Distribution shapes govern the discovery of predictive models for gene regulation

Brian Munsky[a,b,1], Guoliang Li[c], Zachary R. Fox[b], Douglas P. Shepherd[d], and Gregor Neuert[c,e,f,1]

[a]Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523; [b]Keck Scholars, School of Biomedical Engineering, Colorado State University, Fort Collins, CO 80523; [c]Department of Molecular Physiology and Biophysics, School of Medicine, Vanderbilt University, Nashville, TN 37232; [d]Department of Pharmacology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045; [e]Department of Biomedical Engineering, School of Engineering, Vanderbilt University, Nashville, TN 37232; and [f]Department of Pharmacology, School of Medicine, Vanderbilt University, Nashville, TN 37232

Edited by Herbert Levine, Rice University, Houston, TX, and approved May 31, 2018 (received for review March 10, 2018)

Despite substantial experimental and computational efforts, mechanistic modeling remains more predictive in engineering than in systems biology. The reason for this discrepancy is not fully understood. One might argue that the randomness and complexity of biological systems are the main barriers to predictive understanding, but these issues are not unique to biology. Instead, we hypothesize that the specific shapes of rare single-molecule event distributions produce substantial yet overlooked challenges for biological models. We demonstrate why modern statistical tools to disentangle complexity and stochasticity, which assume normally distributed fluctuations or enormous datasets, do not apply to the discrete, positive, and nonsymmetric distributions that characterize mRNA fluctuations in single cells. As an example, we integrate single-molecule measurements and advanced computational analyses to explore mitogen-activated protein kinase induction of multiple stress response genes. Through systematic analyses of different metrics to compare the same model to the same data, we elucidate why standard modeling approaches yield nonpredictive models for single-cell gene regulation. We further explain how advanced tools recover precise, reproducible, and predictive understanding of transcription regulation mechanisms, including gene activation, polymerase initiation, elongation, mRNA accumulation, spatial transport, and decay.

single cell | transcription | quantitative | prediction | modeling

Systems biology seeks to integrate quantitative data with models to predict complex behaviors, such as how cells will react to environmental perturbations (1, 2), how mutations will affect cell phenotypes (3, 4), or how human diseases will respond to drug combinations (5). This goal comprises two steps: "Fitting" is choosing mechanisms and parameters to minimize differences between models and existing experimental data, and "prediction" is using previously fixed models to predict outcomes for untested conditions. Fitting models to data has become commonplace in systems biology, but unfortunately a good fit to one experiment does not guarantee good predictions for new biological conditions (6, 7). Many would argue that predictive modeling is prevented by inescapable biological complexity and the prevalence of randomness or noise (6), while others argue that predictive understanding could be achieved through quantification of model uncertainties (7).

The first argument focuses on the data and models individually and has driven rapid single-cell experimental and computational advances to measure and model individual biomolecules (i.e., DNA, RNA, and protein) in single cells with outstanding spatiotemporal resolution (8–19). Such experiments have characterized many intriguing aspects of biological complexity and variation (3), while capturing these phenomena with stochastic models has improved insight into gene regulation mechanisms and their parameters (1, 20–23). Despite these experimental and computational advances, most biological models still underperform expectations when used to predict new behaviors (6). By attributing such failures to "poor models" or "insufficient data," systems biology has traditionally sought to elucidate more detailed mechanisms or to collect higher-resolution data. However, success in predictive modeling may be limited not only by the quantity and quality of data and the appropriateness of the model but also by the rigor of comparison between models and measurements.

The second argument promotes more rigorous use of Bayesian data analyses to estimate uncertainty and quantify the value of a model given available data (7). Such approaches have attracted growing attention in biological investigations (7, 22), but model inference techniques that suffice in other fields may be inappropriate when applied to biological data. Specifically, most data–model integration techniques assume that measurement errors are continuous Gaussian random variables (24). For example, minimizing the logarithm of Gaussian errors is the theoretical basis for fitting a line to data by minimizing the sum-of-squared differences (least squares fit). For most engineered systems, this Gaussian assumption is justified by the Central Limit Theorem (CLT), which states that if one takes enough quantitative observations from the same underlying distribution, then the average of those observations would be normally distributed with a deviation given by the standard error of the mean (SEM) (25). Practical examples demonstrate that 20 observations are enough to invoke the CLT, but only if the underlying distribution is not extremely asymmetric (25). However, unlike most engineered systems, biological fluctuations are dominated by rare, discrete, and stochastic events (8–20, 23, 26), even to the extent that a single molecule of DNA,

Significance

Systems biology seeks to combine experiments with computation to predict biological behaviors. However, despite tremendous data and knowledge, biological models make less-accurate predictions compared with other fields. By analyzing single-cell, single-molecule measurements of mRNA during yeast stress response, we explore why and how the shapes of experimental distributions control prediction accuracy. We show how asymmetric data distributions with long tails cause standard modeling approaches to yield excellent fits but make meaningless predictions. We show how these biases arise from the violation of fundamental assumptions in standard modeling approaches. We demonstrate how advanced computational tools solve this dilemma and achieve predictive understanding of spatiotemporal mechanisms of transcription control including RNA polymerase initiation and elongation and mRNA accumulation, transport, and decay.

Author contributions: B.M. and G.N. designed research; B.M., G.L., Z.R.F., D.P.S., and G.N. performed research; B.M., Z.R.F., and G.N. contributed new reagents/analytic tools; B.M. and G.N. analyzed data; and B.M. and G.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: munsky@colostate.edu or gregor.neuert@vanderbilt.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804060115/-/DCSupplemental.

