Fort Collins, CO 80523, USA
[2]School of Biomedical Engineering, Colorado State University,
Fort Collins, CO 80523, USA
[3]Department of Molecular Physiology and Biophysics, Department of Biomedical Engineering,
and Department of Pharmacology, Vanderbilt University, Nashville, TN, 37232, USA
[4]Department of Physics, Pediatric Heart Lung Center, and Department of Pediatrics,
University of Colorado Denver, Denver, CO 80217,

[*]Senior authors to whom correspondence should be addressed;
E-mail: munsky@colostate.edu; gregor.neuert@vanderbilt.edu

Mechanistic modeling is more predictive in engineering than in biology, but the reason for this discrepancy is poorly understood. The difference extends beyond randomness and complexity in biological systems. Statistical tools exist to disentangle such issues in other disciplines, but these assume normally distributed fluctuations or enormous datasets, which don't apply to the discrete, positive and non-symmetric distributions that characterize single-cell and single-molecule dynamics. Our approach captures discrete, non-normal

effects within finite datasets and enables biologically significant predictions. Using transcription regulation as an example, we discover quantitatively precise, reproducible, and predictive understanding of diverse transcription regulation mechanisms, including gene activation, polymerase initiation, elongation, mRNA accumulation, transport, and degradation. Our model-data integration approach extends to any discrete dynamic process with rare events and realistically limited data.

Introduction. The ultimate goal of modeling is to integrate quantitative data to understand, predict, or control complex processes. Useful models may be discovered through mechanistic or statistical approaches, but success is always limited by the quantity and quality of data and the rigor of comparison between models and experiments. These issues are largely solved in engineering, where computer analyses routinely enable the design of extraordinarily complex systems. Many would argue that predictive modeling in biology is far behind this capability due to limited experimental data, inescapable randomness or noise, and overwhelming biological complexity. These concerns have driven rapid single-cell experimental and computational advances, which have enabled measurement and modeling of individual biomolecules (i.e., DNA, RNA, and protein) in single cells with outstanding spatiotemporal resolution (*1–12*). Such experiments have allowed the characterization of many intriguing aspects of biological complexity and variation (*13*), while capturing these phenomena with stochastic gene regulation models has improved understanding of mechanisms and their parameters (*14–18*).

Despite experimental and computational advances, most biological models still underperform expectations. While it is tempting to attribute this failure to "poor models" or "insufficient data," a more subtle explanation is that combinations of sufficient data and good models may fail because they haven't been integrated properly. Many standard engineering techniques exist to integrate models with continuous-valued data, but unlike most engineered systems, biologi-

cal fluctuations are dominated by *discrete* events. A single molecule of DNA, RNA, or protein can change the fate of an organism (*19–22*). The resulting positive and discrete distributions violate the most basic assumption of most model inference approaches (i.e., that measurement errors are continuous Gaussian random variables). Moreover, this violation is compounded by the fact that datasets for single-cell imaging and sequencing are usually too small to invoke the central limit theorem (CLT). Consequently, standard data-model integration procedures can fail dramatically. We hypothesize that more exact treatment of discrete biological fluctuations could solve the data-model integration dilemma and enable precise quantitative predictions (Fig. 1).

To test this hypothesis, we aim to quantify and reduce model uncertainty and bias (Fig. 1A). We examine the evolutionary conserved Stress Activated Protein Kinase (p38 / Hog1 SAPK) signal transduction pathway (Figs. 1 and S5), and we quantify its control of transcription mechanisms including RNA polymerase transcription initiation and elongation on target genes as well as mature mRNA export and degradation in *Saccharomyces cerevisiae* during adaptation to hyper-osmotic shock (Fig. 1B) (*23*). We quantify the number of individual mRNA primary transcripts at the site of transcription, in the nucleus, and in the cytoplasm for multiple genes using single-molecule fluorescence *in situ* hybridization (smFISH) (Fig. 1C) (*1, 2*). We collect high-resolution data from more than 65,000 cells, and we quantify single-cell spatiotemporal mRNA distributions that are demonstrably non-normal and non-symmetric. For such distributions, huge data sets would be needed to justify application of the CLT (Fig. 1D). We use computational analyses to integrate these data with a discrete stochastic spatiotemporal model (Fig. 1E), and we show how different computational analyses of the *same experimental data* and *same models* can yield vastly different parameter biases and uncertainties (Fig. 1F,G).

