April 28, 2020

School of Information Science and Engineering
East China University of Science and Technology
130 Meilong Road, Shanghai 200237
Phone: +86-136-5583-9386
Email: zcao@ecust.edu.cn

Department of Chemical Engineering
McGill University
3610 University Street
Montréal, Québec, Canada H3A 0C5

Dear Chair of Search Committee,

I am delighted to apply for the tenure-track position in the Department of Chemical Engineering, McGill University. I am currently a professor of Information Science and Engineering at East China University of Science and Technology (ECUST) working at the interface of machine learning, process system engineering and systems biology.

Previously, I received a B.Eng. degree from Zhejiang University, a top 3 university in China, in 2012 with the National Scholarship of China, and a Ph.D. degree in Chemical and Biomolecular Engineering from the Hong Kong University of Science and Technology (HKUST) with the Chan & Wong Best Postgraduate Award (only one recipient per year), under the supervision of a McGill alumnus – Professor Furong Gao. Later on, I worked as a postdoctoral fellow in the School of Engineering and Applied Sciences, Harvard University, collaborating with the school dean Professor Francis J. Doyle III. After that, I moved to the United Kingdom and took a postdoctoral associate position in the School of Biological Science, the University of Edinburgh. I collaborated with a former McGill professor – Professor Peter Swain and Professor Ramon Grima, pursuing a cutting-edge topic – modeling gene regulatory networks. At this point, *you can see that the McGill spirit has been deeply instilled into my academic genes, which also explains my delightfulness in the beginning*. In 2019, I was awarded a prestigious grant – the Thousand Talent Plan of China with a start-up fund of 1,200,000 Canadian Dollars, and was promoted to professor shortly afterwards at ECUST. More detailed information is included in my curriculum vitae.

Broadly, my research interests pertain to systems biology and machine learning. Transcription, the step of transferring genetic information from DNA to RNA, is of paramount importance, as it constitutes the starting point of myriads of intracellular processes. Thus, gaining a fundamental understanding of transcription and its underlying regulatory mechanism has defined one of the central topics for modern molecular biology. However, distilling such physiological information from single-cell transcriptional data is remarkably daunting, as the underlying details are masked by the transcriptional "noise", a phenomenon principally arising from the low copy number of biological macromolecules. Unlike the state of the art, we intend to present a solution to this outstanding problem by an interdisciplinary approach, making use of methods from machine learning, information theory and statistical physics, and tightly coupling theories with experiments. All such techniques and theories work together to pursue the ultimate goal – to reach a narrative for transcriptional processes that is quantitative, simple and universal. Besides, I am also interested in using machine learning tools to solve problems from synthetic biology, biomedical engineering and classic chemical engineering while closely collaborating with industrial partners. For more technical details, please kindly refer to the attached research statement.

Additionally, I am happy to contribute to the rich and diverse chemical engineering curriculum at McGill by teaching both undergraduate and graduate courses on the topic of machine learning, process control and systems biology. In particular, I am willing to develop two new courses – one for undergraduate students named "Introduction to Machine Learning for Chemical Engineers" (of which a syllabus is attached) and one for graduate students called "Practical Systems Biology".

Given my solid academic training by McGill alumni, I believe I can make a difference to the fields of machine learning, process control and systems biology on the fantastic platform of McGill. Hence I hope that you will find me well suited for the applied position.

Should you have any inquiry or require additional materials, please feel free to contact me via phone (+86-136-5583-9386), email (`zcao@ecust.edu.cn`), or my homepage (`https://edwardcao3026.github.io`).

Sincerely,

Zhixing Cao (Edward), PhD

Professor of Information Science and Engineering
East China University of Science and Technology

130 Meilong Road, Shanghai 200237, China
http://www.ecust.edu.cn

ZHIXING CAO (EDWARD)

+ 86-136-5583-9386 ◇ zcao (at) ecust (dot) edu (dot) cn
Professor of Information Science and Engineering ◇ East China University of Science and Technology
130 Meilong Road ◇ Xuhui ◇ Shanghai ◇ China 200237

PERSONAL INFORMATION

Secondary email:	edwardcao@g.harvard.edu zcaoab@connect.ust.hk edward.cao@ed.ac.uk
Homepage:	https://edwardcao3026.github.io
Skills:	LaTeX, Matlab, Mathematica, C, C++, Linux Python, Julia

ACADEMIC EXPERIENCE

Zhejiang University *June 2008 – July 2012*
B.Eng. in Control Science & Engineering
Overall GPA: 3.95/4.0; Overall Rank: 2/118

Hong Kong University of Science and Technology *August 2012 - July 2016*
Ph.D. student in Chemical & Biomolecular Engineering (CBME)
Supervisor: Prof. Furong Gao (Chair Professor)

Universität Stuttgart *February 2015 - May 2015*
Visiting Scholar at Institute of Systems Thoery and Automatic Control (IST)
Supervisor: Prof. Frank Allgöwer (IFAC president), Prof. Christian Ebenbauer

Harvard University *February 2016 - February 2017*
Postdoctoral Fellow in Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS)
Supervisor: Prof. Francis J. Doyle III (Dean)

University of Edinburgh *June 2017 - April 2019*
Postdoctoral Fellow in Centre for Synthetic and Systems Biology
Supervisor: Prof. Ramon Grima, Prof. Peter Swain

East China University of Science and Technology *April 2019 - present*
Professor of Information Science and Engineering
Jointly affiliated with State Key Laboratory of Bioreactor Engineering

RESEARCH KEYWORDS

Systems biology, process control, process modeling, process optimization, batch process, statistical data analysis, stochastic process, data-driven modeling, iterative learning control, injection molding, physics-informed machine learning, stochastic gene expression, machine learning application in biology and chemical engineering

AWARDS

B.S. period:

- National Scholarship supported by Ministry of Education of China

Ph.D. student period:

- Hong Kong Ph.D. Fellowship supported by Research Grant Council (RGC) in Hong Kong
- Oversea Research Award supported by the Hong Kong University of Science and Technology
- Chan & Wong Best Postgraduate Award at the Hong Kong University of Science and Technology (Award committee chairman: Prof. Guohua Chen, AIChE Fellow)

Postdoctoral period:

- Outstanding Reviewer for *ISA Transactions*
- Outstanding Reviewer for *Chemometrics and Intelligent Laboratory Systems*
- Outstanding Reviewer for *Journal of Process Control*
- Outstanding Reviewer for *Industrial & Engineering Chemistry Research*
- Seal of Excellence, H2020 Marie Skłodowska Curie Actions

Faculty period:

- Thousand Talents Plan in China (Young Professorship Program)
  (start-up fund 1,200,000 Canadian Dollars)

PROFESSIONAL ACTIVITIES

Member of Society of Plastics Engineers

Member of Hong Kong Institute of Science

Guest Editor of *Journal of Control Science and Engineering*

Reviewer of *IEEE Transactions on Automatic Control, IEEE Transactions on Industrial Electronics, IEEE Transactions on Biomedical Engineering, Journal of the Royal Society Interface, ISA Transaction, Chemometrics and Intelligent Laboratory Systems, Industrial & Engineering Chemistry Research, the Canadian Journal of Chemical Engineering, Chemical Engineering Science, Journal of Process Control, Processes, Sensors, Algorithms, Frontiers Cell and Developmental Biology, Biophysical Journal, Mathematical Biosciences, Applied Sciences, American Control Conference, Conference on Decision and Control*

PUBLICATION LIST (PEER REVIEWED)

(*The symbol * stands for corresponding author. The highlighted papers are denoted with ✎. The papers are arranged in authorship and chronological order.*)

Journal Publication

✎ [J1] **Z. Cao**, R. Grima, "Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells". *Proceedings of the National Academy of Sciences*, **117**(9): 4682-4692, 2020.

- The paper presents a stochastic gene expression model including a considerable number of salient features of single-cell biology, such as cell division, replication, mRNA maturation, dosage compensation and growth-dependent transcription. With the model solved analytically, it enables insight into how gene expression fluctuations are modified and controlled by complex intracellular processes.

✎ [J2] **Z. Cao***, J. Yu, W. Wang, H. Lu, X. Xia, H. Xu, X. Yang, L. Bao, Q. Zhang, H. Wang, S. Zhang, L. Zhang*, "Multi-scale data-driven engineering for biosynthetic titer improvement", *Current Opinion in Biotechnology*, (*Accepted*).

- This review aims to summarize the state of the art of biosynthesis titer improvement on different scales separately, particularly regarding the advancement of metabolic pathway rewiring and data-driven process optimization & control.

· [J3] **Z. Cao**, R. Grima. "Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data". *Journal of The Royal Society Interface*, **16**(153): 20180967, 2019.

✎ [J4] **Z. Cao**, R. Grima, "Linear mapping approximation of gene regulatory networks with stochastic dynamics", *Nature Communications*, **9**(1): 3305, 2018.

- A novel gene regulatory network modeling technique named linear mapping approximation is developed to map systems with protein–promoter interactions onto approximately equivalent systems with no binding reactions, allowing systematic exploration of the stochastic properties of nonlinear gene regulatory network in systems and synthetic biology.

· [J5] **Z. Cao**, R. Gondhalekar, E. Dassau, F. J. Doyle III, "Extremum seeking control for personalized zone adaptation in model predictive control for Type 1 diabetes", *IEEE Transactions on Biomedical Engineering*, **65**: 1859-1870, 2018.

· [J6] **Z. Cao**, J. Lu, R. Zhang, F. Gao, "Online average-based system modelling method for batch process", *Computers & Chemical Engineering*, **108**: 128-138, 2018.

· [J7] **Z. Cao**, H.B. Dürr, C. Ebenbauer, F. Allgöwer, F. Gao, "Iterative learning and extremum seeking for repetitive time-varying mappings", *IEEE Transaction on Automatic Control*, **62**(7): 3339-3353, 2017. (Regular paper).

· [J8] **Z. Cao**, J. Lu, Y. Yang, F. Gao, "Iterative learning Kalman filter design for repetitive processes", *Journal of Process Control*, **46**: 92-104, 2016.

· [J9] **Z. Cao**, Y. Yang, H. Yi, F. Gao, "Priori knowledge based online closed-loop identification for batch process with an application to injection molding", *Industrial & Engineering Chemistry Research*, **55**(32): 8818-8829, 2016.

· [J10] **Z. Cao**, R. Zhang, J. Lu, F. Gao, "Online identification for batch processes in closed loop incorporating priori controller knowledge", *Computers & Chemical Engineering*, **90**: 222-233, 2016.

· [J11] **Z. Cao**, R. Zhang, J. Lu, F. Gao, "Two time dimensional recursive system identification incorporating priori pole and zero knowledge", *Journal of Process Control*, **39**: 100-110, 2016.

· **[J12] Z. Cao**, R. Zhang, J. Lu, F. Gao, "Discrete-time robust iterative learning Kalman filtering for repetitive processes", *IEEE Transaction on Automatic Control*, **61**(1): 270-275, 2016.

· **[J13] Z. Cao**, Y. Yang, J. Lu, F. Gao, "Two-time-dimensional model predictive control of weld line positioning in bi-injection molding", *Industrial & Engineering Chemistry Research*, **54**(17): 4795-4804, 2015.

· **[J14] Z. Cao**, Y. Yang, J. Lu, F. Gao,"Constrained two dimensional recursive least squares model identification for batch processes", *Journal of Process Control*, **24**(6): 871-879, 2014.

✎ [J15] J. Lu, **Z. Cao***, C. Zhao, F. Gao. "110th Anniversary: An overview on learning-based model predictive control for batch processes". *Industrial & Engineering Chemistry Research*, **58**(37): 17164-17173, 2019.

- This review summarizes the recent technical advancements during the past two decades, from the perspective of the three different levels of learning mechanisms: control input, model parameter and tracking reference.

· **[J16]** J. Lu, **Z. Cao***, F. Gao, "Multi-point iterative learning model predictive control", *IEEE Transactions on Industrial Electronics*, **66**(8): 6230-6240, 2018.

· **[J17]** J. Lu, **Z. Cao***, R. Zhang, F. Gao, "Nonlinear monotonically convergent iterative learning control for batch processes", *IEEE Transactions on Industrial Electronics*, **65**(7): 5826-5836, 2017.

· **[J18]** J. Holehouse, **Z. Cao**, R. Grima, "Stochastic modeling of auto-regulatory genetic feedback loops: a review and comparative study". *Biophysical Journal*, **118**: 1517-1525, 2020.

· **[J19]** J. Lu, **Z. Cao**, F. Gao, "Batch process control – survey and perspective", *Acta Automatica Sinica*, **43**(6): 933-943, 2017.

· **[J20]** J. Lu, **Z. Cao**, F. Gao, "Ellipsoid invariant set based robust model predictive control for constrained batch processes", *IET Control Theory & Applications*, **10**(9): 1018-1026, 2016.

· **[J21]** R. Zhang, **Z. Cao**, R. Lu, P. Li, F. Gao, "State-space predictive-P control for liquid level in an industrial coke fractionation tower", *IEEE Transaction on Automation Science Engineering*, **12**(4): 1516-1524, 2016.

· **[J22]** J. Lu, **Z. Cao**, Z. Wang, F. Gao, "A two-stage algorithm based on two-dimensional model predictive iterative learning control for non-repetitive disturbances attenuation", *Industrial & Engineering Chemistry Research*, **54**(21): 5683-5689, 2015.

· **[J23]** R. Zhang, **Z. Cao**, C. Bo, P. Li, F. Gao, "New PID controller design using extended nonminimal state space model based predictive functional control structure", *Industrial & Engineering Chemistry Research*, **52**(8): 3236-3245, 2014.

· **[J24]** R. Zhang, **Z. Cao**, P. Li, F. Gao, "Design and implementation of an improved linear quadratic regulation control for oxygen content in a coke furnace", *IET Control Theory & Applications*, **8**(14): 1303-1311, 2014.

· **[J25]** R. Zhang, S. Wu, **Z. Cao**, J. Lu, F. Gao, "A systematic min-max optimization design of constrained model predictive tracking control for industrial processes against uncertainty", *IEEE Transactions on Control Systems Technology*, **26**(6): 2157-2164, 2017.

· **[J26]** R. Zhang, Q. Zou, **Z. Cao**, F. Gao, "Design of fractional order modeling based extended nonminimal state space MPC for temperature in an industrial electric heating furnace", *Journal of Process Control*, **56**: 13-22, 2017.

Submitted Journal Publication

- **[S1] Z. Cao***, J. Zhu, J. Lu, F. Gao, "Comparison study for process monitoring methods to detect abnormalities in gene regulatory networks", Submitted to *IEEE Transactions on Industrial Electronics*.

- **[S2] Z. Cao**, T. Filatova, D.A. Oyarzún, R. Grima. "Multi-scale bursting in stochastic gene expression". Submitted to *Biophysical Journal*.