Published online June 29, 2018.

RNA, or protein can change the fate of an organism (4, 26–28). These single-molecule events can lead to positive and discrete distributions that are far from Gaussian (8–20, 23, 26), and satisfying the CLT for such highly nonsymmetrical data sets may require far more measurements than are standard practice for modern single-cell imaging or sequencing experiments.

The disconnect between single-cell data and standard model-inference techniques raises the possibility that combinations of sufficient data and good models may fail only because they have not been integrated with the right data–model comparison metrics. We hypothesize that more appropriate treatment of discrete biological fluctuations could solve the data–model integration dilemma, reduce uncertainty (i.e., the spread of parameters that match equally well to the data) and bias (i.e., the difference between the best-fit parameters and the true values), and achieve predictive modeling without the need to collect more data or generate new models. For example, in previous work, we demonstrated that the right analyses could systematically examine a large class of models with varying complexity and objectively select the best model to make quantitative predictions (23). Our present goal is not to identify a new model for a new biological system but rather to understand why specific single-cell analyses succeed in the exact same circumstances (i.e., same models, same conditions, and same data) under which standard (mean, variance, and higher moment) analyses yield excellent fits but meaningless predictions.

We seek to characterize and resolve the disconnect between good model fits and poor model predictions. Therefore, we adopt an existing model (Fig. 1*A*) already proven capable to predict precise aspects of *Saccharomyces cerevisiae* transcription in novel combinations of environmental and genetic conditions (23). We expand this model to include additional mRNA dynamics including transcript elongation and intracellular transport. We collect a very large set (>65,000 individual cells) of single-cell and single-molecule data for a different yeast cell line, and we fit the model to these data using many different analytical approaches. All approaches produce excellent fits (Fig. 2), yet standard (mean, variance, and higher moments) data-fitting techniques yield predictions that are wrong by many orders of magnitude (Fig. 1*D*). To explain these errors, we quantify model estimation errors in terms of parameter uncertainty and bias. We then demonstrate why standard single-cell modeling approaches, which assume continuous and normally distributed fluctuations or enough data to invoke the CLT (25) (*Methods* and *SI Appendix*), lead to nonintuitive biases and poor predictions (Fig. 1*D*), especially when mRNA expression is very low. In contrast, we show that improved computational analyses of full single-cell RNA distributions, which do not rely on the CLT, can yield far more precisely constrained, less-biased, more reproducible, and more predictive models (Fig. 1*D*). We also discover important information contained in the intracellular spatial locations of RNA (Fig. 1*B* and *SI Appendix*, Figs. S4 and S5), enabling quantitative predictions for dynamics of gene regulation at multiple scales, including transcription initiation and elongation rates, fractions of actively transcribing cells, and the average number and distribution of polymerases per active transcription site (TS) versus time (Figs. 1–4), which have not been, and could not otherwise be, measured simultaneously in endogenous cell populations.

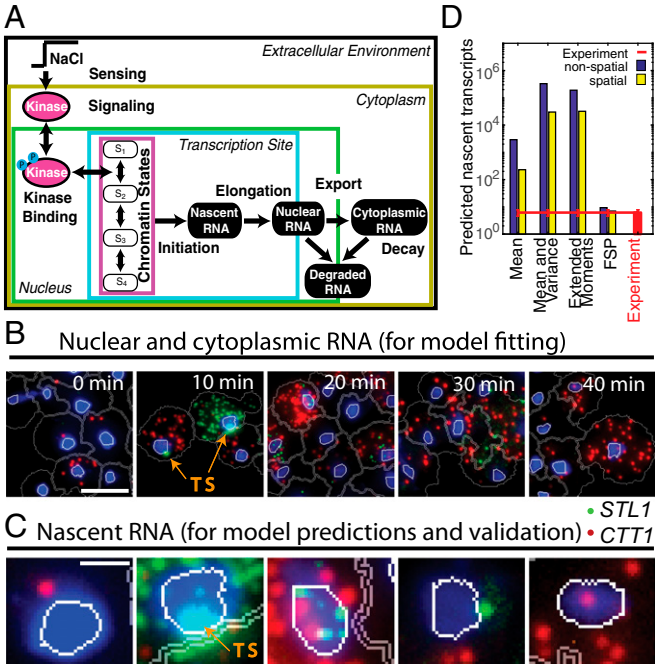

Fig. 1. Discovering stochastic models to predict single-cell gene regulation. (*A*) Scope of the model, including quantitative analysis of MAPK induction and translocation, chromatin reorganization, polymerase initiation and elongation, and mRNA transcription, export, and nuclear/cytoplasmic decay. RNAs in the cytoplasm (yellow) and nucleus (green) are used to constrain parameters. RNAs are predicted at the TS (cyan). Parameterization of the kinase signaling dynamics and the number of chromatin states were previously identified (23) (purple). (*B*) Collection of single-cell spatiotemporal RNA transcription data to fit the model. Cytoplasmic and nuclear transcription quantification for expression of two mRNA species (*CTT1* in red and *STL1* in green). DAPI-stained nucleus in blue. The white line is the nuclear border, and the gray line is the cell boundary after automated segmentation. Representative images of cells exposed to 0.2 M NaCl; 65,454 cells in total have been imaged at 16 time points. (Scale bar: 5 μm.) (*C*) Intensely bright spots within some cell nuclei are identified as TS. TS data are used to determine the number of nascent transcripts and validate model predictions. (Scale bar: 1 μm.) (*D*) Model validation comparing measured (red, experiment) to predicted average number of nascent *STL1* RNA per TS using the same model and same training data but under different modeling assumptions. Non-spatial analyses (blue) use the statistics (means, means and variances, or distributions) of the total number of RNA per cell. Spatial analyses (yellow) use the joint statistics of nuclear and cytoplasmic number of RNA per cell.