We discover that standard single-cell modeling approaches, which assume continuous and normally distributed fluctuations *or* enough data to invoke the CLT (*24*), are not always valid to interpret finite datasets for single-cell transcription responses (Fig. 1D). These approaches can

yield surprising errors and poor predictions, especially when mRNA expression is very low. In contrast, we show that improved computational analyses of full single-cell RNA distributions can yield far more precisely constrained, less biased, and more reproducible models. We also discover new and valuable information contained in the intracellular spatial locations of RNA, enabling quantitative predictions for novel dynamics of gene regulation, including transcription initiation and elongation rates, fractions of actively transcribing cells, and the average number and distribution of polymerases per active transcription site, which could not otherwise be measured simultaneously in endogenous cell populations.

Results. Under osmotic stress, the high osmolarity glycerol kinase, Hog1, is phosphorylated and translocated to the nucleus, where it activates several hundred genes (*23*). For two of these genes (*STL1*, a glycerol proton symporter of the plasma membrane and *CTT1*, the Cytosolic catalase T), we quantified transcription at single-molecule and single-cell resolution (Figs. 1C, S6, and S7), at temporal resolutions of one to five minutes, at two osmotic stress conditions (0.2M and 0.4M NaCl), and in multiple biological replicas. We built histograms to quantify the marginal and joint distributions of the nuclear and cytoplasmic mRNA (Figs. 2D and S6-S9).

We extended a bursting gene expression model (*2, 14, 25*) to account for transcriptional regulation and spatial localization of mRNA (Fig. 1C,E) (*24*). We considered four approaches to fit this model to gene transcription data: First, we used exact analyses of the first moments (i.e., population means) of mRNA levels as functions of time. Second, we added exact analyses of the second moments (i.e.,variances and covariances). Third, we extended the moments analyses to include the third and fourth moments. Finally, we used the finite state projection (FSP, (*26*)) approach to compute the full joint probability distributions for nuclear and cytoplasmic mRNA. *All four approaches provide exact solutions of the same model*, but at different levels of statistical detail (*24*). We used each analysis to compute the likelihood that the measured mRNA data would match the model, and we maximized this likelihood (*24*). As was the case

for previous studies (*17, 27*), we note that the moments-based likelihood computations assume either normally distributed deviations (first and second methods) *or* sufficiently large sample sizes such that the moments can be captured by a normal distribution as guaranteed by the CLT (third method) (*24*). In contrast, the FSP approach (fourth method) makes no assumptions on the distribution shape and has no requirement for large sample sizes.

The four likelihood definitions were maximized by different parameter combinations (Tables S3 and S4), and the fit and prediction results are compared to the measured mean, variance, ON-fraction (i.e., fraction of cells with more than 3 mRNA / cell), and distributions versus time for *STL1* and *CTT1* (Figs. 2, S6, and S7). The different analyses used the same model, and they were fit to the exact same experimental data, but they yielded dramatically different results. When identified using the average mRNA dynamics, the model failed to match the variance, ON-fractions, or distributions of the process (Fig. 2B-D, left). Fitting the response means and variances simultaneously (Fig. 2A-B, center) failed to predict the ON-fractions or probability distributions (Fig. 2C-D, center). In contrast, parameter estimation using the full probability distribution (Fig. 2, right column and Figs. S6 and S7) matched all measured statistics. Importantly, the parameters identified using the FSP approach agree quite well with previous studies (*18*), which indicates strong reproducibility of both experiments and analyses (Tables S3 and S4) and provides more confident predictions for new transcriptional mechanisms as discussed below. In contrast, the moment-based analyses overestimated these rates by multiple orders of magnitude.