- **[S3]** J. Lu, **Z. Cao***, Q. Hu, Z. Xu, F. Gao, "Optimal Iterative Learning Control for Batch Processes in the Presence of Time-varying Dynamics", Submitted to *IEEE Transactions on System, Man and Cybernetics: Systems*.

Conference Publication

- **[C1] Z. Cao**, E. Dassau, R. Gondhalekar, F. J. Doyle III, "Extremum seeking control based zone adaptation for zone model predictive control in type 1 diabetes", in *Proceedings of IFAC World Congress, Toulouse, France*: 15639-15644, 2017.

- **[C2] Z. Cao**, F. Gao, J. Lu, Y. Yang, "From single batch process control to multiple batch processes control: a review and a perspective for injection molding", in *Proceedings of SPE Annual Technical Conference*: 2271-2276, 2014.

- **[C3] Z. Cao**, Y. Yang, J. Lu, F. Gao, "Two dimensional recursive least squares for batch processes system identification", in *Proceedings of IFAC International Symposium on Dynamics Control Process Systems*: 780-784, 2013.

- **[C4] Z. Cao**, Y. Yang, F. Gao, "New approaches of 2D recursive least squares system identification for batch processes", in *Proceedings of SPE Annual Technical Conference*: 2223-2228, 2013.

- **[C5]** J. Lu, **Z. Cao**, F. Gao, "A stable two-time dimensional (2D) Model Predictive Control with zero terminal state constraints for constrained batch processes", in *Proceedings of IFAC International Symposium on Advanced Control of Chemical Processes*, **48**(8): 514-519, 2015.

- **[C6]** J. Lu, **Z. Cao**, F. Gao, "A repetitiveness index-based adaptive two dimensional iterative learning model predictive control", in *Proceedings of IFAC World Congress, Cape Town, South Africa*: 3092-3097, 2015.

- **[C7]** J. Lu, **Z. Cao**, F. Gao, "A two-dimensional iterative learning model predictive control method for injection molding based on mixed integer quadratic programming", in *Proceedings of SPE Annual Technical Conference*: 2311-2316, 2014.

- **[C7]** B. Zhang, **Z. Cao**, X. Li, H. Su, X. Chen, "High temperature proton exchange membrane fuel cell hybrid power system monitoring platform designing", in *Proceedings of International Conference on Automatic Control Artificial Intelligence*: 1129-1132, 2012.

- **[C9]** J. Lu, D. Li, **Z.Cao**, and F. Gao, "Rejection of periodic disturbances based on adaptive repetitive model predictive control", in *Proceedings of IFAC International Symposium on Dynamics Control Process Systems*: 768-773, 2013.

- **[C10]** X. Chen, X. Li, H. Su, **Z. Cao**, B. Zhang, "Multiple model predictive control for oxygen starvation prevention of fuel cell system", in *Proceedings of International Conference on Automatic Control Artificial Intelligence*: 1368-1371, 2012.

INVITED TALKS

- **[T1]** Tenure-track interview invited by Eric Croiset, Department of Chemical Engineering, University of Waterloo, Feb. 2019.
- **[T2]** Tenure-track interview invited by Peter Engelzos, Department of Chemical and Biological Engineering, University of British Columbia, Jul. 2018.
- **[T3]** College of Electrical Engineering and Control Science, Nanjing University of Technology, Jun. 2018.
- **[T4]** Young Talent Forum on Intelligent Connection, Computing & Control held by East China University of Science and Technology, May. 2018
- **[T5]** Mathematical Biology Seminar (funded by Higgs Center), University of Edinburgh, Oct. 2017.
- **[T6]** Stephen Duncan's Group Meeting, University of Oxford, Jun. 2017.
- **[T7]** Seminar of Department of Chemical & Biomolecular Engineering, Hong Kong University of Science and Technology, Feb. 2017.
- **[T8]** Johan Paulsson's Group Meeting, Harvard Medical School, Jan. 2017.
- **[T9]** Francis J Doyle III's Group Meeting, Harvard University, Jan. 2016.
- **[T10]** Frank Allgöwer's Group Meeting, Universität Stuttgart, Apr. 2015.

CONFERENCE PRESENTATIONS

- **[P1]** *From single batch process control to multiple batch processes control: a review and a perspective for injection molding* , SPE Annual Technical Conference, Las Vegas, NV, USA, 2014.
- **[P2]** *New approaches of 2D recursive least squares system identification for batch processes*, SPE Annual Technical Conference, Cincinnati, OH, USA, 2013.

TEACHING EXPERIENCE

- **Department of Chemical and Biomolecular Engineering, HKUST**
  Teaching Assistant – 2013-2014 Fall & Spring, 2014-2015 Fall
  Responsible for lectures and recitations on the course CENG4120 "Process Dynamics and Control"
- **Department of Chemical and Biomolecular Engineering, HKUST**
  Master Student Project Supervisor – 2014-2015
  Supervising 4 master student
- **Fok Ying Tung Graduate School, HKUST**
  Project Manager – 2017
  Supervising 1 PhD student and 2 Postdocs (The supervision resulted in joint publications J15, J16, J17 that were correspondingly authored by me.)

· **School of Information Science and Technology, ECUST**
Supervisor of Undergraduate Final Year Project – 2020
– Project A: *Machine-learning based solver for delayed chemical master equation*
– Project B: *Inference of transcriptional kinetic parameters from single-cell nascent mRNA counts*

· **School of Information Science and Technology, ECUST**
Postgraduate Course Organizer – 2020
– Organizing the Ph.D. course "Intelligence Theory and Applications"

· **School of Information Science and Technology, ECUST**
Undergraduate Core Course Organizer – 2020
– Organizing the undergraduate course "Introduction to Machine Learning for Chemical Engineers"

COLLABORATORS

· **Prof. Furong Gao**
Hong Kong University of Science and Technology

· **Prof. Francis J Doyle III**
Harvard University

· **Prof. Frank Allgöwer**
University of Stuttgart

· **Prof. Ramon Grima**
University of Edinburgh

· **Prof. Peter Swain**
University of Edinburgh

· **Prof. Diego A. Oyarzún**
University of Edinburgh

· **Prof. Christian Ebenbauer**
University of Stuttgart

· **Prof. Shen Zeng**
Washington University in St. Lious

· **Prof. Shuyang Lin**
New York University

· **Prof. Philipp Thomas**
Imperial College London

· **Prof. Yuan Yao**
National Tsing Hua University

· **Prof. Lixin Zhang**
East China University of Science and Technology

- **Prof. Chunhui Zhao**
  Zhejiang University

- **Prof. Xi Chen**
  Zhejiang University

- **Dr. Yi Yang** (Industrial partner)
  Shenzhen Time High-Tech equipment Co. Ltd (for drying equipment of lithium battery)

- **Mr. Shengjun Li** (Industrial partner)
  Acting CEO of Tongkun Group Co. Ltd (manufacturing polyester filament yarn)

REFERENCES

- **Prof. Furong Gao** (*McGill alumnus*)
  Supervisor; Chair Professor, Department of Chemical and Biological Engineering
  Hong Kong University of Science and Technology
  Email: kefgao@ust.hk
  Phone: +852-2358-7139
  Address: RM 4558, HKUST, Clear Water Bay, Kowloon, Hong Kong

- **Prof. Guohua Chen** (*McGill alumnus*)
  Former Head of Department; Associate Vice President, Department of Mechanical Engineering
  Hong Kong Polytechnic University
  Email: guohua.chen@polyu.edu.hk
  Phone: +852-2766-6647
  Address: FG 606, Hong Kong Polytechnic University, Hong Hom, Kowloon, Hong Kong

- **Prof. Peter Swain** (*Former McGill faculty*)
  Professor, School of Biological Sciences
  University of Edinburgh
  Email: peter.swain@ed.ac.uk
  Phone: +44-131-650-5451
  Address: SynthSys, Max Born Crescent, Edinburgh EH9 3BF, Scotland, UK

- **Prof. Ramon Grima**
  Professor, School of Biological Sciences
  University of Edinburgh
  Email: ramon.grima@ed.ac.uk
  Phone: +44-131-650-9060
  Address: RM 3.03, SynthSys, Max Born Crescent, Edinburgh EH9 3BF, Scotland, UK

Research Statement

Zhixing Cao (Edward)

Section 1 of this statement is devoted to the summary of my past research projects, and it is followed by a detailed presentation of an ongoing project in Section 2, which constitutes the central piece of one of my proposals submitted to *the National Natural Science Foundation of China*. Some interesting problems promisingly leading to high-profile results are proposed in Sections 3-5.

1. Previous research

Previously, as a Ph.D. student, I was asked to systematically improve the control performance of a batch-process control system, which comprises four modules – model identifier, state estimator, controller and set-point optimizer. To this end, I harnessed the innate repeatable operational pattern of batch processes and developed an algorithmic framework, wherein the performance of each module was remarkably enhanced. The framework was successfully applied to the injection molding process with encouraging results. At Harvard, I worked with Professor Francis J Doyle III and developed an intelligent *artificial pancreas* (*AP*) to enable automatic personalization of insulin therapy, thereby minimizing physicians' intervention and therapeutic cost. After moving to Edinburgh, I shifted my focus onto seeking the "hidden variables" underpinning cellular heterogeneity in gene expression by modeling the stochastic dynamics therein. The project was jointly advised by Professor Peter Swain and Professor Ramon Grima.

After becoming a PI, I am working on several projects simultaneously, most of which are centered around a common theme – discovering and harnessing hidden patterns underlying fundamental cellular processes while tightly coupling experiment and theory and using interdisciplinary approaches including machine learning, information theory, control theory and statistical physics. I would continue to work on it, if honorably given the opportunity at McGill. My vision and plan for the future 5 years are summarized below in greater detail.

2. Ongoing project: Finding transcriptional "latent variables"

2.1 Introduction

Gaining quantitative insight into the functioning of living cells, particularly the diverse fundamental intracellular processes, has received increasing attention recently and is deemed a key biological research topic for the next decades. Transcription, the step of transferring genetic information from DNA to RNA, is of paramount importance, as it constitutes the starting point of myriads of intracellular processes. **In this study, we aim to reveal the underlying transcriptional regulatory machinery by mathematically modeling the dynamics.** Using mathematical models to reflect the possible underlying mechanism is one of the pervasive approaches in modern molecular biology: a good match between the predictions of a mathematical model and experimental data implies that the model offers a potential explanation of the observation, while a bad match implies that further refinement of the model (and probably further experiments) is necessary. Important though mathematical models are, building a good one for transcriptional dynamics is remarkably daunting. Such a task is fundamentally challenged by transcriptional heterogeneity. Researchers used to believe that it principally arises from the low copy number of biological macromolecules, but it has been suggested by the recent advancements of single-cell experiments that other previously uncharacterized physiological factors ("latent variables") are of equal importance at least. As long as those factors remain hidden, we are prone to interpret the transcriptional regulatory machinery falsely.

To this end, our theories will be tightly coupled with experiments: they will be motivated and conceived from experiments, and used to distill experimental results into novel biological insights. The experiment results we need are from either the *in vivo* RNA visualization technique named *Pepper* [1] or *native elongating transcription sequencing* (*NET-Seq*) [2]. To obtain the former, we are collaborating with the Pepper developers at East China University of Science and Technology and intend to later set our own wet lab and perform experimental measurements by ourselves. Meanwhile, we are working with a team led by Dr. Chris Sibley at the University of Edinburgh to have NET-Seq data. Besides, the theoretical work will be conducted in an interdisciplinary framework using the toolkits of *machine learning* (*ML*), process control theory, information theory and statistical physics. All such techniques and toolkits work together to pursue the ultimate goal, as always in physics, to reach a narrative for transcriptional processes that is quantitative, simple and universal. More specific details will be elaborated in Section 2.3.

2.2 Our previous efforts

One of the commonly accepted frameworks for modeling transcriptional dynamics is *chemical master equations* (*CMEs*), which is comprised of a set of difference-differential equations suitably capturing the two dynamical characteristics of transcription – state discreteness (mRNA numbers are integers) and temporal randomness (CMEs assume the time between any two successive reactions complies with exponential distributions). Thus, CMEs were and will be the central piece that our studies rest upon. **Here we highlight two of our previous attempts for modeling transcriptional dynamics**: the first one [3] is the development of *linear mapping approximation* (*LMA*) for analytically solving CMEs in the presence of feedback, whereas the second one [4] presents a stochastic model and its analytical solution for stochastic transcriptional dynamics with a considerable number of salient features of cellular physiology.

LMA: Despite the great modeling capability of CMEs, it is also notorious for the intractability of analytical solutions except a few special cases [8–11], thereby hindering the efficient and accurate interpretation of experimental data. Such a problem becomes acute in the presence of transcriptional feedback wherein a transcriptional factor binds to a promoter to activate or inhibit the transcription (see Fig. 1A top). To resolve such a problem, we developed the LMA method wherein a nonlinear *gene regulatory network* (*GRN*) is mapped onto a linear one, while the solution to the CMEs of the linear is readily available.

Specifically, we use the auto-regulatory feedback GRN (Fig. 1A top) to illustrate the fundamental ideas of LMA. The feedback GRN is comprised of two gene states (G and G^*) with different transcription rates (ρ_{u} and ρ_{b}). The two states can be switched from one to the other by binding or unbinding a protein. The method LMA maps the feedback GRN on to a linear GRN (Fig. 1A bottom) where all reactions are of first order and whose solution to the protein number distribution at steady state is denoted as $P(\bar{\sigma}_{\mathrm{b}})$. Hence, the protein number distribution of the feedback GRN can be determined by plugging the re-parametrized effective rate $\bar{\sigma}_{\mathrm{b}}$ into $P(\bar{\sigma}_{\mathrm{b}})$, and $\bar{\sigma}_{\mathrm{b}}$ can be determined by solving the average number of proteins conditional on the promoter being in state G. Specifically, we solve the conditional mean $\langle n_p|n_g=1\rangle$ (where n_p is the number of proteins, $n_g=1$ means the state G and $\langle\cdot\rangle$ is the expectation operator) from the moment equations of the linear GRN.

$$\begin{cases} \partial_t\langle n_p\rangle = \rho_{\mathrm{u}}\langle n_g\rangle + \rho_{\mathrm{b}}(1-\langle n_g\rangle) - \langle n_p\rangle, \\ \partial_t\langle n_g\rangle = -\bar{\sigma}_{\mathrm{b}}\langle n_g\rangle + \sigma_{\mathrm{u}}(1-\langle n_g\rangle), \\ \partial_t\langle n_p n_g\rangle = \rho_{\mathrm{u}}\langle n_g\rangle + \sigma_{\mathrm{u}}\langle n_p\rangle - (1+\bar{\sigma}_{\mathrm{b}}+\sigma_{\mathrm{u}})\langle n_p n_g\rangle, \end{cases} \Rightarrow \langle n_p|n_g=1\rangle = \frac{\rho_{\mathrm{u}}+\rho_{\mathrm{b}}\bar{\sigma}_{\mathrm{b}}+\rho_{\mathrm{u}}\sigma_{\mathrm{u}}}{1+\sigma_{\mathrm{u}}+\bar{\sigma}_{\mathrm{b}}} = f(\bar{\sigma}_{\mathrm{b}}). \quad (1)$$

Then we solve the effective rate $\bar{\sigma}_{\mathrm{b}}$ by equating $\bar{\sigma}_{\mathrm{b}} = \sigma_{\mathrm{b}} f(\bar{\sigma}_{\mathrm{b}})$ and plug the yielded solution into $P(\bar{\sigma}_{\mathrm{b}})$ of the linear GRN, giving us an approximate closed-form solution to the protein number distribution of the nonlinear GRN. It is shown in Fig. 1B that LMA prediction agrees well with that of the *stochastic simulation algorithm* (*SSA*) over vast swathes of kinetic parameter space. Integrated with Magnus series expansion, LMA is also able to predict time-dependent distributions accurately (see Fig. 1C).