Results

To elucidate the importance of the data–model integration approach, rather than just the data or model alone, we analyzed single-cell transcription activation under the control of hyperosmotic stress in *S. cerevisiae* (Fig. 1*A* and *SI Appendix*, Fig. S1). Specifically, we analyzed the high-osmolarity glycerol kinase Hog1, which is a well-characterized homolog of the human p38 kinase that helps regulate differentiation and apoptosis. Under osmotic stress, Hog1 is phosphorylated and translocated to the nucleus, where it activates several hundred genes (29). We used a fluorescent protein reporter and time-lapse fluorescence microscopy to quantify Hog1p dynamics at 1-min resolution throughout the stress-adaptation response (*SI Appendix*, Fig. S1).

We then quantified transcription activity for two Hog1p-activated genes: *STL1*, a glycerol proton symporter of the plasma membrane, and *CTT1*, the cytosolic catalase T. For both genes, we used single-molecule RNA FISH (8, 9), along with a nuclear stain, and custom image processing software to quantify simultaneously the number of individual mRNA primary transcripts at the site of transcription, in the nucleus, and in the cytoplasm, all at temporal resolutions of 1 to 5 min, at two osmotic stress conditions (0.2 M and 0.4 M NaCl), in multiple biological replicas, and for more than 65,000 cells (Fig. 1 *B* and *C* and *SI Appendix*, Figs. S2 and S3). With these datasets of unprecedented spatial and temporal detail, we built histograms to quantify the marginal and joint distributions of the nuclear and cytoplasmic mRNA (Fig. 2 *D* and *E* and *SI Appendix*, Figs. S2–S5). The resulting distributions are demonstrably nonnormal and nonsymmetric (*SI Appendix*, Figs. S2 and S3).

In previous work, we searched hundreds of different model topologies to find and validate a simple model that consists of four states in a linear chain and which captured and quantitatively predicted Hog1-activated gene expression for several yeast genes including *CTT1* and *STL1* and in multiple genetic and environmental

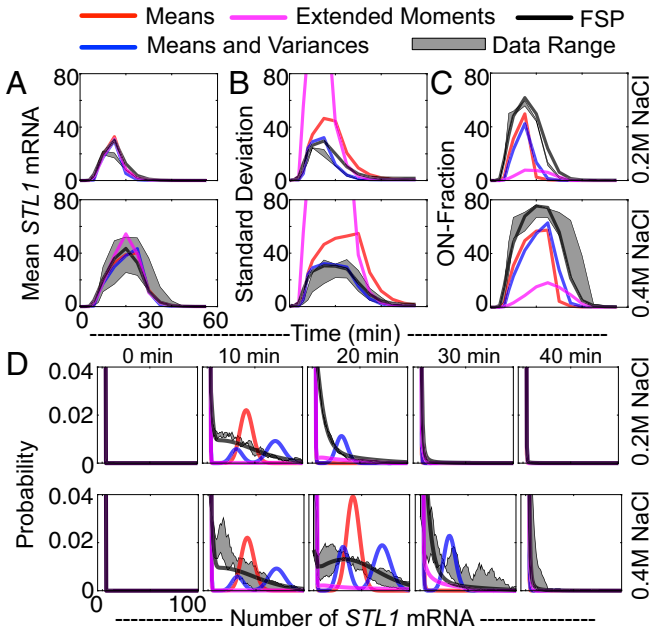

Fig. 2. Different computational analyses result in matches to different data statistics. (*A*) Mean number, (*B*) SD, (*C*) ON-fraction (cells with more than three mRNAs), and (*D*) temporal distributions of *STL1* mRNA copy number. In each panel, data for 0.2 M NaCl (two biological replicas) and 0.4 M NaCl (three biological replicas) are shown in top and bottom rows, respectively. Data range is shown in gray. Colors denote models identified using analyses of mean (red), means and variances (blue), extended moments (magenta), and full distributions (black). Extended results for spatial fits and *CTT1* expression fits are shown in *SI Appendix*, Figs. S2–S5.

conditions (23). Our current study considers these specific genes and conditions so that we can now explore the robustness and reproducibility of model parameter estimation even when applied to different cell lines and utilizing different sets of laboratory and microscopy equipment. To test the ability of our approach to capture and predict transcription regulation mechanisms on different scales, we also extended our previous model to consider transport of mRNA from nucleus to cytoplasm, nuclear and cytoplasmic mRNA decay, as well as mRNA elongation dynamics (Fig. 1*A*) (*Methods* and *SI Appendix*).