We considered three explanations for the failure of moment-based parameter estimation approaches: (i) the model parameters could be unidentifiable from the considered moments; (ii) the parameters could be too weakly constrained by those moments; or (iii) the moments analyses could have introduced systematic biases due to a failure of the CLT. To eliminate the first explanation, we computed the Fisher Information Matrix (FIM) defined by the moments-based analyses (*24*). Because the computed FIM has full rank, we conclude that the model should

be identifiable. If the second explanation were true (i.e., if the moments-analyses had produced weakly constrained models), then changing the parameters to those selected by the FSP analysis should have only a small effect on the moment-based likelihood. In such a case, the FSP parameters would lie within large parameter confidence intervals identified by the moments-based analyses (i.e., as depicted in Fig. 1F as opposed to Fig. 1G). However, using the experimental *STL1* data, we computed that the FSP parameter set was $10^{2,720}$ less likely to have been discovered using means, $10^{14,100}$ less likely to have been discovered using means and variances, and 10^{665} less likely to have been discovered using the extended moments analysis (Table S5). Thus, we conclude that failure of the moments-based analyses to match the distributions in Fig. 2 cannot be explained by uncertainty alone.

To test the third explanation for parameter estimation failure (i.e., systematic bias), we used the FSP parameters and generated simulated data for the mean (Fig. 3A), standard deviation (Fig. 3B), and the ON-fraction (Fig. 3C) versus time for *STL1* mRNA under an osmotic shock of 0.2M NaCl. As shown in Fig. 3A,B, the *median* of the simulated data sets (magenta) matches the experimental data (red and cyan) at all times, but at later times ($>$20 minutes) both are consistently less than the theoretical values (black). This mismatch is due to finite sampling from asymmetric distributions especially at later time points (Fig. 3D,E). The first two moments analyses, which do not account for this asymmetry, specify a tight and nearly-symmetric likelihood function (Fig. 3E, magenta lines), which is inconsistent with the broad and highly-asymmetric likelihood function computed using the FSP (Fig. 3E, blue lines). As a result, these likelihood functions deleteriously overfit the low mRNA expression at late time points, and resulted in an excessively confident overestimation of the mRNA degradation rate (Table S3). Conversely, the extended moments analysis, which allows for excessively large third and fourth moments, is too poorly constrained by the data (Fig. S10).

To confirm the tradeoff between uncertainty and bias, we applied the Metropolis Hastings

algorithm (MHA) to analyze parameter variation for the different likelihood functions and to estimate parameter uncertainty and bias (Fig. 4A-C) (*24*). Comparing the parameter variations for the transcription initiation rate, k_{i3}, and the mRNA degradation rate, γ, illustrates that extending the analysis from the means to means and variances can affect the parameter identification bias much more than the parameter uncertainty (Fig. 4A). Moreover, this effect is not always advantageous; inclusion of variances in the analysis led to substantially increased parameter bias (compare red and blues ellipses in Fig. 4A and see Figs. 4C and S13). In contrast, analyses using the FSP consistently reduced both uncertainty and bias for both *STL1* and *CTT1* analyses (Figs. 4A-C, S13, S14, and S15A-C).

Having established that different stochastic fluctuation analyses attain different levels of uncertainty and bias, we asked if more information could be extracted from spatially-resolved data. Using a nuclear stain, we quantified the numbers of *STL1* and *CTT1* mRNA in the nucleus and cytoplasm (*24*). We then extended the model and our analyses to consider the joint cytoplasmic and nuclear mRNA distributions (Fig. S8 and S9). From these analyses, we observed that spatial data reduced parameter bias for the models, despite the addition of new parameters and model complexity (Fig. 4A-C, S13-15).

We next explored how well the identified models could be used to predict the elongation dynamics of nascent mRNA at individual *STL1* or *CTT1* transcription sites (TS, Fig. 1C). We quantified the TS intensity for *CTT1*, and we used an extended FSP model for *CTT1* regulation to estimate the Polymerase II elongation rate to be 63 ± 14 nt/s (*24*), a value consistent with published rates of 14-61 nt/s (*28, 29*). We assumed an identical rate for the *STL1* gene, and we used the FSP model for *STL1* gene regulation to predict the *STL1* TS activity (Figs. 4D-H). The spatial (non-spatial) FSP model predicts an average of 6.8 (9.2) full length *STL1* mRNA per active TS, a value that matches well to our measured value of 4.2-7.5 *STL1* mRNA per active TS. However, predictions using parameters identified from moments-based analyses were

incorrect by several orders of magnitude (Fig. 4D). In addition to predicting the average number of nascent mRNA per active TS, the FSP model also accurately predicts the fraction of cells that have an active *STL1* TS versus time (Figs. 4E,F) as well as the distribution of nascent mRNA per TS (Figs. 4G,H).