Detailed stochastic transcriptional model: The two-state model (Fig. 1A top) is the standard model of stochastic mRNA dynamics in eukaryotic cells and the most commonly used one to fit experimental data. Tough the two-state model well depicts the reactive components (e.g. mRNA degradation), the detailed biological knowledge gleaned from single-cell experiments suggests that non-reactive components plays an equally important role for cellular stochasticity. Such knowledge has not been incorporated in numerous models in the literature, thereby making us prone to interpret experimental data inaccurately and lead to erroneous biological conclusions. To address such an issue, we extended the two-state model to include mRNA maturation, cell division, gene replication, dosage compensation, and growth-dependent transcription (see Figs. 1E& 1F). Despite the model's complexity, we still did derive expressions for the time-dependent distributions of nascent mRNA and mature mRNA numbers, provided two assumptions hold: 1) nascent mRNA dynamics are much faster than those of mature mRNA; and 2) gene-inactivation events occur far more frequently than gene-activation events. We confirmed that thousands of eukaryotic genes satisfy these assumptions by using data from yeast, mouse, and human cells. Surprisingly, we even found that the time-dependent distributions predicted by our model are generally well approximated by the negative binomial distribution (see Fig. 1G), thereby enabling detailed quantitative insight into how cells coordinate diverse physiological processes to regulate the fluctuations of gene expression.

2.3 Main aim of the current proposal

The novelties of this proposed research come from: (i) to date there is no approximation method which solves time-dependent solutions for delayed CMEs; (ii) transcriptional bursting is ubiquitous to a vast set of genes, yet the underlying reason remains elusive; (iii) transcriptomic data indicates there exists a optimal length of mRNA, but what accounts for such optimality is not clear. Hence, the research program is divided into three objectives accordingly, each one allowing us to look into transcription through different lenses.

Objective 1: Modeling nascent mRNA dynamics by physics-informed machine learning techniques. To understand the characteristics of the transcription process, it is common to map mature mRNA counts to the transcriptional

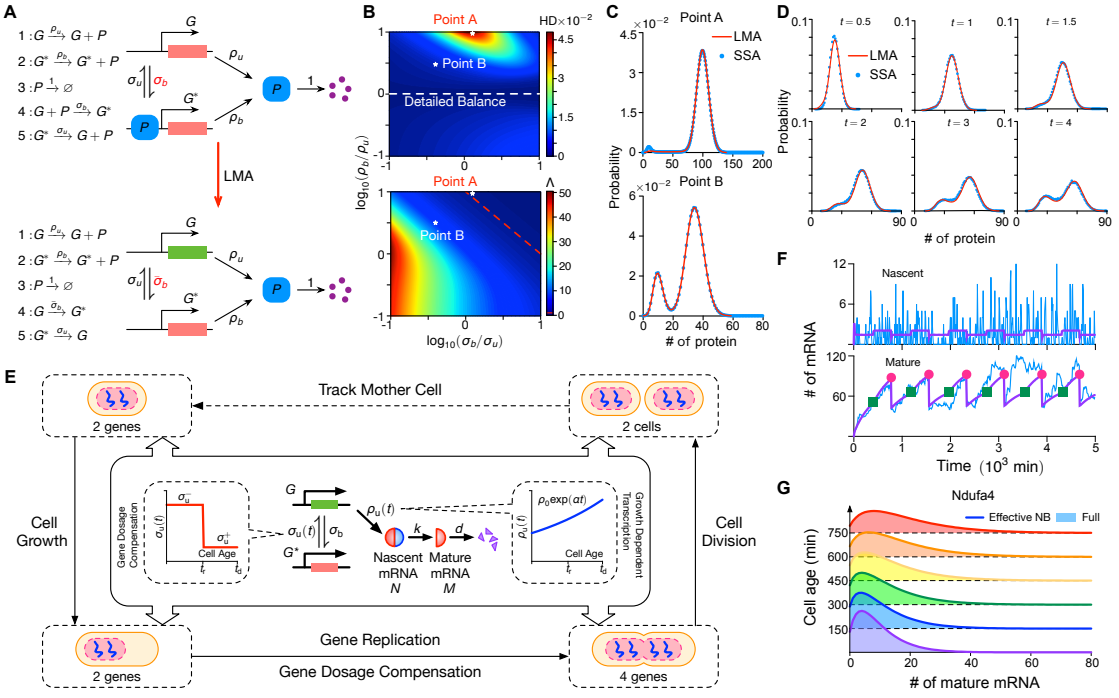


Figure 1: A sketch of ideas in previous efforts. (A) Illustration of the main idea of LMA approximating the binding (nonlinear) reaction of protein (blue squares) and promoter in the nonlinear GRN by a first-order (linear) reaction with a re-parametrized effective reaction rate ($\bar{\sigma}_b$) in the linear GRN. (B) The top shows the LMA prediction error of steady-state protein number distribution against the exact solution [5] of a GRN in panel A as a function of leakage production rate ρ_b and inactivation rate σ_b. The error is quantified in *Hellinger distance* (*HD*). The bottom shows the ratio (Λ) of the two eigenvalues of the Jacobian of the deterministic rate equations of the nonlinear GRN in steady state conditions, indicating that the approximation accuracy of LMA is independent of the time scales in a system. The red broken line denotes the contour line of $\Lambda = 1$. Parameters: $\rho_u = 10$ and $\sigma_u = 0.01$. (C) We show that the LMA and SSA predictions match well for Points A and B in panel B, while Point A corresponds to the point with the largest HD in panel B. (D) We show that the LMA is able to predict protein number distribution as a function of time with impressive accuracy. Parameters: $\rho_u = 60$, $\rho_b = 25$, $\sigma_b = 0.004$ and $\sigma_u = 0.25$. (E) Illustration of the detailed stochastic transcriptional model. Nascent mRNA is shown as joined blue and red semicircles, illustrating its unspliced nature (blue for introns and red for exons), while mature mRNA being composed of only exons is shown as red semicircles. The model is composed of nonreactive components (dosage compensation, replication, cell division, and growth-dependent transcription) and reactive components; the latter are shown in the central boxes. (F) We show stochastic simulations of the model in panel E using the SSA, where the purple lines denote the mean, and a typical time series is shown in blue. The green squares and red dots indicate the gene-replication time (t_r) and cell-division time (t_d), respectively. We use parameters measured for *Nanog* in mouse embryonic stem cells [6]. (G) We surprisingly found that the negative binomial (NB) distribution can accurately predict the mRNA number distribution of model in panel E well in time for a vast set of genes. The gene is *Ndufa4* in mouse embryonic stem cells [7].

kinetics predicted by a model. Nevertheless, the stochasticity of such mature mRNA counts may be masked by additional processes downstream of transcription such as cell division. By contrast, the nascent mRNA – the mRNAs still actively transcribed at the gene may not be subject to these effects, thereby bearing more closely the signature of the transcription process.

However, the stochastic modeling of nascent mRNA dynamics is still at its infancy, as the process is remarkably impacted by delays. As shown in Fig. 2A top, a nascent mRNA needs some elongation time before its detachment from the gene and becoming a mature mRNA. Indeed, delays exist in the other cellular processes pertinent to nascent mRNA as well, such as nascent mRNA degradation by RNases. Therefore, the delays must be incorporated in the framework of CMEs, whereas CMEs have to be appropriately modified to accommodate the non-Markovian dynamics induced by the delays. Specifically, let us consider a simplified model for nascent mRNA dynamics shown in Fig. 2B bottom, wherein only two types of reactions can occur

$$\varnothing \xrightarrow{\rho} N, \qquad N \overset{d}{\underset{\tau}{\Rightarrow}} \varnothing. \tag{2}$$

The wide arrow denotes the delayed reaction in which the maturation process is initiated at rate d and requires time τ to complete, and the letter N stands for nascent mRNA. The corresponding modified CME is

$$\partial_t \underbrace{P(n,t)}_{\substack{\text{Marginal}\\ \text{distribution}}} = \rho(\mathbb{E}^{-1}-1)P(n,t) + d\sum_{m=0}^{\infty} m(\mathbb{E}-1)H(n)\underbrace{P(n,t;m,t-\tau)}_{\substack{\text{Joint}\\ \text{distribution}}}, \tag{3}$$

where the $\mathbb{E}$ is the unitary shift operator, $\mathbb{E}P(n,t) = P(n+1,t)$, $P(n,t;m,t-\tau)$ is the joint probability of having n molecules at time t and m molecules at time $t-\tau$, and $H(n)$ is the Heaviside function. Notably, Eq. (3) is not closed, as the one-point probability distribution is determined by the two-point probability distribution. The only possibility to solve Eq. (3) is to use the delayed version of SSA [12], while such an SSA method is endowed with the drawback that all Monte-Carlo methods have – overwhelming computational cost. Therefore, it calls for an efficient and accurate solution to pave the way for interpreting single-cell data of nascent mRNA counts. On the other hand, it is noted from the Bayes' theorem that the joint distribution can be decomposed as $P(n,t;m,t-\tau) = P(n,t)P(m,t-\tau|n,t)$, which further allows us to compactly rewrite Eq. (3) in the discrete-time form:

$$\mathbf{P}(t+\Delta t) = f[\mathbf{P}(t)] + g[\mathbf{P}(t), \mathbf{P}(t-\tau|t)], \tag{4}$$

with $\mathbf{P}(t) = [P(0,t), \cdots, P(N,t)]^\top$ for some large integer N. Computing the histogram of adequate realizations in SSA to solve Eq. (3) is indeed a "big-data" approach demanding substantial computational resources, and so is the approach wherein a purely data-driven model is built using machine learning tools and asks for plenty of data to fix the large set of hyperparameters in the model. By contrast, we intend to harness the information offered by the physical model and combine "small data" to solve Eq. (3): the "small data" is obtained by simulating SSA for a small number of realizations, while the lack of information is thereafter compensated by the physical model. Specifically for our problem, we will assume a neural-network prior for the mapping $\mathbf{P}(t) \mapsto \mathbf{P}(t-\tau|t)$, which is trained by the small amount of SSA data (see Fig. 2B). In doing so, we are able to solve the delayed CMEs semi-analytically using tractable amount of computational resources, and this method is expected to be scalable to help us understand how regulation functions in a large-scale gene network.

Objective 2: Seeking the origins of long periods of gene transcriptional silence. Tremendous single-cell experimental data has unveiled that the transcription process of most genes has the pattern that the genes stay silent for most of the time but produce a burst of transcripts within a short active window. Such an observation is called *transcriptional bursting* as shown in Fig. 2C. Despite that a number of phenomenological models account for such dynamical patterns (including the two-state model presented in Fig. 1A bottom), there is still a lack of a narrative physical model to offer us simple but universal understanding. We will fill in the gap by interrogating the possible origins of the transcriptional burstiness. On the other hand, it is noted that most human behaviors, such as sending emails, exhibit similar burstiness [15], in which the priorities of different tasks play a critical role. Hence, it inspires us to conjecture that this may be the case for the bursty gene expression patterns as well.

Specifically, a set of genes (maybe the whole genome) have been assigned different expression priorities as per the function of their genetic products: house-keeping genes may be assigned with a high priority due to that house-keeping proteins are essential to the basic functioning of a cell, whereas the genes pertinent to differentiation or

regulation may be of low priorities (see Fig. 2D). Transcriptional resources including RNA polymerases and energy in the form of ATP or alike are finite or even scarce in a cell, and are assumed to only be allocated to one certain gene