Our primary goal is to quantify and explain the intricate effects that different statistical data analyses have on the uncertainty, bias, and resulting predictive capabilities of gene regulation models. Toward this end, we considered four approaches to fit the extended model to the measured gene transcription data (Fig. 2, full details are given in *Methods* and *SI Appendix*). First, we used exact analyses of the first moments (i.e., population means) of mRNA levels as functions of time. This is the standard approach to fit dynamical models to time-varying data (24). 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; ref. 30) approach to compute the full joint probability distributions for nuclear and cytoplasmic mRNA. All four approaches provide exact solutions of the same model as functions of time during the adaptation response, but with different levels of statistical detail (*Methods* and *SI Appendix*). We used each analysis to compute the likelihood that the measured mRNA data would match the model, and we maximized these analysis-dependent likelihood functions (*Methods* and *SI Appendix*). As was the case for previous studies (22, 31), 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 first two moments could be captured by a multivariate normal distribution as guaranteed by the CLT (third method) (ref. 22 and *Methods* and *SI Appendix*). In contrast, the FSP approach (fourth method) makes no assumptions on the distribution shape and has no requirement for large sample sizes.

Different Exact Analyses of the Same Model and Same Data Yield Dramatically Different Results. All four approaches produced excellent fits to the corresponding features in the experimental training data (Fig. 2). However, the four likelihood functions were maximized by different parameter combinations (*SI Appendix*, Tables S3 and S4), and the resulting models were compared with the measured mean, variance, ON-fraction (i.e., fraction of cells with more than three mRNAs per cell), and distributions versus time for *STL1* and *CTT1* (Fig. 2 and *SI Appendix*, Figs. S2 and S3). When identified using the average mRNA dynamics (Fig. 2*A*, red), the model failed to match the variance, ON-fractions, or distributions of the process (Fig. 2 *B*–*D*, red). Fitting the response means and variances simultaneously (Fig. 2 *A* and *B*, blue) failed to predict the ON-fractions or probability distributions (Fig. 2 *C* and *D*, blue). Extending the moments-based likelihood analysis to include third and fourth moments led to very poor fits to the variances (Fig. 2*B*, magenta) and provided no improvement to the distribution predictions (Fig. 2*D* and *SI Appendix*, Figs. S2 and S3). In contrast, parameter estimation using the full probability distributions (Fig. 2, black and *SI Appendix*, Figs. S2 and S3) matched all measured statistics. Importantly, key conserved parameters identified using the FSP approach agree well with previous studies (23). For example, decay rate estimates for *CTT1* (0.0053 s^{-1}) and *STL1* (0.0021 s^{-1}) changed only 5% and 8% compared with our previously reported values (*SI Appendix*, Tables S3 and S4). This agreement, which indicates strong reproducibility of both experiments and analyses, provides more confident predictions for new transcriptional mechanisms as discussed below. In contrast, the moment-based analyses led to far less consistent results, in many cases overestimating these rates by multiple orders of magnitude (*SI Appendix*, Tables S3 and S4).

Standard Modeling Identification Procedures Fail due to Bias in Moment Estimation. We considered three explanations for why standard moment-based parameter estimation approaches failed: (*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 systematically evaluate these three explanations, we quantified the posterior uncertainty and bias in parameters after fitting to single-cell data under each modeling approach and for different aspects of quantified single-cell data (Figs. 3 and 4 and *SI Appendix*, Figs. S9–S11). To eliminate the first explanation, we computed the Fisher information matrix (FIM) defined by the moments-based analyses (*Methods* and *SI Appendix*). 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 the FSP parameters would lie within large parameter confidence intervals identified by the moments-based analyses. However, using the experimental *STL1* data, we computed that the FSP parameter set was $10^{2,750}$ less likely to have been discovered using means and $10^{14,500}$ less likely to have been discovered using means and variances (*SI Appendix*, Table S5). In other words, the means-based analysis resulted in the worst-case scenario of a high confidence estimate of poor parameters, and inclusion of variances in the analyses only exacerbated the issue. Thus, we conclude that failure of the moments-based analyses to match the distributions in Fig. 2 and *SI Appendix*, Figs. S2 and S3 cannot be explained by model 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. 3*A*), SD (Fig. 3*B*), ON-fraction (Fig. 3*C*), and distributions (Fig. 3*D*) versus time for *STL1* mRNA under an osmotic shock of 0.2 M NaCl and for the other combinations of genes and conditions (*SI Appendix*, Fig. S7). Each panel shows the exact theoretical prediction (black), 20 sets