Discussion. Integrating stochastic models and single-molecule and single-cell experiments can provide a wealth of information about gene regulatory dynamics (*14*). In previous work, we discussed the importance to *choose the right model* to match the single-cell fluctuation information and achieve predictive understanding (*18*). Here we have shown how important it is to *choose the right computational analysis* with which to analyze single-cell data. We showed how model identification based solely upon average behaviors can lead to substantial parameter uncertainty and bias, potentially resulting in poor predictive power (Figs. 4, S15). We showed how single-molecule experiments often yield discrete, asymmetric distributions that are demonstrably non-Gaussian (Figs. 1D, 3D, S6, and S7), and how model extensions to include hard-to-measure variances and covariances may exacerbate biases (Fig. 4C) leading to greatly diminished predictive power (Fig. 4D). We stress that this deleterious effect occurs even for models for which exact equations are known and solvable for the statistical moment dynamics. For more complex and nonlinear models, where approximate moment analyses are required, these effects are likely to be exacerbated further. This issue is expected to be even more relevant in mammalian systems, which exhibit greater bursting (*1, 2, 21*) and for which data collection may be limited to smaller sizes (e.g., by increased image processing difficulties for complex cell shapes or by small numbers of cells, as available from an organ, a tissue from a biopsy, or for a rare cell type population).

Because most single-cell modeling investigations to date have used only means or means and variances from finite data sets to constrain models, it is not surprising that many biological models fail to realize predictive capabilities. Conversely, full consideration of the single-

molecule distributions enabled discovery of a comprehensive model that quantitatively captures transcription regulation with biologically realistic rates and interpretation for transcription initiation, transcription elongation, and mRNA export and nuclear and cytoplasmic mRNA degradation (Fig. 4I). We argue that the solution is not to collect increasingly massive amounts of data, but instead to develop computational tools that utilize the full, unbiased spatiotemporal distributions of single-cell fluctuations. By addressing the limitations of current approaches and relaxing requirements for normal distributions or large sample sizes, our approach should have general implications to improve mechanistic model identification for any discipline that is confronted with non-symmetric datasets and finite sample sizes.