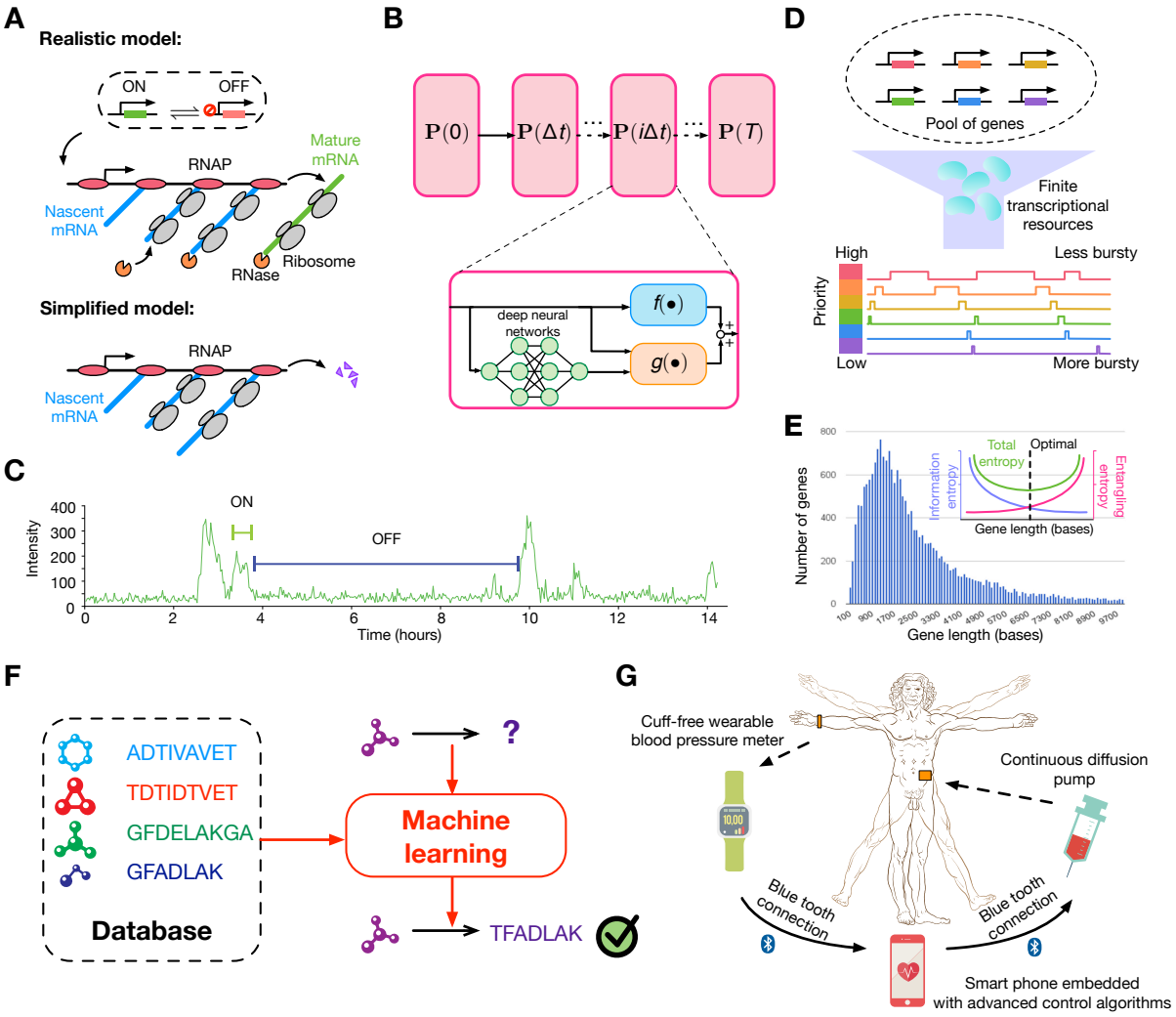


Figure 2: Illustration of proposed ideas. (A) The top shows a model for mRNA kinetics. The promoter stochastically switches between active and inactive states. In the active state, stochastic transcription initiation is followed by mRNA elongation at a certain speed. Once the transcript is complete, mature mRNA is released from the gene into the cytoplasm. The decay of mRNA needs some time to complete as well. The bottom shows a simplified model for mRNA kinetics, wherein only the mRNA initiation and maturation are considered. (B) A sketch of the idea of physics-informed machine learning technique for solving Eq. (3) in the time interval $[0, T]$. The time interval is divided into multiple sections of size Δt, in each of which the state vector $P(i\Delta)$ is assumed constant and propagated into the next state as per Eq. (4). The mapping $\mathbf{P}(t) \mapsto \mathbf{P}(t-\tau|t)$ is modeled by a deep neural network, whose hyperparameters will be determined by minimizing the predicted error against a set of SSA trajectories. (C) Live-cell imaging technique of *TTF1* loci in human breast cancer cells shows that the repressive period lasts remarkably long [13]. (D) It shows a pool of genes with diverse priorities competing for finite transcriptional resources. Owing to limited access to the resources, the genes of lower priorities tends to be expressed in a more bursty way with longer waiting times. (E) Distribution of coding gene lengths of *C. elegans*. The data obtained from WormBase release WS237 suggests the distribution peaks around the length of 1700 bases [14]. The inset shows both the information and entangling entropy are monotonic functions of gene length, indicating the existence of an optimum in total entropy. (F) Machine learning approaches afford the opportunity to rapidly and inexpensively explore the vast protein-sequence spaces *in silico*. (G) Schematics of blood pressure regulator.

at one time. This is indeed the case experimentally [16]. Hence, the resources must be delicately dispensed amid the set of genes to gain maximal cellular benefits, and the priority-driven transcriptional regulation is a plausible approach for such an end. Intuitively, the genes of low priorities have less accessibility to the transcriptional resources and have to wait a long period of time prior to the resources being available, thereby resulting the bursty patterns. We will use the queueing theory and renewal theory to quantitatively describe the priority-driven mechanism and may shed light on the cause of heavy tails of mRNA number distributions [17].

Objective 3: Finding the optimal size of mRNA. The high number of genes yields the smooth histogram of gene length and gene numbers. Such a histogram of *C. elegans* suggests that most of the genes are centered around the length of 1700 bases (see Fig. 2E). Driven by curiosity, it naturally raises the interesting questions: 1) is there any universal law governing such gene-length distributions? 2) where does the number 1700 come from?

Our study intends to seek answers to such questions by an interdisciplinary approach – integrating polymer physics and information theory. From the perspective of information theory, the longer the mRNA is, the more information can be encoded; in other words, the information entropy is lower. On the other hand, the molecules of DNA and mRNA being transcribed can be imagined as two ribbons tied at one end (the polymerase), while the length of DNA ribbon is fixed for different genes. As the mRNA ribbon gets longer, the probability of entanglement occurrence becomes higher. Such entanglement is not favorable to transcription. The polymerase needs to break through the knot to resume transcription, during which may break either the mRNA or DNA calling for the repairing unit. As such, there must be a trade-off between the two adverse effects, and we believe such a trade-off is optimal given millions of years of biological evolution. We will develop a novel theory to answer these questions by integrating polymer physics and information theory, and later verify it by using bioinformatics to build a database of the gene-length distributions for different cell lines.

This ongoing project is expected to achieve within 3-4 year with the production of 10 publications among which 4 or 5 will be submitted to high-profile journals such as *Nat. Commun.*, *Phys. Rev. Lett.*, *Proc. Natl Acad. Sci. USA* and *Nature*.

3. Project II: Machine-learning based biosynthetic process

The terpenoids are a large and diverse class of naturally occurring organic chemicals derived from terpenes. About 60% of known natural products are terpenoids. More importantly, terpenoids are precursors of a multitude of drug of high clinical values, for instance anti-cancer drug paclitaxel. The aim of the project is to find the enzyme synthesizing a particular terpenoid of interest and subsequently engineer strains to improve titers in industrial-scale bioreactors. To achieve the former, we will first build a curated database of terpenoid-enzyme mappings, base on which the unprecedented power of machine learning will be harnessed to explore vast protein-sequence spaces that is beyond the reach of current experimental approaches (see Fig. 2F). The project is ongoing while we are collaborating with the team of Professor Lixin Zhang at ECUST, which has accumulated substantial experience on biosynthetic process.

4. Project III: Biomedical titration control

Biomedical titration is a newly emerging treatment that takes advantage of continuous subcutaneous infusion pump to deliver medication into human body, for further regulating or stabilizing one or more clinical indices. This concept is applicable to a vast set of elementary biomedical systems, for example, dosing insulin to regulate blood glucose, employing antihypertensive medication to control systolic blood pressure, and using tamoxifen to minimize the number of malignant cells in patients with breast cancer. The study on this kind of treatment is highly motivated by the huge clinical and economical value behind. Diabetes alone, cost 101 billion USD in diagnosis and treatment in 2013 across the United States soil, ranking as one of the most expensive condition by CNBC [18]. There are approximately 1.25 million people suffering diabetes in the United States, and this number is showing an increasing trend [19]. The state-of-the-art investigations into this topic are merely limited to blood glucose regulation or artificial pancreas, and results on blood pressure regulation or tumor cell minimization are rarely reported. In the future, I will organize the work into three parts.

4.1 High-performance controller for biomedical titration

Compared with the systems of traditional chemical engineering process, the system of biomedical titration admits a remarkably higher degree of complexity, and suffer enormously more disturbances. Taking blood glucose as an

example, the irregularity of meal intake and exercise and even emotional activities affect the evolution of blood glucose. Additionally, there are more stringent constraints imposed on the control of biomedical titration. For diabetic treatment, the condition that blood glucose level is less than 60 mg/dL is clinically deemed hypoglycemia, which may lead to immediate and severe clinical consequences, such as clumsiness, seizure, coma, and even death. Hence, the dosage of insulin should be delivered with great delicacy. In short summary, it is challenging to design controllers for biomedical titration process due to its strong disturbances and stringent constraints. Model predictive control, which is good at handling constraints and has been widely applied in industry, may be a plausible option for regulating clinical titration indices and easing the pain of patients. Another feature that distinguish biomedical titration process from the others is that all the quantities and variables therein are positive. For instance, the dosage of insulin must be positive, which means we cannot remove insulin from human body. The control theory on this type of systems is usually overlooked and now calls for a systematic study.

4.2 Automatic therapy personalization

The high degree of demographic and genetic heterogeneity leads to tremendous pharmacokinetic/pharmacodynamic differences from person to person. A well-tuned titration controller for one patient may act catastrophically for another, which motivates methods for controller tuning and optimization. Recently, in [20], we presented an elegant framework to personalize the setting adaptively by minimizing a composite clinical risk index merely according to blood glucose measurements. Compared to the best tuning ever, our method is able to significantly reduce the glycemic risk by at most 9.7%. In fact, there exist multiple indices to pursue in clinical practice and there may be conflicts between them. For instance, the percentage time in range of the intervals $[180\ \mathrm{mg/dL}, +\infty)$ and $(0{,}60\ \mathrm{mg/dL}]$ are two different indices that are of interest to be minimized in blood glucose regulation. This calls for methods to achieve a delicate tradeoff among multiple indices, which sets the goal for my group.

4.3 Development of continuously precise wearable measuring device

The major technical obstacle stands in the way towards automatic biomedical titration device is the lack of devices that provide continuous reliable measurements with high precision. For example, the most popular method to measure blood pressure is the cuff method that asks for inflating the cuff to occlude the upper arm artery. Unfortunately, this feature compromises the capability to provide continuous readings, thus possibly leading to unalarmed fatal hyper- or hypotension. The inflation requires the compressor component, which can hardly be miniaturized and degrades the wearability. I intend to devise a cuff-free wristband that delivers continuous blood pressure readings by measuring the pulse transmitting time along artery. A mathematical model will be established to translate the time measurements into blood pressure readings, in the mean time calibrating the effect of arm position with the help of gyro sensors. This development constitutes the last missing piece of the puzzle towards the most desired fully automated blood pressure regulator as shown in Fig. 2G.

5. Project IV: Plant-wide optimization for organo-silicon process

The organo-silicon process is an excellent example of continuous and batch processes integrated together: the upstream processes continuously produce intermediate organo-silicon monomers such as methylchlorosilane, whereas the downstream processes use the monomers to synthesize diverse organo-silicon products in batches. Such end-products are usually customized by orders and of high commercial value. I am working with Elkem Ltd., a world-leading organo-silicon producer, and we aim to use machine learning techniques to optimize the plant-wide process operation to reach an equilibrium where energy efficiency is maximized and the most of orders are delivered timely. Unfortunately, I am not allowed to disclose more details due to business confidentiality.

References

[1] Chen, X. *et al.* Visualizing RNA dynamics in live cells with bright and stable fluorescent RNAs. *Nat. Biotechnol.* **37**, 1287–1293 (2019).

[2] Nojima, T. *et al.* Mammalian NET-seq reveals genome-wide nascent transcription coupled to rna processing. *Cell* **161**, 526–540 (2015).

[3] Cao, Z. & Grima, R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **9**, 1–15 (2018).

[4] Cao, Z. & Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl Acad. Sci. USA* **117**, 4682–4692 (2020).

[5] Elf, J. & Ehrenberg, M. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.* **13**, 2475–2484 (2003).

[6] Ochiai, H., Sugawara, T., Sakuma, T. & Yamamoto, T. Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells. *Sci. Rep.* **4**, 1–9 (2014).

[7] Larsson, A. J. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).

[8] Kumar, N., Platini, T. & Kulkarni, R. V. Exact distributions for stochastic gene expression models with bursting and feedback. *Phys. Rev. Lett.* **113**, 268105 (2014).

[9] Grima, R., Schmidt, D. R. & Newman, T. J. Steady-state fluctuations of a genetic feedback loop: An exact solution. *J. Chem. Phys.* **137**, 035104 (2012).

[10] Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* **79**, 031911 (2009).

[11] Pendar, H., Platini, T. & Kulkarni, R. V. Exact protein distributions for stochastic models of gene expression using partitioning of poisson processes. *Phys. Rev. E* **87**, 042720 (2013).

[12] Bratsun, D., Volfson, D., Tsimring, L. S. & Hasty, J. Delay-induced stochastic oscillations in gene regulation. *Proc. Natl Acad. Sci. USA* **102**, 14593–14598 (2005).

[13] Rodriguez, J. *et al.* Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell* **176**, 213–226 (2019).

[14] Girard, L. R. *et al.* Wormbook: the online review of caenorhabditis elegans biology. *Nucleic Acids Res* **35**, D472–D475 (2007).

[15] Barabasi, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).

[16] Hamperl, S. & Cimprich, K. A. Conflict resolution in the genome: how transcription and replication make it work. *Cell* **167**, 1455–1467 (2016).

[17] Ham, L., Brackston, R. D. & Stumpf, M. P. Extrinsic noise and heavy-tailed laws in gene expression. *Phys. Rev. Lett.* **124**, 108101 (2020).

[18] Ferris, R. Diabetes costing americans more than any other disease. *CNBC report* (2016). http://www.cnbc.com/2016/12/27/diabetes-costing-americans-more-than-any-other-disease.html (Retrieved on: Apr. 15, 2020).

[19] American Diabetes Association. Statistics about diabetes (2016). Http://www.diabetes.org/diabetes-basics/statistics/ (Retrieved on: Apr. 15, 2020).

[20] Cao, Z., Gondhalekar, R., Dassau, E. & Doyle, F. J. Extremum seeking control for personalized zone adaptation in model predictive control for type 1 diabetes. *IEEE Trans. Biomed. Eng.* **65**, 1859–1870 (2017).

Teaching Statement

Zhixing Cao (Edward)

Since the first day of my PhD, throughout eight-year academic life at diverse top-tier research institutes, I was first asked and is later continuously urging myself to uncompromisingly strive for seeking interesting problems either arising from great needs or driven by curiosity, and then solving them from a unique perspective that others may overlook. As a modeler and theorist, I have firm faith that the most fascinating part about mathematical models and theories lies in their elegant simplicity and remarkable predictability – they distill knowledge in the form of simple laws from overwhelming complexities, but are able to cast accurate and even surprising predictions on realities, thereby forming a closed loop. Such an academic spirit (what I named the McGill spirit) was first passed by my supervisor Professor Furong Gao, and may be traced back to his supervisor Professor Musa R. Kamal at McGill. The McGill spirit not only underlies my daily research working but also serves as the central piece of academic characteristics that I would like generations of my academic descendants to have as well. Meanwhile, it also greatly impacts my teaching commitment and philosophy: one of the key priorities as an instructor is to be a bridge that connects theories and realities as well as knowledge and students. Besides, instructors should not play a role of a knowledge reservoir but the drive that keeps knowledge "flow" amid students.