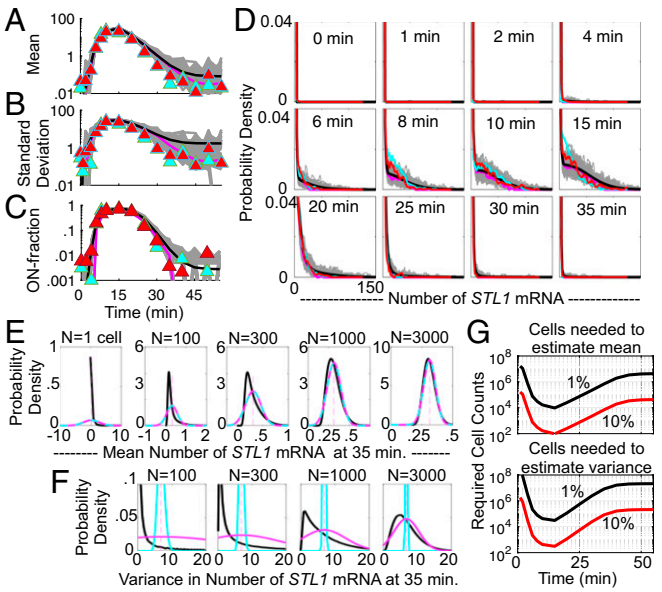

Fig. 3. Violation of the CLT due to discrete positive distributions leads to failure of moment estimation. (*A*) Mean, (*B*) SD, (*C*) ON-fraction, and (*D*) full distributions of *STL1* mRNA versus time for an osmotic shock of 0.2 M NaCl applied at time $t = 0$. Theoretical values are in black, representative 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* and *F*) Expected distribution of sample mean (*E*) and sample variance (*F*) for *STL1* at 35 min computed using a Gaussian approximation (cyan), an extended moment analysis with exact knowledge of the third and fourth moments (magenta), or exact sampling from the FSP (black) for population sizes of 1, 100, 300, 1,000, and 3,000 cells. (*G*) Expected number of cells required to estimate the mean (*Top*) or variance (*Bottom*) within SEs of 10% (red) or 1% (black) for *STL1*. The dependence on time is due to the changing distribution shapes shown in *D*.

of independent 200-cell population simulations (gray), the overall median statistic from these simulations (magenta), and experimental results for biological replicas (red and cyan). As shown in Fig. 3 *A* and *B*, the median of the simulated datasets (magenta) matches the experimental data (red and cyan) at all times, but at later times (>20 min) both are consistently less than the theoretical value (black). This mismatch is due to finite sampling from highly asymmetric distributions, particularly at later time points (Fig. 3*D*). The Gaussian assumption applied to the first two moments analyses (*Methods* and *SI Appendix*), which does not account for asymmetry, imposes narrow and nearly symmetric likelihood functions for the sample mean and sample variance (cyan lines in Fig. 3 *E* and *F*, respectively). These moment-based likelihood functions are inconsistent with the actual sample statistic distributions (Fig. 3*F*, compare cyan and black lines). Because the mRNA distributions are very broad at late time points (Fig. 3*D*), one would need to measure 10^5 or 10^7 cells to estimate the variance within 10% or 1%, respectively (Fig. 3*G*). Furthermore, because the mRNA distributions are asymmetric, measurements are likely to repeatedly underestimate the mean summary statistics (Fig. 3 *A* and *B*; compare magenta lines or red/cyan triangles to the black line). Moreover, when the moment-based likelihood functions were constrained to match underestimated mRNA expression at late time points, the analyses resulted in excessively confident overestimation of the mRNA decay rate (*SI Appendix*, Table S3). In principle, if exact third and fourth moments were known a priori, then the extended moments analysis would have been able to capture the correct likelihood function for the sample statistics. However, in practice, higher moments are even more difficult to measure, and all moments had to be computed by the same model. Thus, the extended moment analyses led to much greater uncertainty (*SI Appendix*, Fig. S6).

Full Distribution Analyses Substantially Reduce Model Uncertainty and Bias. To confirm the trade-off 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. 4 *A*–*C*, *Methods*, and *SI Appendix*, Figs. S9–S11). Comparing the parameter variations for the transcription initiation rate, k_{i3}, and the mRNA decay rate, γ, illustrates that extending the analysis from the means to means and variances affected the parameter identification bias much more than the parameter uncertainty (Fig. 4*A*). Moreover, this effect could be deleterious; analysis of variances led to substantially increased parameter bias for *STL1* (compare red and blues ellipses in Fig. 4 *A* and *C* and see *SI Appendix*, Fig. S9) and relatively little change for *CTT1* (*SI Appendix*, Figs. S10 and S11). Although extension to third and fourth moments improved estimation of k_{i3} and γ (Fig. 4*A*), the higher moments led to an increase in overall uncertainty (Fig. 4*B*). In contrast, analyses using the FSP consistently reduced both uncertainty and bias for both *STL1* and *CTT1* analyses (Fig. 4 *A*–*C* and *SI Appendix*, Figs. S9–S11).