References and Notes

1. A. M. Femino, F. S. Fay, K. Fogarty, R. H. Singer, *Science* **280**, 585 (1998).

2. A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, S. Tyagi, *Nature Methods* **5**, 877 (2008).

3. S. C. Bendall, *et al.*, *Science* **332**, 687 (2011).

4. P. Hammar, *et al.*, *Nature genetics* **46**, 405 (2014).

5. J. T. Gaublomme, *et al.*, *Cell* **163**, 1400 (2015).

6. N. Battich, T. Stoeger, L. Pelkmans, *Cell* **163**, 1596 (2015).

7. J. D. Buenrostro, *et al.*, *Nature* **523**, 486 (2015).

8. L. A. Sepúlveda, H. Xu, J. Zhang, M. Wang, I. Golding, *Science* **351**, 1218 (2016).

9. T. Morisaki, *et al.*, *Science* **352**, 1425 (2016).

10. J. R. Moffitt, *et al.*, *Proceedings of the National Academy of Sciences* **113**, 11046 (2016).

11. L. Bintu, *et al.*, *Science* **351**, 720 (2016).

12. K. L. Frieda, *et al.*, *Nature* **541**, 107 (2017).

13. R. M. Kumar, *et al.*, *Nature* **516**, 56 (2014).

14. B. Munsky, G. Neuert, A. van Oudenaarden, *Science* **336**, 183 (2012).

15. C. Zechner, M. Unger, S. Pelet, M. Peter, H. Koeppl, *Nature Methods* **11**, 197 (2014).

16. A. Hilfinger, T. M. Norman, J. Paulsson, *Cell systems* **2**, 251 (2016).

17. J. Ruess, A. Milias-Argeitis, J. Lygeros, *Journal of The Royal Society Interface* **10**, 20130588 (2013).

18. G. Neuert, *et al.*, *Science* **339**, 584 (2013).

19. L. S. Weinberger, J. C. Burnett, J. E. Toettcher, A. P. Arkin, D. V. Schaffer, *Cell* **122**, 169 (2005).

20. G. M. Suel, *et al.*, *Science* **315**, 1716 (2007).

21. A. Raj, S. A. Rifkin, E. Andersen, A. van Oudenaarden, *Nature* **463**, 913 (2010).

22. G. Balázsi, A. van Oudenaarden, J. J. Collins, *Cell* **144**, 910 (2011).

23. H. Saito, F. Posas, *Genetics* **192**, 289 (2012).

24. See Materials and Methods.

25. J. Peccoud, B. Ycart, *Theoretical Population Biology* **48**, 222 (1995).

26. B. Munsky, M. Khammash, *The Journal of Chemical Physics* **124**, 044104 (2006).

27. M. Komorowski, M. J. Costa, D. A. Rand, M. P. H. Stumpf, *Proceedings of the National Academy of Sciences* **108**, 8645 (2011).

28. D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, R. H. Singer, *Science* **332**, 475 (2011).

29. P. B. Mason, K. Struhl, *Molecular Cell* **17**, 831 (2005).

Acknowledgments. BM and ZF designed and implemented the computational analyses. GL and GN designed and implemented the experimental analyses. BM, ZF, DS and GN wrote the manuscript. BM and ZF were funded by the W.M. Keck Foundation, DTRA FRCALL 12-3-2-0002 and CSU Startup Funds. GL and GN were funded by NIH DP2 GM11484901 and Vanderbilt Startup Funds. The authors like to thank Luis Aguilera, Anthony Weil, Alexander Thiemicke, Dustin Rogers, Benjamin Kesler and Rohit Venkat for comments on the manuscript.