Admittedly, I feel very much privileged that I haven been exposed to the McGill spirit and given the chance to apply it in real teaching at a very early stage. First I worked as a teaching assistant at the Hong Kong University of Science and Technology, and then later as a course organizer at the East China University of Science and Technology, where I actively attended seminars about teaching skills including online lecturing. Such opportunities allow me to accumulate teaching experiences and polish instructing skills. In the teaching practice, I always asked myself to present the fundamental concepts in the clearest possible way, highlight the core ideas and motivations, and emphasize the colorful connections between concepts. To this end, in class, (i) I frequently illustrated the abstract concepts by examples from real world and even from scientific frontiers so that students can picture the concepts themselves to gain a visible understanding. For instance, when I introduced the first-order transfer function (FOTF) models in the class of "Process Dynamics and Control", I correlated the FOTF's capability of buffering noise with the effect of compartmentation on stochastic gene expression in eukaryotic cells, which is a recent finding published in *Cell*. (ii) I firmly believe that practice is of paramount importance in all teaching activities; hence, I would like students to "get their hands dirty". In class, I used to encourage students to interpret the concepts through their own lenses so that they can grasp the concepts easily and efficiently, and also memorize them better. Project is the alternative approach to familiarize them with the course contents. For example, the students in the course "Introduction to Machine Learning for Chemical Engineers" were asked to use machine learning tools to solve problems specifically in the field of chemical engineering and biology. In the course of doing a project, one can connect the "dots" of the course contents, polish programming skills, get their first bites on research and ultimately be primed for doing cutting-edge research in the future. In summary, so as to tightly connect between theory and reality, I shaped my classes to be both example- and practice-oriented.

If given the opportunity at McGill, I am happy to contribute to the rich and diverse chemical and biological engineering curriculum at McGill by teaching both undergraduate and graduate courses on either process control, machine learning or systems biology. In particular, I will be happy to instruct undergraduate courses "Process Dynamics and Control" and "Introduction to Machine Learning for Chemical Engineers" (of which the syllabus is attached) or something alike. For both courses, I am already able to organize a whole series of lectures, design the complete curriculum and prepare exercises for students. Besides, I am also very glad to instruct a graduate course "Practical Systems Biology" to show the students how mathematics and biology come into "chemical" reactions.

The role of a principal investigator (PI) is akin to that of a soccer club manager working on a big picture. Being a marvelous PhD student who can code awesome programs and write perfect papers does not necessarily result in being a successful PI. To be so, one should think smartly and strategically and solve the following set of tasks: (i) one should wisely use the resources at hand including collaborations, fundings, helpful suggestions and innovative ideas to help lab members to reach their full academic potential and maximal research outputs. (ii) One then should make full use of the outputs to convince investors including research council and industrial partners, thereby forming a closed loop to sustainably develop a vibrant and dynamic research team into a world-leading powerhouse of innovation.

Syllabus of "Introduction to Machine Learning for Chemical Engineers"

Zhixing Cao (Edward)

1 Basic Information

Course name	Introduction to Machine Learning for Chemical Engineers
Course type	Core course for Year 3 undergradate
Credit	2
Number of lectures	32 (16 weeks)
Prerequisites	Differential equations, probability, statistics, linear algebra, basic programming skills

2 Course Overview

With the ever increasing amounts of data generating in chemical engineering practice and biological experiments, the need for automated methods for data analysis continues to grow. The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest. The course is designed as a self-contained introduction to machine learning, including deep learning, Bayesian methods, unsupervised learning, etc. The course aims for Year 3 undergraduate students with a suitably mathematical background to provide them with fundamental understanding of machine-learning problem formulations and enhance their capability to interpret data from reality.

3 Course Objectives

The knowledge, abilities and skills that students are expected to gain from this course are:

A) Fundamental theories and key techniques of machine learning, including deep learning, supervised learning, Bayesian models and methods, data preprocessing, model training, hyperparameters optimization and cross validation, as well as the use of novel programming language with high computational efficiency – Julia;

B) Ability to analyze problems independently, particularly those from chemical engineering and biology, and capability of analyzing data thereof, building mathematical models and subsequently concluding;

C) Knowledge about the framework of machine-learning field, ability to read cutting-edge research papers and establishment of a sense of green and sustainable engineering practice.

4 Course Calendar

Teaching methods: a. Lecturing, b. Examples/Case illustrating, c. Off-class reading, d. Discussion.

Topic	Contents	Time (class)	Objectives	Teaching Methods
1. Introduction	a. Basic concepts and an overview of ML	0.5	A	a,b
	b. History and perspectives of ML	0.5	A	a,b,c
2. IO data processing	a. Data sampling, cleaning and outlier handling	1	A	a,b,c
	b. PCA, PLS	1.5	A	a,b,c
3. Algorithms	a. Julia tutorial	1.5	A	a,b,c
	b. Concepts of regression & linear models	1	A	a,b,c
	c. Concepts of classification	1	A	a,b,c
4. Model evaluation	a. Cross validation	1	A	a,b,c,d
	b. Hyperparameter selection	1	A	a,b,c,d
	c. Computational cost estimation	1	A	a,b,c,d
4. Probabilistic model	a. Latent variable models & EM algorithms	1	A,B	a,b,c,d
	b. Graphic models	2	A,B	a,b,c,d
	c. Gaussian process	1	A,B	a,b,c,d
5. Deep learning	a. Deep neural network training & evaluation	1	A,B	a,b,c,d
	b. Convolutional neural network	1	A,B	a,b,c,d
	c. Autoencoders	1	A,B	a,b,c,d
	d. Deep network with stochastic depth	1	A,B	a,b,c,d
6. Case study	a. Process monitoring & fault diagnosis	2	A,B,C	a,b,c,d
	b. Multi-omics data analysis	6	A,B,C	a,b,c,d
	c. Protein design	2	A,B,C	a,b,c,d
7. Case discussion	a. Student presentation	4	A,B,C	b,c,d

5 Recommended Textbooks

- Murphy, K.P. *Machine learning: A probabilistic perspective*. MIT Press, 2012.
- Geron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow:*. O.Reilly, 2017.

Class notes and handouts will be provided as necessary on the web.

6 Grading

The grade of students consists of four parts: mid-term exam, final exam, projects and assignments, whose weights and criteria are summarized in the following tables.

Evaluation Methods	Scores
Mid-term Exam	20
Final Exam	50
Project & presentation	15
3 Assignments	5 each

Evaluation Methods	Criteria	Grade			
		A	B	C	F
Assignments & Exams	Work independence	●●●	●●●	●●○	●○○
	Knowledge awareness	●●●	●●●	●●○	●○○
	Accuracy	●●●	●●○	●●○	●○○
Projects & Presentations	Novelty	●●●	●●●	●●●	●●●
	Techniques	●●●	●●●	●●●	●●●
	Speech	●●●	●●●	●●●	●●●

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05822-0 OPEN

Linear mapping approximation of gene regulatory networks with stochastic dynamics

Zhixing Cao [1] & Ramon Grima[1]

The presence of protein–DNA binding reactions often leads to analytically intractable models of stochastic gene expression. Here we present the linear-mapping approximation that maps systems with protein–promoter interactions onto approximately equivalent systems with no binding reactions. This is achieved by the marriage of conditional mean-field approximation and the Magnus expansion, leading to analytic or semi-analytic expressions for the approximate time-dependent and steady-state protein number distributions. Stochastic simulations verify the method's accuracy in capturing the changes in the protein number distributions with time for a wide variety of networks displaying auto- and mutual-regulation of gene expression and independently of the ratios of the timescales governing the dynamics. The method is also used to study the first-passage time distribution of promoter switching, the sensitivity of the size of protein number fluctuations to parameter perturbation and the stochastic bifurcation diagram characterizing the onset of multimodality in protein number distributions.

[1] School of Biological Sciences, the University of Edinburgh, Mayfield Road, Edinburgh EH9 3JH Scotland, UK. Correspondence and requests for materials should be addressed to R.G. (email: ramon.grima@ed.ac.uk)

Gaining detailed quantitative insight into the dynamics of single living cells is one of the main goals of modern molecular biology. It is well acknowledged that a systems biology approach, whereby alternating cycles of mathematical modeling and experiments lead to refined understanding of the biological system is ideal[1]. In such an approach, the prediction of a mathematical model is contrasted with experimental data: a good match implies that the model offers a potential explanation of the observations (and potentially an estimation of the parameters) while a bad match implies that further refinement of the model (and probably further experiments) is necessary. The output of the experiments is often the number of fluorescently tagged proteins as a function of time from which one can calculate the probability distribution and its associated moments such as the mean and variance in the protein numbers. Clearly then a mathematical model is useful in this systems biology approach to living cell dynamics, if it can accurately predict the distribution of protein numbers and herein lies a problem: exact solutions of the stochastic description of gene regulatory networks (GRNs) have only been reported for a few simple cases. In this article, we describe a novel method which circumvents the aforementioned problem by deriving approximate but accurate solutions to the probability distributions of protein numbers of a wide variety of GRNs.

Before we describe the method, we summarize the state of the art in the mathematical modeling of GRNs. It is well known that such networks suffer from noise principally due to the low copy number of genes, mRNA and of some protein molecules inside single cells[2,3]. Hence, a stochastic mathematical framework is necessary to describe the dynamics of GRNs. The accepted modern-day framework is the Chemical Master Equation (CME), which is a set of differential equations describing the probabilistic evolution of states of the GRN[4]. Exact solutions of the CME have only been reported for a few simple GRNs: (i) the time-dependent solution of the CME of a GRN involving the reversible switching between two promoter states, the production of mRNA by the active state and the degradation of mRNA[5]; (ii) the time-dependent solution of the CME of a GRN involving the transcription of mRNA by an active promoter, the translation of the mRNA into protein and the decay of both protein and mRNA[6]; (iii) the steady-state solution of a GRN of a negative or positive feedback loop, whereby a promoter can produce proteins with a certain rate in the inactive state and with a different rate in the active state and it switches from the inactive to the active state by binding a protein molecule. This model also includes protein degradation[7]; (iv) the same as in (iii) but with the production rate occurring in bursts[8]. In models (i) and (ii), every reaction is either zero or first-order and hence we shall refer to these as linear GRNs since by the law of mass action, the rate of every reaction is linear in the concentrations. We shall refer to models (iii) and (iv) as nonlinear GRNs because there is a second-order reaction involving the binding of protein to the promoter, whose rate is nonlinear in the concentrations. Note that the exact time-dependent solution has only been obtained for linear GRNs; for the nonlinear GRNs only the steady-state solution is known. It is also the case that none of these model an external time-varying stimulus to the GRN, a commonly observed feature, e.g., circadian clocks.

Notwithstanding these difficulties, some have devised methods to obtain expressions for the approximate probability distribution solutions of the CME for nonlinear GRNs under various assumptions: (i) the fluctuations in copy numbers are very small and the distribution is Gaussian[9–11]; (ii) that there exists time-scale separation, e.g. slow promoter switching[12–16]; (iii) the promoter states are uncorrelated[17]; (iv) the volume of the cell is large enough that the CME can be approximated by a few terms in the system-size expansion[18]. All of these methods generate approximate time evolving distributions for GRNs with second-order reactions (for a comprehensive recent review see[19]) and hence circumvent the issues of exact solutions of the CME. The disadvantages of these methods are however considerable because of their limiting assumptions: (i) distributions measured in vivo are often highly skewed and sometimes multimodal, i.e. non-Gaussian; (ii) timescale separation occurs in a few cases but is not generally the case in nature; (iii) promoter states are often correlated due to the presence of feedback loops; (iv) it is impossible to a priori estimate how many terms are needed in the system-size expansion to obtain an accurate result. There are also methods which compute the approximate distribution numerically without explicit analytical expressions (see for e.g.[20–23]); of these the Stochastic Simulation Algorithm (SSA)[20] is of particular importance because the approximation error is equal to the sampling error and hence can be made arbitrarily small.

In this article, we devise a novel type of approximate solution of the CME which provides analytical or semi-analytical expressions for the time-dependent and steady-state solution of common nonlinear GRNs without making a priori assumptions on the form of the distribution or invoking timescale separation and which is even applicable to GRNs with an external time-varying stimulus.

Results

Illustrating the linear mapping approximation by an example. The solution of the CME of linear GRNs is typically easier than the solution of the CME of nonlinear GRNs. This observation leads to the question: is it possible to map, in an approximate way, a nonlinear GRN onto an equivalent linear GRN such that the exact solution of the latter gives an approximate solution of the former?

We shall first develop the method on a simple nonlinear feedback loop which is schematically shown in Fig. 1a (upper). A promoter switches between two states G and G^*, and each state is associated with a different rate of protein production. The switch from G to G^* occurs through the binding of a protein molecule to G and the protein can also decay. This is a rudimentary form of a feedback loop: if $\rho_u > \rho_b$, then it functions as a negative-feedback loop (protein represses its own expression) and otherwise it is a positive feedback loop (protein activates its own expression). Here we do not explicitly model the mRNA for simplicity purposes. This nonlinear GRN can be transformed in to a linear GRN (Fig. 1a lower) by removing the second-order reaction between protein and state G. Specifically, we replace the reversible reaction $G + P \rightleftharpoons G^*$ by $G \rightleftharpoons G^*$. Note that all parameters between the two models are the same except for σ_b and $\bar{\sigma}_b$. The question now is: given the nonlinear GRN with a certain set of parameters, how can we select the free parameter $\bar{\sigma}_b$ in the linear GRN such that the solution of this system well approximates the solution of the original nonlinear GRN?