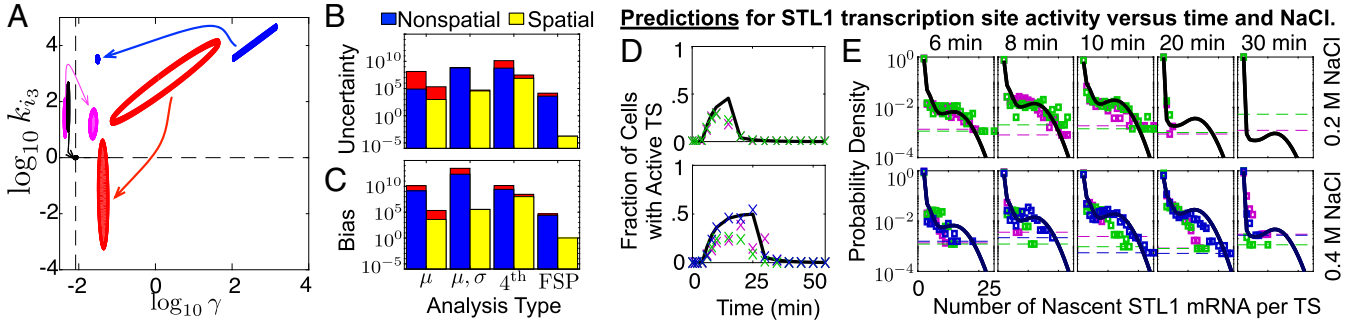

Fig. 4. Stochastic and spatial fluctuation information reduce uncertainty and bias in parameter estimation to enable precise quantitative predictions. (*A*) Ninety percent confidence ellipses for the decay rate (γ) and the maximal transcription initiation rate (k_{i3}) using the means only [μ(t), red], means and variances [μ(t),Σ(t) blue], extended moment analyses (fourth, magenta), or the full FSP distributions [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 nonspatial (blue) and spatial (yellow) analyses. The red regions show the difference between independent MHA chains. (*D*) FSP predicted (black) and measured (magenta, green, and blue crosses) fractions of cells with active *STL1* TS versus time at 0.2 M (*Top*) NaCl and 0.4 M (*Bottom*) NaCl osmotic shock. (*E*) FSP predicted (black) and measured (magenta, green and blue) distributions of nascent *STL1* mRNA per TS at different times following 0.2 M (*Top*) NaCl and 0.4 M (*Bottom*) NaCl osmotic shock. Magenta, green, and blue horizontal lines correspond to the minimum detection limit ($1/N_c$, where N_c is the number of cells measured at that time for the corresponding biological replica). All predictions are made using fixed parameters estimated previously from the mature, spatial mRNA distributions.

Using Spatial Fluctuations Improves Model Identification. 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. We then extended the model and our analyses to consider the joint cytoplasmic and nuclear mRNA distributions (*SI Appendix*, Figs. S4 and S5). From these analyses, we observed that spatial data reduced parameter bias and uncertainty for the models, despite the addition of new parameters and model complexity (Fig. 4 *A*–*C* and *SI Appendix*, Figs. S9–11).

Measuring and Predicting Transcription Site Dynamics. We next explored how well the identified models could be used to predict the elongation dynamics of nascent mRNA at individual *STL1* or *CTT1* TS (Fig. 1 *B* and *C*). 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 ± 13 nt/s (*Methods* and *SI Appendix*), a value consistent with published rates of 14–61 nt/s (32, 33). 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. 1*D* and 4*D*). The spatial (nonspatial) FSP model predicts an average of 7.0 (9.3) 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. 1*D*). In addition to predicting the average number of nascent mRNAs per active TS, the FSP model also accurately predicts the fraction of cells that have an active *STL1* TS versus time (Fig. 4*D*) as well as the distribution of nascent mRNA per TS versus time (Fig. 4*E*).

Discussion

Integrating stochastic models and single-molecule and single-cell experiments can provide valuable information about gene regulatory dynamics (20). We previously discussed the importance of choosing the right model to match the single-cell fluctuation information and achieve predictive understanding (23). Here we showed why and how important it is to choose the right computational analysis with which to analyze single-cell data. We showed that model identification based solely upon average behaviors can lead to substantial parameter uncertainty and bias, potentially resulting in poor predictive power (Figs. 1–4). We showed how single-molecule experiments often yield discrete, asymmetric distributions that are demonstrably non-Gaussian (Figs. 2 *D* and *E* and 3 and *SI Appendix*, Figs. S2 and S3), and how model extensions to include hard-to-measure variances and covariances may exacerbate biases (Fig. 4*C*), leading to greatly diminished predictive power (Fig. 1*D*). By taking into account the full distribution shapes, one can correct these deleterious effects and obtain parameter estimates and predictions that are improved by orders of magnitude, even when applied to the same model and same data (Figs. 1*D*, 2, and 4). We stress that this concern occurs even for models for which exact equations are known and solvable for the statistical moment dynamics. For more complex and nonlinear systems or for models where cellular communication or selective growth induce non-Markovian dynamics (34), approximate analyses are required, and these effects are likely to be exacerbated further. These issues are expected to be even more relevant in mammalian systems, which exhibit greater bursting (8, 9, 28) 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).

Most biological modeling investigations to date have used only means or means and variances from finite datasets to constrain models, so it is not surprising that many models fail to realize predictive capabilities. Conversely, our 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 decay (Fig. 1*A*). 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, such approaches should have general implications to improve mechanistic model identification for any discipline that is confronted with nonsymmetric datasets and finite sample sizes.

Methods

Yeast Strain, Growth Condition, and Sample Preparation. *S. cerevisiae BY4741* (*MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*) was used for FISH experiments, and Hog1 was tagged on the C terminus with YFP for live-cell imaging (2, 35). Cells were grown in a flow chamber or in a culture flask in the presence of minimal media (CSM) with or without 0.2 M or 0.4 M NaCl. For RNA-FISH, cells were fixed between 0–55 min after osmotic stress in time intervals of 1, 2, and 5 min and spheroplasted, RNA-FISH probes were hybridized, and cells were imaged.

Microscopy Setup, Image Acquisition, and Image Analysis for Time-Lapse and Single-Molecule RNA-FISH Imaging. Cells were imaged with a Nikon Ti-E epifluorescent microscope. Live-cell time-lapse microscopy was performed in flow chambers by taking bright-field and YFP fluorescent images. These images are used to track cells over time and to segment cells automatically. The final time-lapse microscopy dataset consists of 246 (0.2 M NaCl) and 167 (0.4 M NaCl) cells containing biological duplicates or triplicates. Single-molecule RNA-FISH microscopy was done in z-stacks of fixed yeast cells, for which the nucleus and cell boundary was segmented and the fluorescent RNA spots were counted automatically. The total RNA-FISH dataset consists of a total of 65,454 single cells (25,511 at 0.2 M NaCl and 39,943 at 0.4 M NaCl) with cells expressing *STL1* and *CTT1* mRNA. From these datasets the marginal distributions, the joint probability distributions of nuclear and cytoplasmic RNA, and the fraction of cells with more than three mRNA molecules (ON-cells) and the number of nascent transcripts were determined.