Supplementary Materials

Materials and Methods

Figs. S5 to S16

Tables S1 to S6

References 31 to 48

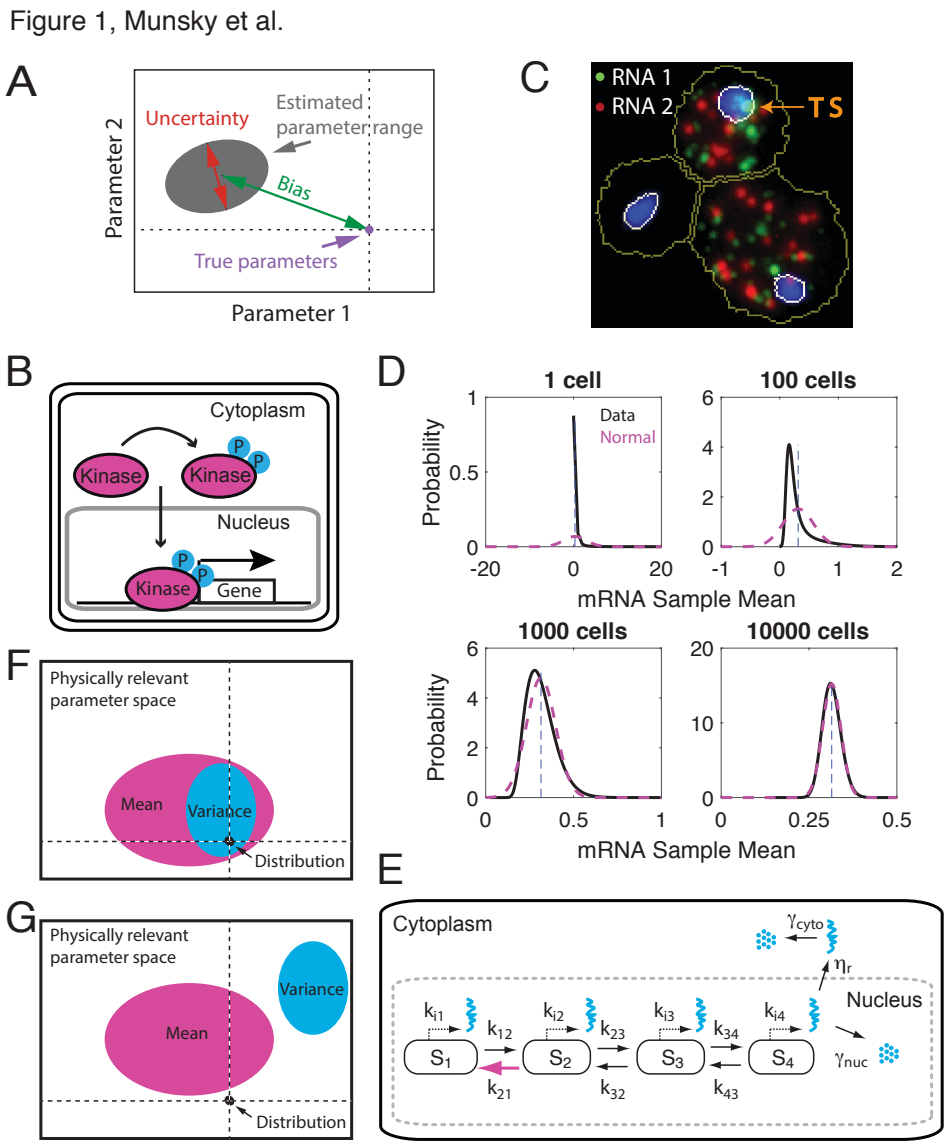


Figure 1: **Estimating stochastic models from single-cell data.** A) Model identification errors can be quantified in terms of uncertainty and bias. B) Simplified Hog1 kinase signal transduction and transcription pathway. C) Cytoplasmic, nuclear, and nascent transcription quantification for expression of two mRNA species (red and green). White line is the nuclear border and the dark yellow like is the cell boundary after automated segmentation. Intensely bright spots within the cell nucleus are identified as transcription sites (TS). D) Under the central limit theorem, the sample mean of a non-normal distribution converges to a normal distribution. For long tail distributions, this can take a very large number of cells ($\gg$ 1000 in this case). E) Stochastic model of different chromatin states (S_1-S_4 with Hog1-dependent k_{21} rate) (*18*) that predicts mRNA transcription initiation rates (k_{i1}-k_{i4}) and elongation, mRNA export (η_r), and nuclear and cytoplasmic mRNA degradation (γ_{nuc}, γ_{cyto}). F,G) Different analyses of the same data can provide very different effects on parameter uncertainty and bias.