First we write the exact moment equations for the linear GRN (which can be straightforwardly obtained from the CME—see Methods):

$$\begin{aligned}
\partial_t \langle n_p \rangle &= \rho_u \langle n_g \rangle + \rho_b \left(1 - \langle n_g \rangle\right) - \langle n_p \rangle, \\
\partial_t \langle n_g \rangle &= -\bar{\sigma}_b \langle n_g \rangle + \sigma_u \left(1 - \langle n_g \rangle\right), \\
\partial_t \langle n_p n_g \rangle &= \rho_u \langle n_g \rangle + \sigma_u \langle n_p \rangle - (1 + \bar{\sigma}_b + \sigma_u) \langle n_p n_g \rangle,
\end{aligned} \tag{1}$$

where ∂_t denotes the time derivative, n_p is the number of molecules of protein P, n_g is a Boolean variable taking the value of 1 if the promoter is in state G and the value 0 if it is in state G^*

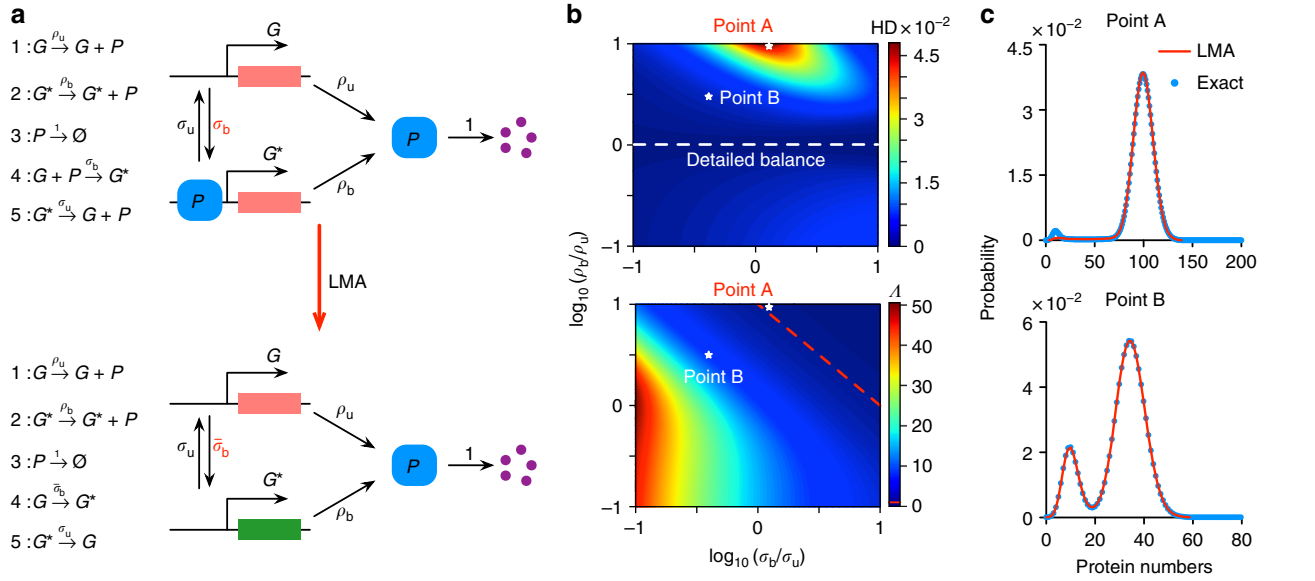


Fig. 1 Linear mapping approximation (LMA) and its application to steady-state conditions. **a** Illustration of the main idea behind the LMA namely to approximate the reversible (nonlinear) reaction between protein and promoter in the nonlinear GRN by a first-order (linear) reaction with an effective reaction rate in a linear GRN. **b** The upper figure shows a heatmap of Hellinger distance (HD) between the LMA and the exact probability distribution of protein numbers in steady-state conditions with parameters $\rho_u = 10$, $\sigma_u = 0.01$ for the nonlinear GRN shown in **a**. The exact distribution is reported in[7]. The bottom figure shows a heatmap of Λ, which is the ratio of the values of the two eigenvalues of the Jacobian of the deterministic rate equations of the nonlinear GRN in steady-state conditions. The red broken line denotes the contour line of $\Lambda = 1$. Note that the value of the HD is very small over a wide range of the ratio of time scales Λ indicating that the LMA's accuracy is independent of time-scale separation. **c** A comparison of the LMA and exact steady-state distributions for Points A and Point B, marked as white stars, on the heatmap in **b**; note that Point A corresponds to the parameter set with the largest HD ($\rho_b = 100$, $\sigma_b = 0.0126$ with HD of 0.0478). Point B corresponds to $\rho_b = 35$, $\sigma_b = 0.004$ with HD of 0.0032

and bracket $\langle\cdot\rangle$ is the expection operator. Next we note that the first-order reaction $G \to G^*$ in the linear GRN maps onto the second-order reaction $G + P \to G^*$ if we select $\bar{\sigma}_b = \sigma_b\left\langle n_p | n_g = 1\right\rangle$ where $n_p \mid n_g = 1$ is the instantaneous number of proteins n_p given the promoter is in state G. The simplest approximation is to use the expectation value of this stochastic rate such that we have:

$$\bar{\sigma}_b = \sigma_b\left\langle n_p | n_g = 1\right\rangle = \sigma_b \frac{\left\langle n_p n_g\right\rangle}{\left\langle n_g\right\rangle}. \quad (2)$$

This is a mean-field assumption and is expected to be accurate when the size of the fluctuations in the number of proteins, given the promoter is in state G, are small compared to the mean number of proteins conditional on the same state. Since experiments show that the standard deviation of the fluctuations divided by the mean molecule number roughly scales as the inverse square root of the mean molecule number[24], it follows that the mean-field assumption should be accurate provided the mean protein molecule numbers in state G are not too small.

Substituting Eq. (2) in Eq. (1) and solving the resultant coupled set of differential equations, we obtain a time-dependent solution for the moments. These solutions can then be substituted in Eq. (2) to obtain our estimate for the effective rate parameter in the linear (mapped) GRN:

$$\bar{\sigma}_b = f\left(t, \rho_u, \rho_b, \sigma_u, \sigma_b\right), \quad (3)$$

where f denotes function of. This function can generally be obtained by numerical solution of the aforementioned modified differential equations; in steady-state conditions an explicit formula can also be obtained:

$$\bar{\sigma}_b = \frac{-1 + \rho_b\sigma_b - \sigma_u + \sqrt{\left(1 - \rho_b\sigma_b + \sigma_u\right)^2 + 4\rho_u\sigma_b(1 + \sigma_u)}}{2}. \quad (4)$$

The last and remaining question is how can we use this parameter estimate to build the full time-dependent solution of the nonlinear GRN. We observe that the time-dependent probability distribution solution of the CME of the linear GRN with general time-dependent $\bar{\sigma}_b$ is likely impossible to obtain in closed-form. However it is possible to solve if $\bar{\sigma}_b$ were a constant independent of time (see "Methods" section); let this general probability distribution solution be denoted as $\mathcal{S}_{\mathrm{FLTD}}(\bar{\sigma}_b, t)$. We shall then make the assumption that the time-dependent probability distribution solution of the CME of the linear GRN with general time-dependent $\bar{\sigma}_b$ given by Eq. (3) is well approximated by $\mathcal{S}_{\mathrm{FLTD}}\left(\bar{\sigma}_b^*, t\right)$ where $\bar{\sigma}_b^*$ is the time-average of Eq. (3):

$$\bar{\sigma}_b^* = \frac{\int_0^t f\left(t', \rho_u, \rho_b, \sigma_u, \sigma_b\right)\mathrm{d}t'}{t}. \quad (5)$$

A rigorous theoretical justification of this assumption can be found in the Methods. Hence the linear mapping approximation (LMA) of the probability distribution of the nonlinear feedback loop at time t is given by $\mathcal{S}_{\mathrm{FLTD}}\left(\bar{\sigma}_b^*, t\right)$. Note that the time-averaging assumption is only needed if one wants to calculate the distribution in finite time; in steady-state, there is no need of the assumption since then $\bar{\sigma}_b$ is constant (and equal to Eq. (4)) and the steady-state probability distribution is directly given by $\mathcal{S}_{\mathrm{FLTD}}(\bar{\sigma}_b, t \to \infty)$.

To summarize, the LMA procedure to find the approximate time-dependent probability distribution of protein numbers at time t in a general nonlinear GRN involves the following steps: (i)

find the linear GRN by replacing any reversible promoter-protein reaction in the nonlinear GRN by a reversible pseudo first-order reaction between promoter states with stochastic rates; (ii) write the closed-set of moment equations for the linear GRN with the stochastic rates replaced by their means, solve for the moments at time t and use the latter to obtain the approximate value of the rate parameter/s at time t characterizing the pseudo first-order reaction/s in the linear GRN; (iii) calculate the time-average of these parameters over the time interval $[0, t]$; (iv) obtain the time-dependent probability distribution solution of the CME of the linear GRN assuming the rate parameter/s characterizing the pseudo first-order reactions are time-independent constants; (v) the approximate time-dependent probability distribution of the nonlinear GRN at time t is then given by replacing the "constant" rate parameter/s characterizing the pseudo first-order reactions solution in step (iv) by the time-averaged parameters calculated in step (iii).

Steps (i) to (iii) can always be performed but steps (iv) and (v) require the existence of a closed-form solution for the linear GRN and this is the major limitation of the method. When such a solution exists, then for a nonlinear GRN with N protein–promoter binding reactions, the approximate time-dependent probability distribution given by the LMA is a closed-form distribution with N effective parameters to be determined numerically. In practice, this leads to a considerable computational advantage over purely numerical methods such as the SSA[20] and the Finite State Projection method[21] (see Supplementary Note 3 for details) simply because the closed-form distribution is composed of well-known functions that can be evaluated by standard symbolic packages in fractions of a second.

If one is only interested to find an approximate steady-state probability distribution of protein numbers then the procedure is considerably simpler. Step (i) is as before. Step (ii) is the same but now the moments are found in steady-state. The final approximate solution is then obtained by substituting the effective rate parameters found in Step (ii) in the steady-state probability distribution solution of the CME of the linear GRN. In many cases, these steps can be done analytically and hence the output is an approximate solution in closed-form.

Note that independent of whether we are interested in the time-dependent or steady-state problem, when a closed-form solution for the linear GRN does not exist, the method still gives approximate expressions for all the moments of the nonlinear GRN using steps (i) and (ii); in this case, its output is similar to moment-closure methods (see ref. [19] for a recent review) but with the advantage that we have made no implicit assumption on the form of the probability distribution solution of the chemical master equation.

The LMA of common nonlinear GRNs. Next we will test the accuracy of this method for various nonlinear GRNs using both exact results and stochastic simulations. In particular, we want to clearly show that the LMA accurately predicts probability distributions for protein numbers which are unimodal or bimodal, Gaussian or skewed, in steady-state or evolving in time and independent of timescale separation.

In Fig. 1, we show the high accuracy of the LMA in predicting the probability distribution of protein numbers for the feedback loop in steady-state conditions. In particular, Fig. 1b (upper) shows a heat map of the Hellinger distance (HD) between the exact steady-state probability distribution of the nonlinear GRN (reported in[7]) and the approximate probability distribution given by the LMA (as described earlier). Note that the HD has the properties of being symmetric and satisfies the triangle inequality, thus implying that it is a distance metric on the space of probability distributions (unlike for example the commonly used Kullback–Leibler divergence). Since it returns a number between 0 and 1, it is clear that the distance between the exact and approximated distributions is very small for both negative ($\rho_b < \rho_u$) and positive feedback ($\rho_b > \rho_u$). This is further confirmed by explicitly showing in Fig. 1c, the distribution for two points in the heat map in Fig. 1b (upper): the LMA distribution with the largest Hellinger distance (Point A) is barely noticeably different from the exact distribution and the LMA does extremely well even when the distribution is bimodal (Point B). It can also be easily proved that the LMA distribution is exact when the system is in detailed balance conditions. This is since in such conditions, $\rho_u = \rho_b$[7], which implies that the protein distribution is unaffected by the bimolecular reaction at the heart of promoter switching and hence the system acts as a linear GRN in this special case. In Fig. 1b (lower), we further confirm the hypothesis that the LMA does well in steady-state conditions independent of the existence of timescale separation conditions: a comparison of the heat plots in Fig. 1b upper and lower shows that while the ratio of the gene and protein timescales (Λ) varies considerably (0.1 to 50) over the region of parameter space considered, there is very little corresponding change in the HD (0 to 4.5×10^{-2}). There is also no correlation between the two heat plots. Λ is the ratio of the two eigenvalues obtained from the Jacobian of the deterministic rate equations. Hence to sum up, in steady-state the LMA predictions for the feedback loop are accurate independent of the type of feedback (positive or negative), modality of the distribution and of timescale separation conditions.

Next we test the accuracy of the LMA for predicting the time-evolution of the probability distribution of proteins in four common types of nonlinear GRNs (or motifs): the feedback loop (Fig. 1a upper), the feedback loop with protein bursting (Fig. 2a), the feedback loop with cooperative protein binding (Fig. 2b) and the feedback loop with oscillatory transcription rates (Fig. 2c). Details of these loops, their master equation formulation and corresponding LMA can be found in Methods and the Supplementary Information. Note that the decay rate of proteins in all cases is set to unity; this is not an arbitrary choice but rather stems from the fact that the time in the master equation can always be non-dimensionalised using the actual value of the protein decay rate k_d. Hence all times shown in the graphs should be understood to be non-dimensional and equal to the real time multiplied by k_d while all other parameters (ρ_u, ρ_b, σ_u, σ_b) should be understood to also be non-dimensional and equal to the real value of the parameter divided by k_d. Bursting (production of proteins in bursts), cooperativity (multiple proteins binding the promoter) and time-varying transcription rates are common features observed in many GRNs in both eukaryotic and prokaryotic cells (see for example ref. [25–27]). Note that an implicit description of mRNA exists in the model with protein bursting because protein burst sizes distributed according to a geometric distribution are obtained when the protein is produced by a fast intermediate mRNA, a common scenario in bacteria and yeast[15]. Note also that while all the three systems are composed of reactions with mass-action propensities, in certain limits they reduce to systems composed of effective reactions with non-mass action propensities e.g. under quasi-equilibrium conditions between promoter and protein, the model of a feedback loop with cooperative binding reduces to an effective model describing protein production with a Hill-type propensity[28,29].

Figure 2d shows that in all cases the LMA distribution agrees very well with that obtained from stochastic simulations using the SSA[20]. In particular the LMA precisely captures the change in shape of the distribution with time from Gaussian at short times to a skewed unimodal distribution at intermediate times to