Hog1-Kinase Model. Parameters of an existing model (23) were fit to the measured Hog1p nuclear enrichment levels as functions of time and osmolyte concentrations (*SI Appendix*, Fig. S1 and Table S2). This time-varying signal was used as an input to the gene regulation models.

Gene Regulation Model. To capture the spatial stochastic expression of *STL1* or *CTT1* mRNA, an existing four-state *Hog1p*-activated gene expression model (23) was extended to account for spatial localization of mRNA in the nucleus or cytoplasm (Fig. 1*A*). In total, there are 13 nonspatial or 15 spatial parameters in the model.

Computation of Moments. Moment dynamics were analyzed using sets of coupled linear time-varying ordinary differential equations, which provide exact expressions for the dynamics of the model's means, variances, covariances, and higher moments (36). All reaction rates are all linear, these moment equations are closed, and the moments can be computed exactly.

Computation of Full Distributions. Distributions were computed using the FSP approach (30) to solve the chemical master equation. The FSP is a finite set of linear, time-varying ordinary differential equations, whose solutions provide guaranteed bounds on the model-predicted probability distributions at all finite times.

Computation of Moment-Based Likelihood Functions. Three approaches were derived to compute the likelihood of observed moments. First, to estimate likelihoods of the average data, fluctuations were assumed to be Gaussian, with the model-generated means and the measured sample (co)variances. Second, to estimate likelihoods of measured sample variance (nonspatial) or covariance matrix (spatial), the χ^2 distribution (nonspatial) or Wishart distribution (spatial) was used to approximate the likelihood of the measured sample means and variances, given the model (37). Third, the CLT and the first four model moments were used to approximate the joint likelihood for joint sample means and sample covariance matrix (22).

Computation of the Full Distribution Likelihood. The log-likelihood of the full distribution data was computed using the FSP approach (23, 38), using the formula $\log(L) = \sum d_i P_i$, where d_i and P_i are the measured number and FSP-generated probability of cells with i mRNA at the appropriate time.

Parameter Searches to Maximize Likelihood. Local and global parameter searches to maximize likelihood functions used multiple starting parameter guesses and totals of $>4 \times 10^7$ evaluations of the means and moments analyses, $>5 \times 10^6$ evaluations of the nonspatial FSP distributions, and $>5 \times 10^5$ evaluations for the extended moments and the spatial FSP distributions.

Quantification of Parameter Uncertainties. The MHA (39) was used to quantify parameter uncertainties. All parameter explorations were conducted in logarithmic parameter space, and all analyses (with moments or distributions, both spatial and nonspatial) used the same proposal distributions as described in *SI Appendix*. MHA chain lengths were >1 million (means and simpler moments), >120,000 (extended moment analyses), >250,000 (nonspatial FSP), and >15,000 (spatial FSP). *SI Appendix*, Fig. S8 shows the similarity of distributions of likelihood values for two independent MHA runs for each gene and analysis. The total bias and total uncertainty (Fig. 4 *B* and *C* and *SI Appendix*, Fig. S11 *B* and *C*) were computed as described in *SI Appendix*. Cross-validation was applied to verify that the most important parameters identified using the FSP were insensitive to specific biological replica data and that the MHA results applied to all data were consistent with parameter variations under biological replica studies (*SI Appendix*, Fig. S12).

Predictions of TS Activity. Two analyses were developed to predict TS activity: a simplified theoretical analysis of average active TS activity and an extended FSP analysis of distributions of polymerases on a given TS. In the simplified analysis, it was assumed than an active TS would correspond to one gene at steady state with the maximum transcription rate, $k_{\text{i-max}}$. Under this assumption, the average number of elongating polymerases is given by $\langle n_{\text{pol}} \rangle = k_{\text{i-max}}\tau_{\text{elong}} = k_{\text{i-max}}L/k_{\text{elong}}$. The average nascent mRNA was assumed to be half the length of a mature mRNA, and the average nascent mRNA was assumed to exhibit half the brightness of a mature mRNA. An extended FSP approach (40) was used to compute the distribution for the number of polymerases at the TS as described in *SI Appendix*. The distribution of TS spot intensities with N_{poly} polymerases was found through the convolution of N_{poly} independent random variables, each with a uniform distribution between zero and one. The FSP analysis was confirmed using stochastic simulation as described in *SI Appendix*, Fig. S13. TS sites were labeled as ON if their predicted or measured intensities were greater than twice the intensity of a single mature mRNA.

Identification of mRNA Elongation Rate. The transcription elongation rate was found by computing the TS intensity distribution for *CTT1* at each point in time for 0.2 M and 0.4 M NaCl osmotic shock using the previously identified parameters (*SI Appendix*, Table S4) and one free constant to describe the average elongation rate, k_{elong}. The probability that the observed distributions of *CTT1* TS intensities could have originated from this model was computed for all time points and conditions, and as a function of k_{elong}. This likelihood was maximized for the different biological replicas and NaCl concentrations to determine the uncertainty in this parameter. The simplified theoretical model, which does not account for transitions between active and inactive periods, provided an upper bound on the *CTT1* elongation rates to be 91 ± 9 nt/s. The more detailed spatial FSP approach determined the *CTT1* elongation rates to be 63 ± 14 nt/s. For both cases, the uncertainty is given as the SEM using the five experimental replicas (two for 0.2 M NaCl and three for 0.4 M NaCl). The elongation rate was then fixed to be 63 nt/s, and this rate was used in conjunction with the previously identified parameters to predict the TS intensity distributions for *CTT1* and *STL1* as functions of time in both osmotic shock conditions (Figs. 1*D* and 4 *D* and *E* and *SI Appendix*, Fig. S11 *D–H*).