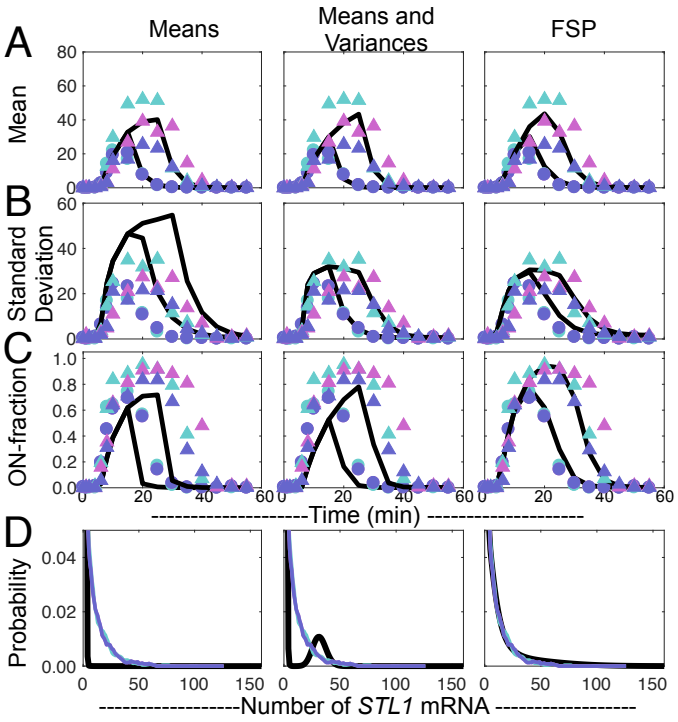


Figure 2: **Different computational analyses result in matches to different data characteristics**. A) Mean number, B) standard deviation, C) ON-fraction (cells with $\geq$ 3 mRNA), and D) distributions of *STL1* mRNA copy number. In each panel, data for 0.2M NaCl (circles, two biological replica) and 0.4M NaCl (triangles, three biological replica) are shown in magenta, cyan, and violet, and model results are shown in black. Distributions are measured at 0.2M NaCl after 20 minutes (cyan and violet are biological replica). The left column corresponds to the best fit to the measured mean; the center column corresponds to the best fit to the measured means and variances; and the right column corresponds to the best fit to the full measured distributions.

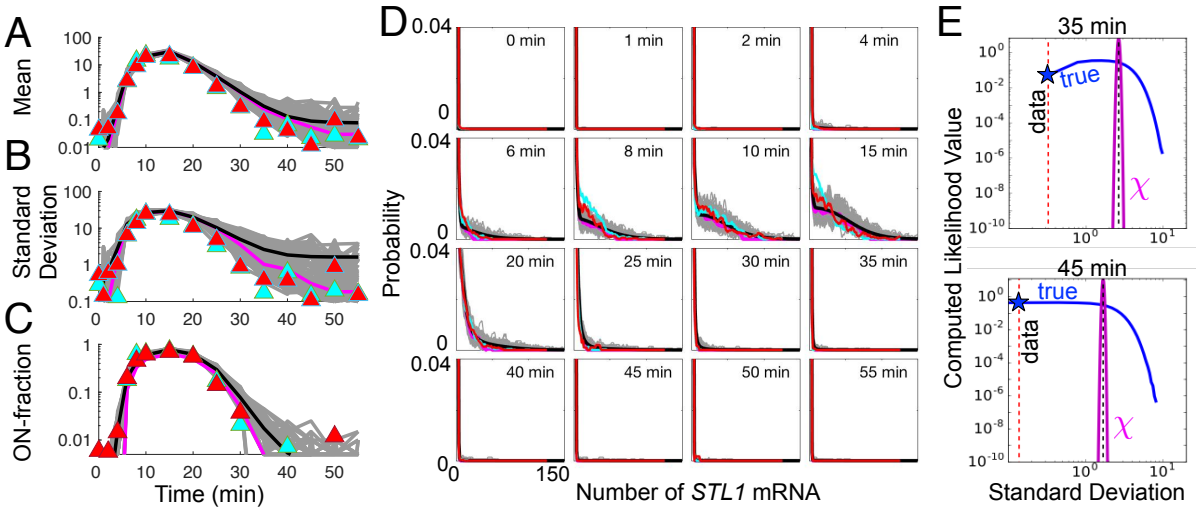
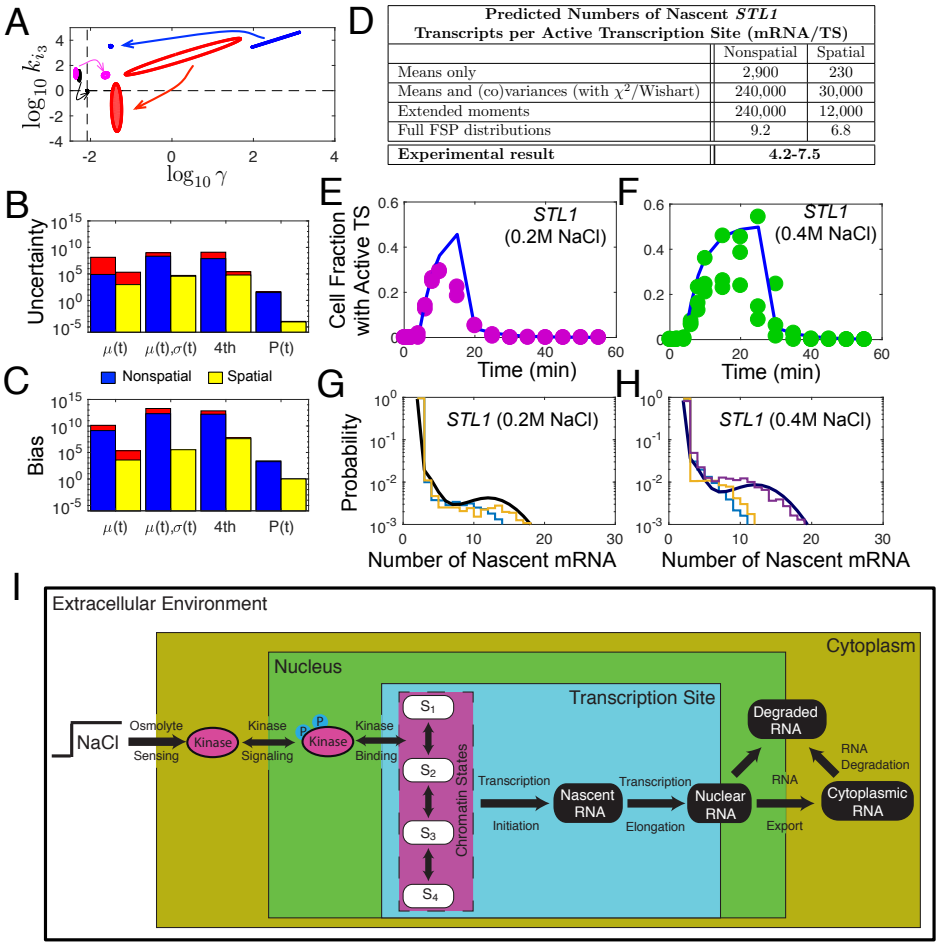