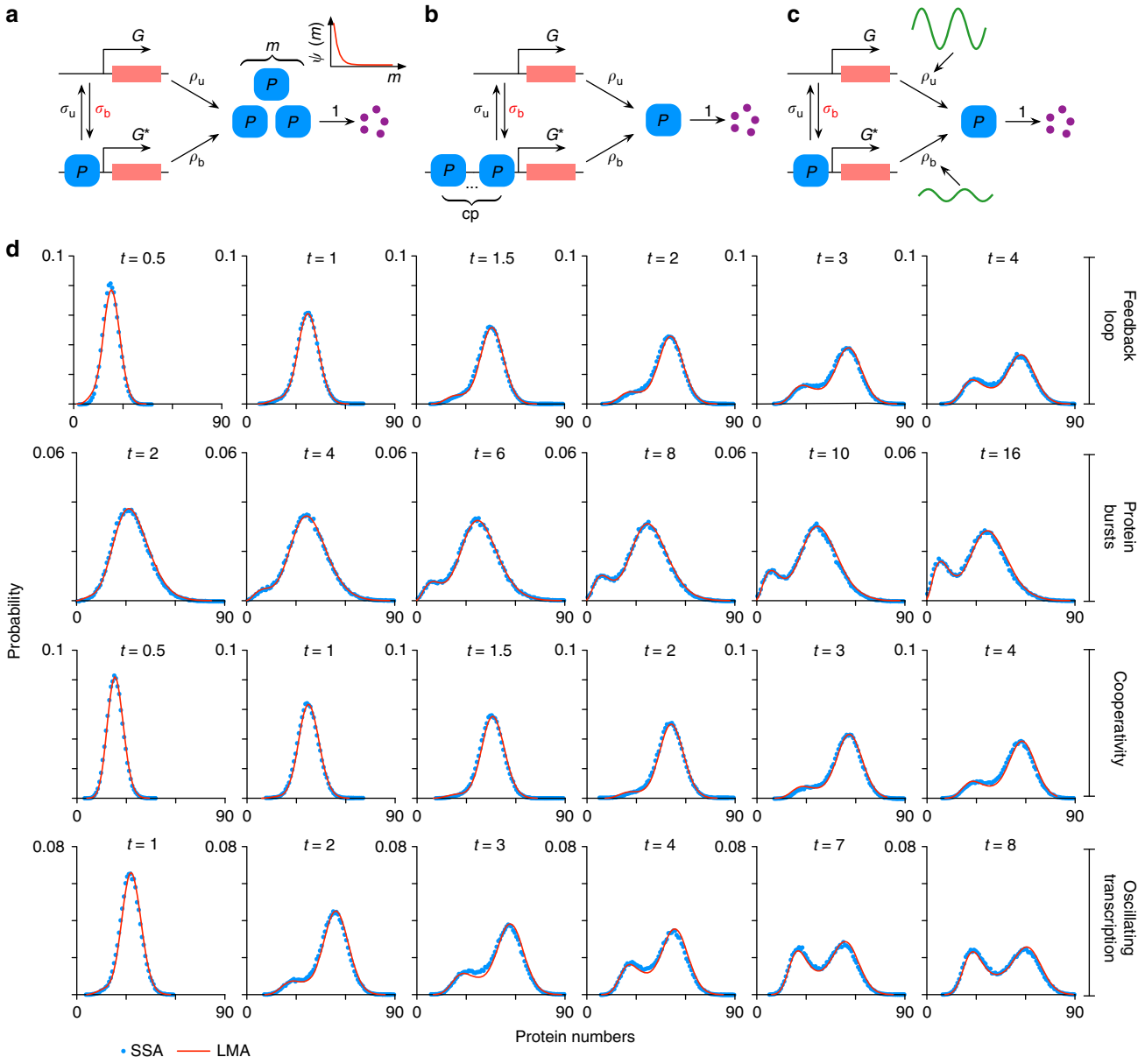


Fig. 2 LMA approximation for the time-dependent probability distribution of protein numbers for various nonlinear GRN. The feedback loop shown in Fig. 1a upper, a feedback loop with protein bursting (**a**), a feedback loop with cooperativity (**b**) and a feedback loop with oscillatory transcription (**c**). The inset of **a** shows the probability distribution $\psi(m)$ of the protein burst size m: we consider a geometric distribution, the discrete analog of the exponential distribution which has been measured in experiments[15, 25]. The number of proteins binding to the promoter for the cooperative loop in **b** is given by cp. In **d**, we show snapshots of the protein number distribution at various times. The LMA approximation (red line) agrees with the results of stochastic simulations using the SSA (blue dots), and is able to capture the transition from unimodality at short times to bimodality at long times. The parameters are: for feedback loop $\rho_u = 60$, $\rho_b = 25$, $\sigma_b = 0.004$, $\sigma_u = 0.25$; for protein bursts $\rho_u = 20$, $\rho_b = 5$, $\sigma_b = 10^{-3}$, $\sigma_u = 0.1$ and the mean burst size $b = 2$; for cooperativity $\rho_u = 60$, $\rho_b = 25$, $\sigma_b = 5 \times 10^{-5}$, $\sigma_u = 0.25$ and cooperativity order cp = 2; for oscillating transcription, the parameters are $\rho_u = 60$, $\rho_b = 25$, Am = 0.3, $k = 1.33$, $\sigma_b = 0.004$, $\sigma_u = 0.25$. The SSA result in each snapshot is obtained by averaging over 80,000 realizations and the sampling error is <2% for each point in the SSA probability distribution. In all cases, the initial conditions are zero protein in promoter state G

bimodal at long times. Furthermore for the oscillatory transcription feedback loop, it can be shown that the LMA correctly captures the oscillatory nature of the mean and variance in protein numbers and accurately predicts the phase difference between the oscillations in the mean protein numbers and in the transcription rate (see Supplementary Fig. 1).

Next, we seek to understand the dependence of the error in the LMA predictions with parameter values and the intuitive reasons underlying such relationships. In Fig. 3a, we show the HD (between the SSA calculated distribution and the LMA distribution) as a function of time for the feedback loop with cooperative protein binding with parameters $\rho_u = 60$, $\rho_b = 25$ and for various values of σ_b. While there is no apparent relationship between HD and σ_b in steady-state, it is clear that the maximum of the HD (over time) increases with σ_b. The corresponding probability distributions for these three maxima (A, B and C) are shown in Fig. 3c upper. The degree of nonlinearity in the GRN is controlled by the rate σ_b of the only nonlinear (second-order) reaction in the nonlinear GRN and hence one would expect our approximate linear mapping to be less accurate as σ_b increases

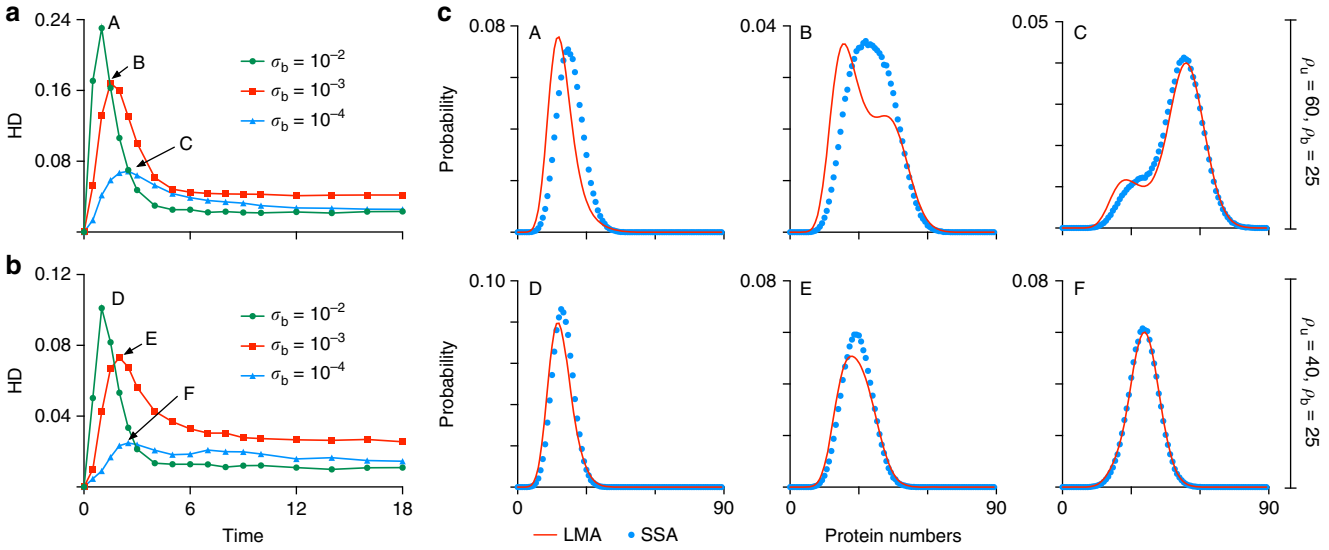


Fig. 3 Dependence of the LMA approximation error on parameters in non-oscillatory feedback loops. **a**, **b** Show the variation of the Hellinger distance (HD) between LMA and SSA distributions for protein numbers in the cooperative feedback loop as a function of time, σ_b and $\rho_u - \rho_b$. The results show that the HD reaches its maximum at intermediate times and that this value is an increasing function of σ_b (the rate parameter controlling the degree of nonlinearity in the nonlinear GRN) and of the difference in gene expression between the two promoter states ($\rho_u - \rho_b$). The other parameters are: cp = 2, $\sigma_u = 0.25$. **c** Compares the LMA predictions and SSA distributions for six time points (at which the HD maximizes) as indicated in **a**, **b**

(for a more precise explanation see the subsection on the justification of the time-averaging assumption in the "Methods" section). In Fig. 3b, we repeat the same analysis but now using parameters $\rho_u = 40$, $\rho_b = 25$. The same relationship is seen between the maximum of the HD (over time) and σ_b however the absolute values of the error are now reduced by about half. This can be explained due to the fact that the difference between ρ_u and ρ_b is smaller in this case than for the one shown in Fig. 3a and we already know that as ρ_u approaches ρ_b, the bimolecular reaction behind promoter switching becomes irrelevant and the LMA becomes exact. We also note that in all cases, the maximum HD was obtained at intermediate times (rather than in steady-state). This is since the HD must be zero initially since we start with the same initial conditions in the LMA and SSA, it must be significant in finite time because both assumptions (mean-field and time averaging) are being used while it must be small in steady-state because only the mean-field assumption is being used. These results are typical of the three nonlinear GRN studied (feedback loop, feedback loop with bursting and feedback loop with cooperativity) which all possess non-zero, non-oscillating moments of protein molecule numbers at steady-state. In summary, the error in the LMA prediction is typically smaller in steady-state compared to time-evolution and it achieves a maximum whose value increases with the rate parameter controlling the nonlinear protein–promoter binding reaction and with the difference between the protein production rates of the two promoter states.

We studied also the error in the LMA predictions for the feedback loop with oscillatory transcription which leads to oscillating moments in the protein numbers in steady-state and hence is in a different class than the previous three nonlinear GRNs. One can think of this nonlinear GRN as an oscillating input signal passing through a filter (composed by the interacting molecular components) with output given by the mean protein numbers. Depending on the type of filter, one would expect certain frequencies will be more attenuated than others. Indeed this is what is observed in a plot (Fig. 4) of the amplitude in the oscillations in the mean protein numbers as a function of the frequency of the oscillating transcription (the input frequency). The SSA predicts the amplitude to gently peak at a frequency of about 0.2; the rate equations predict the same albeit with a different protein amplitude. The LMA however does not capture the peak and simply predicts a decreasing amplitude with increasing frequency which agrees very well with the SSA for frequencies larger than that of the peak. In fact it can be proved (see Supplementary Note 2 Eq. (12)) that independent of parameter values and of the amplitude and frequency of the oscillatory transcription, the LMA predicts the amplitude of the mean protein oscillations to decrease monotonically with increasing frequency. This behavior of the LMA is due to its time-averaging assumption: the amplitude of the time-average of a sinusoidal function is inversely proportional to the frequency. In summary the accuracy of the LMA's predictions is likely low for input frequencies close to the intrinsic resonant frequency of a general nonlinear GRN and high otherwise.

The master equations studied thus far have been limited to GRNs with two promoter states, reactions with mass-action propensities and no explicit description of mRNA. While an implicit description of mRNA exists in the feedback loop with protein bursting model, an explicit description has the advantage that it gives information about both mRNA and protein and can hence can be useful to interpret experiments producing such type of data (see for example ref. [30]). Also we earlier mentioned that an implicit description of effective non-mass action propensities of the Hill-type exists in the model of a feedback loop with cooperativity; an explicit description in the sense of using directly Hill-type propensities in the master equation can sometimes be helpful when we want to work with a reduced model in terms of few parameters. In Figs. 5 and 6, we show the application of the LMA to master equations describing systems with more than two promoter states, effective reactions with non mass-action propensities and including mRNA dynamics. Specifically we find that the LMA accurately captures: (1) the time-evolution of the protein distributions for the 4 promoter state toggle switch (Fig. 5a, b), which involves the expression and mutual repression of two different proteins P and M; (2) the steady-state protein distribution in a two promoter feedback loop where the protein decays via the (non-mass action) Michaelis–Menten like

propensity function (Fig. 5c, d); (3) the mRNA steady-state distribution in a two promoter feedback loop which models both mRNA transcription and protein translation (Fig. 6a, b). Details of the LMA for these three systems can be found in the Supplementary Notes 5–7.

Note that for the two promoter feedback loop modeling transcription and translation, the mRNA distribution can also be computed in time; however, the protein distribution cannot currently be obtained from the LMA in time or steady-state. The reason is that the LMA maps this GRN on to a linear network (also called three-stage gene expression in ref. [15]), for which there is an exact analytical solution for the marginal distribution of mRNA numbers but no solution is currently known for the marginal distribution of protein numbers. However, note that nevertheless the LMA does give all the moments of the protein distribution and these are shown in Fig. 6c to be very accurate compared to those obtained from the SSA, independent of the ratio of the timescales of protein and mRNA—this is particularly relevant to the description of mammalian gene expression[31] where the ratio of timescales varies widely. Analytical solutions for the linear network are known for the case of timescale separation of protein and mRNA lifetimes[15] (conditions compatible with gene expression in bacteria and yeast) and thus in this case by use of the LMA, one can obtain the corresponding analytical solutions for both the mRNA and protein marginal distributions for the feedback loop shown in Fig. 6a.

Further applications of the LMA. Having verified the high accuracy of the LMA, we shall next use it to shed light on how the stochastic properties of a feedback loop are affected by cooperativity and protein bursting. In particular we are interested in how these two features affect: the first-passage time distribution of switching from one promoter state to the other, the sensitivity of the coefficient of variation squared to a change in the parameter values and the stochastic bifurcation diagram.