ACKNOWLEDGMENTS. We thank Luis Aguilera, Anthony Weil, Bill Tansey, Roger Colbran, Alexander Thiemicke, Dustin Rogers, Benjamin Kesler, Rohit Venkat, and Amanda Johnson for comments on the manuscript. This work was supported by W. M. Keck Foundation Grant DTRA FRCALL 12-3-2-0002 and NIH Grant R35GM124747 (to B.E.M. and Z.R.F.) and by NIH Grants DP2 GM11484901 and R01GM115892 and Vanderbilt Startup Funds (to G.L. and G.N.).

1. Zechner C, Unger M, Pelet S, Peter M, Koeppl H (2014) Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat Methods* 11:197–202.
2. Muzzey D, Gómez-Uribe CA, Mettetal JT, van Oudenaarden A (2009) A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell* 138:160–171.
3. Kumar RM, et al. (2014) Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516:56–61.
4. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* 122:169–182.
5. Michor F, et al. (2005) Dynamics of chronic myeloid leukaemia. *Nature* 435:1267–1270.
6. Brenner S (2010) Sequences and consequences. *Philos Trans R Soc Lond B Biol Sci* 365:207–212.
7. Kirk PDW, Babtie AC, Stumpf MPH (2015) Systems biology (un)certainties. *Science* 350:386–388.
8. Femino AM, Fay FS, Fogarty K, Singer RH (1998) Visualization of single RNA transcripts in situ. *Science* 280:585–590.
9. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5:877–879.
10. Bendall SC, et al. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332:687–696.
11. Hammar P, et al. (2014) Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. *Nat Genet* 46:405–408.
12. Gaublomme JT, et al. (2015) Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell* 163:1400–1412.
13. Battich N, Stoeger T, Pelkmans L (2015) Control of transcript variability in single mammalian cells. *Cell* 163:1596–1610.
14. Buenrostro JD, et al. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523:486–490.
15. Sepúlveda LA, Xu H, Zhang J, Wang M, Golding I (2016) Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science* 351:1218–1222.
16. Morisaki T, et al. (2016) Real-time quantification of single RNA translation dynamics in living cells. *Science* 352:1425–1429.
17. Moffitt JR, et al. (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc Natl Acad Sci USA* 113:11046–11051.
18. Bintu L, et al. (2016) Dynamics of epigenetic regulation at the single-cell level. *Science* 351:720–724.
19. Frieda KL, et al. (2017) Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541:107–111.
20. Munsky B, Neuert G, van Oudenaarden A (2012) Using gene expression noise to understand gene regulation. *Science* 336:183–187.
21. Hilfinger A, Norman TM, Paulsson J (2016) Exploiting natural fluctuations to identify kinetic mechanisms in sparsely characterized systems. *Cell Syst* 2:251–259.
22. Ruess J, Milias-Argeitis A, Lygeros J (2013) Designing experiments to understand the variability in biochemical reaction networks. *J R Soc Interface* 10:20130588.
23. Neuert G, et al. (2013) Systematic identification of signal-activated stochastic gene regulation. *Science* 339:584–587.
24. Ljung L (1999) *System Identification, Theory for the User* (Prentice Hall, Upper Saddle River, NJ), 2nd Ed.
25. Krzywinski M, Altman N (2013) Points of significance: Importance of being uncertain. *Nat Methods* 10:809–810.
26. Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: From microbes to mammals. *Cell* 144:910–925.
27. Süel GM, Kulkarni RP, Dworkin J, Garcia-Ojalvo J, Elowitz MB (2007) Tunability and noise dependence in differentiation dynamics. *Science* 315:1716–1719.
28. Raj A, Rifkin SA, Andersen E, van Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. *Nature* 463:913–918.
29. Saito H, Posas F (2012) Response to hyperosmotic stress. *Genetics* 192:289–318.
30. Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys* 124:044104.
31. Komorowski M, Costa MJ, Rand DA, Stumpf MPH (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc Natl Acad Sci USA* 108:8645–8650.
32. Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH (2011) Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* 332:475–478.
33. Mason PB, Struhl K (2005) Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol Cell* 17:831–840.
34. Nevozhay D, Adams RM, Van Itallie E, Bennett MR, Balázsi G (2012) Mapping the environmental fitness landscape of a synthetic gene circuit. *PLoS Comput Biol* 8:e1002480.
35. Ferrigno P, Posas F, Koepp D, Saito H, Silver PA (1998) Regulated nucleo/cytoplasmic exchange of HOG1 MAPK requires the importin β homologs NMD5 and XPO1. *EMBO J* 17:5606–5614.
36. Hespanha JP, Singh A (2005) Stochastic models for chemically reacting systems using polynomial stochastic hybrid systems. *Int J Robust Nonlinear Control* 15:669–690.
37. Wishart J (1928) The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* 20A:32–52.
38. Fox Z, Neuert G, Munsky B (2016) Finite state projection based bounds to compare chemical master equation models using single-cell data. *J Chem Phys* 145:074101.
39. Chib S, Greenberg E (2012) Understanding the metropolis-hastings algorithm. *Am Stat* 49:327–335.
40. Senecal A, et al. (2014) Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep* 8:75–83.