Figure 3: **Skewed measurements of summary statistics introduce parameter bias errors**. (A) Mean, (B) standard deviation, (C) ON-fraction, and (D) full distributions of *STL1* mRNA versus time for an osmotic shock of 0.2M NaCl applied at time $t = 0$. Theoretical values are in black, representative simulated samples of 200 cells each are in gray, median statistics of the simulated samples are in magenta; and experimental biological replica data are in red and cyan. (E) CLT-based approximation (magenta) and FSP-based computation (blue) of the likelihood of the standard deviation at 35 and 45 minutes (945 and 1348 cells, respectively) using the model identified by the FSP approach. The star denotes the true likelihood of the measured sample variance. In contrast the χ^2 approximation underestimates the likelihood by many hundreds of orders of magnitude.



Predicted Numbers of Nascent *STL1* Transcripts per Active Transcription Site (mRNA/TS)		
	Nonspatial	Spatial
Means only	2,900	230
Means and (co)variances (with χ^2/Wishart)	240,000	30,000
Extended moments	240,000	12,000
Full FSP distributions	9.2	6.8
Experimental result	**4.2-7.5**	

Figure 4: **Stochastic and spatial fluctuation information improve parameter estimation and yield greater predictive power**. (A) Ninety percent confidence ellipses for the degradation rate (γ) and the maximal transcription initiation rate (k_{i3}) using the means only ($\mu(t)$, red), means and variances ($\mu(t), \sigma(t)$, blue), extended moment analyses (4th, magenta), or the full FSP distributions ($\mathbf{P}(t)$, black). Arrows show the effect of adding spatial information to the analyses. The dashed black lines show the fit parameters for the spatial FSP *STL1* model. (B) Total parameter uncertainty and (C) bias for the four analyses using non-spatial (blue) and spatial (yellow) analyses. The red regions show the difference between independent MHA chains. (D) Predictions and experimental data for the average number of nascent mRNA per active *STL1* transcription site using each analysis. (E,F) Predicted (blue) and measured (magenta and green circles) for the fraction of cells with active *STL1* TS versus time at (E) 0.2M NaCl and (F) 0.4M NaCl osmotic shock. (G,H) Predicted (black) and measured (orange, blue, purple) distributions of nascent *STL1* mRNA per TS. (I) Summarized scope of the final model and experimental data, including quantitative analysis of MAPK induction and translocation, chromatin reorganization, polymerase initiation and elongation, and mRNA production, transport and degradation.