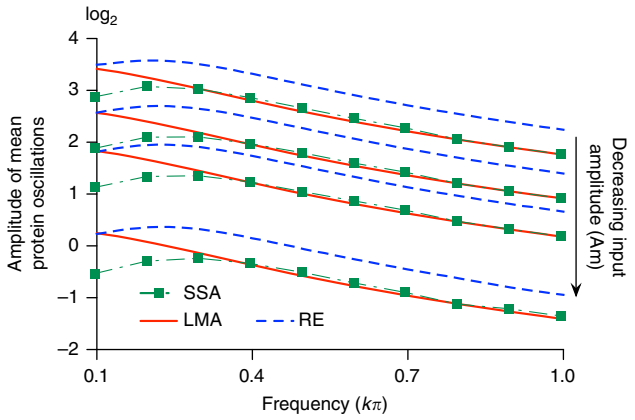


Fig. 4 Dependence of the LMA approximation error on parameters in the oscillatory feedback loop. The figure shows a plot of the amplitude of the oscillations in mean protein number as a function of the frequency of the oscillatory transcription ($k\pi$) and of its amplitude (Am). This is obtained using the LMA (red solid line), the SSA (green squares) and the deterministic rate equations (RE, dashed blue line). The results shows that the rate equations cannot predict the amplitude precisely but capture the weak resonance phenomenon while the LMA gives precise predictions at high frequencies, but cannot capture the resonance. The frequency selectivity of the LMA is due to its time-averaging assumption. The oscillatory feedback loop is the one shown in Fig. 2c with $\rho_u = 20$, $\rho_b = 0$, $\sigma_b = 0.04$, $\sigma_u = 0.25$, and Am is selected to be 0.9, 0.5, 0.3, 0.1 (from top to bottom). Note that the transcription rate in the active promoter state is $\rho_u[1 + \text{Am}\cos(k\pi t)]$

Practically, all GRNs involve multiple promoter states with different post-translational pathways enabled by each state. Hence the switching from one state to another is important to understand from the perspective of cellular decision-making, e.g., a cell's response to a stimulus may require the quick switching on of certain biochemical machinery. This can be mathematically characterized using the first-passage time (FPT) distribution which is the probability distribution of the time it takes to switch between two promoter states given initially one of the states. The switch from G^* to G occurs via $G^* \to G + P$, which is a linear reaction with rate σ_u and hence it can be easily shown that the FPT for the promoter switching from state G^* to G is simply exponential distribution with mean σ_u^{-1}. Hence, cooperativity and bursting have no effect on this switch. The switch from G to G^* occurs via $G + P \to G^*$, which is a nonlinear reaction with rate σ_b; in this case it is much more difficult to obtain the FPT because the process is nonlinear (most FPT theory is for linear reactions though there are exceptions[32]) and since there is a dependence on the instantaneous protein number that is affected by many different processes (transcription, degradation, bursting, cooperativity, etc). However, the LMA maps the above nonlinear reaction to a linear one, and thus enables us to obtain an approximate non-exponential expression for the FPT of switching from state G to G^* (given no protein initially in state G) for nonlinear GRNs (see "Methods" section). In Fig. 7a, we show the LMA's estimate of the FPT distribution for the feedback loop (Fig. 1a upper); the feedback loop with cooperativity, specifically two protein molecules binding the promoter in state G (Fig. 2b) and the feedback loop with protein bursting and a mean burst size of two protein molecules (Fig. 2a) (the four parameters which are common to all three GRNs are fixed for comparison purposes; see Fig. 7 caption for details). The estimates are close to the FPT calculated using the SSA thus verifying the LMA's accuracy. The mean time to switch from G to G^* in a feedback loop is decreased considerably by cooperativity and slightly by bursting; this was observed for all parameter sets, which we studied. We also found out using the LMA that over a large region of parameter space, the mean first-passage time τ is approximately described by a simple power law in two parameters (Fig. 7b):

$$\tau \propto \sigma_b^{-3/5}\rho_u^{-4/5}, \quad \text{feedback loop with and without bursting} \qquad (6)$$

$$\tau \propto \sigma_b^{-1/3}\rho_u^{-4/5} \quad \text{feedback loop with cooperativity} \qquad (7)$$

For the case of the feedback loop with no cooperativity, it is possible to derive an exact solution for the FPT of switching from state G to G^* in steady-state conditions and hence this provides another means to evaluate the accuracy of the LMA (see Supplementary Note 4 for details). This comparison is shown in Fig. 7c, where we show that the error in the LMA's estimate of the FPT distribution (measured by the HD) increases with σ_b (the rate parameter controlling the degree of nonlinearity in the GRN), in agreement with our previous error analysis for time-dependent protein distributions. Nevertheless, the high accuracy of the LMA for predicting first-passage time distributions is visually discernible in all cases.

Next we turn our attention to the sensitivity of nonlinear GRNs to noise. The coefficient of variation of protein number fluctuations defined as the ratio of the standard deviation of the fluctuations and the mean protein numbers is a common measure of the size of intrinsic noise. It is often the case that noise needs to be controlled such that the smooth performance of a certain cellular function is guaranteed[33]. The question then is: which parameter tweaking leads to the largest and smallest changes in

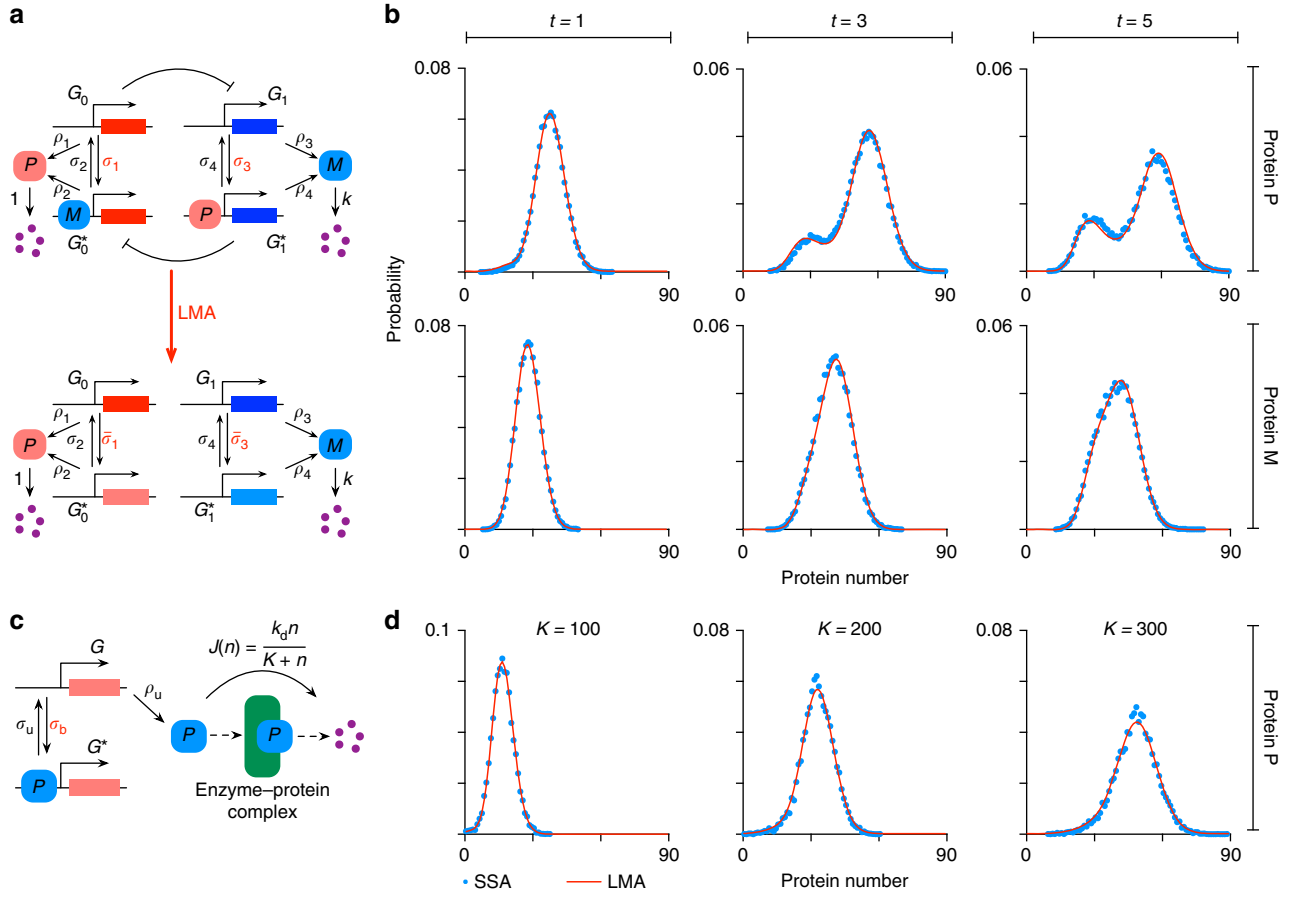


Fig. 5 LMA for the toggle switch and for a feedback loop with nonlinear protein degradation. **a** (upper panel) Illustrates a toggle switch (with two active and two inactive promoters), whereby two proteins *P* and *M* are expressed and mutually repress each other. (lower panel) illustrates how the toggle switch system is decoupled by the LMA into two independent linear networks. **b** Shows the LMA predictions for time-evolution of the marginal distributions of the two proteins, *P* and *M*, in the toggle switch for three different time points. The parameters are: $\rho_1 = 60$, $\rho_2 = 25$, $\rho_3 = 45$, $\rho_4 = 30$, $\sigma_1 = \sigma_3 = 0.004$, $\sigma_2 = \sigma_4 = 0.25$ and $k = 1$. **c** Shows the feedback loop with nonlinear protein degradation via an effective Michaelis–Menten reaction. **d** Shows the LMA predictions for steady-state protein distributions of the system shown in **c** for three different values of the Michaelin–Menten constant *K*. The other parameters are: $\rho_u = 3$, $\rho_b = 0$, $\sigma_b = 4 \times 10^{-4}$, $\sigma_u = 0.25$ and $k_d = 20$

the coefficient of variation squared in steady-state conditions? Whilst this is computationally very expensive to answer using the SSA over large areas of parameter space, with the LMA it can be addressed straightforwardly. We used the LMA to compute the logarithmic sensitivity[34] of the coefficient of variation squared to the four rate parameters common to all three non-oscillatory loops (namely ρ_u, ρ_b, σ_b and σ_u) over a large swath of parameter space. The results are summarized in pie chart form in Fig. 8. For all feedback loops, independent of whether the protein was repressing or activating its own production, the most sensitive parameter was in the vast majority of cases ρ_b while the least sensitive parameter was one of the other three parameters (typically was either σ_u or σ_b). Hence in summary, control of the size of the protein fluctuations can be most efficiently obtained by tweaking gene expression in state G^*.

Finally, we study differences in the stochastic bifurcation diagrams of the three types of feedback loop. The LMA reveals that, for some parameter values, the system has a unimodal steady-state distribution, whereas for other values it has a bimodal distribution, i.e., the noise causes the system to switch between two distinct states. This phenomenon is referred to as noise-induced bistability (NIB) since the deterministic rate equations do not show bistability[35–37]. We explored how the region in parameter space where NIB is observed depends on cooperativity, protein bursting, the type of loop (positive or negative feedback) and the existence of timescale separation. The results are summarized in Fig. 9. Each of the subfigures in Fig. 9 is generated by calculating the steady-state protein distribution over parameter space: the white then indicates a unimodal distribution while a shade of red indicates a bimodal distribution. The three shades of red indicate three different parameter sets as described in the figure caption where a lighter shade of red indicates a larger distance from detailed balance conditions. The calculation is done using the LMA and direct numerical integration of the master equation (as in ref. [7]) and differences between the two are shown in black. For negative feedback, the black regions show where numerical integration predicts unimodality, whereas the LMA (incorrectly) predicts bimodality whereas for positive feedback, the black regions show the opposite situation. In all cases, the black regions are very small thus showing the accuracy of the LMA in capturing NIB. The plots comparing positive (activator) and negative (repressor) feedback show a slight increase in the region of space where there is NIB when there is positive feedback. A comparison of the three shades of red shows that the major factor determining NIB is not the feedback type but rather the difference between the rates of protein production in the two promoter states, i.e., the distance from detailed balance. The larger the difference between the two rates, the larger is the region

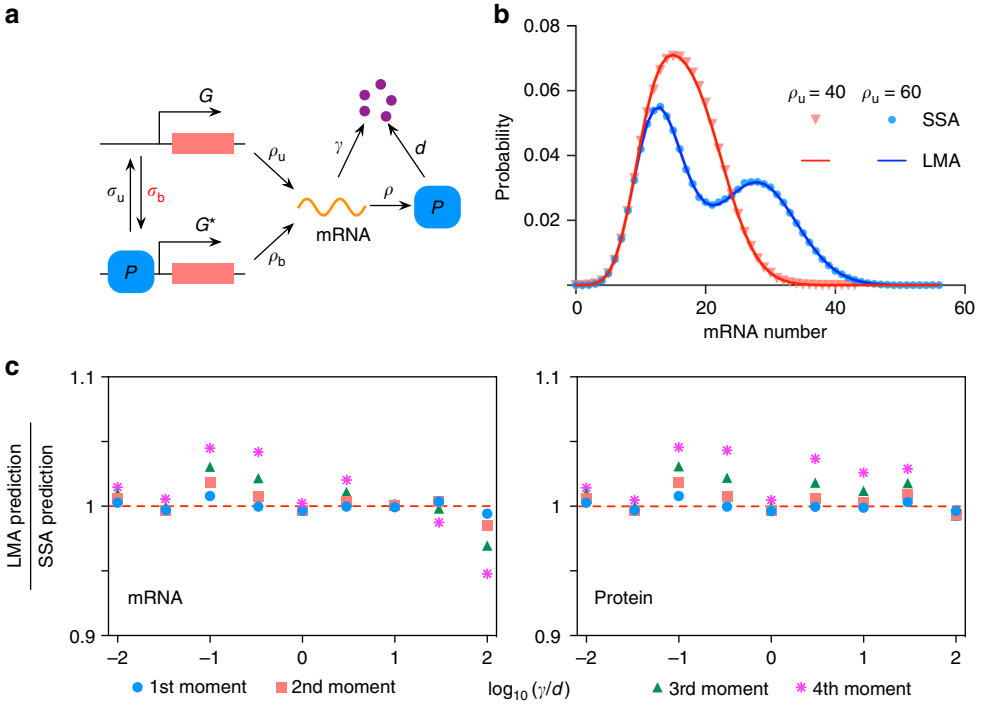


Fig. 6 LMA for the feedback loop with explicit modeling of both mRNA transcription and protein translation. **a** Illustrates the nonlinear feedback loop. **b** Shows the LMA and SSA predictions for steady-state mRNA distribution (See Supplementary Note 7), showing that the LMA is able to capture the change in modality induced by a change in ρ_u. The rest of the parameters are: $\rho_b = 25$, $\sigma_b = 0.004$, $\sigma_u = 0.25$, $\rho = 3$, $\gamma = 2$ and $d = 1$. **c** Compares LMA and SSA predictions for the first to fourth moments (about zero) of both mRNA and protein numbers over a wide range of the mRNA-protein degradation ratio γ/d. The result indicates that the accuracy of the LMA is independent of time-scale separation assumptions. The parameters used are: $\rho_u = 5$, $\rho_b = 0$, $\rho = 100$, $\sigma_u = 1$, $\sigma_b = 1.5 \times 10^{-5}$. The nine pairs of ($\gamma$, d) are: (0.02, 2), (0.02, 0.6), (0.05, 0.5), (0.05, 0.15), (0.2, 0.2), (0.15, 0.05), (0.2, 0.02), (0.15, 0.005) and (0.2, 0.002), constituting a logarithmic span from −2 to 2. The ratio γ/d varies over a range consistent with mammalian gene expression[31]

of parameter space where NIB is observed; for example for the negative feedback loop, the fraction of parameter space where NIB occurs is 59%, 36% and 5% for $(\rho_u, \rho_b) = (60, 25)$, (50, 25) and (40, 25), respectively (see Supplementary Table 1 for data of all cases shown in Fig. 9). We found that both cooperativity and bursting significantly reduce the size of this space and thus have an adverse effect on NIB. In Fig. 9b, we repeat the same exploration as in Fig. 9a, but using a method in the literature that assumes slow promoter switching, i.e., the timescale of promoter switching is much larger than the timescale of protein dynamics[12]. A comparison of Fig. 9a, b shows that the assumption of timescale separation tends to significantly overestimate the size of parameter space where NIB exists, though capturing some of the major observed trends. The comparison also confirms that the LMA is free of underlying assumptions of timescale separation.

Discussion

In this paper, we have introduced a new modeling framework, the LMA, based on a mapping of a nonlinear gene regulatory network to an approximately equivalent linear network. The approach rests on the following two assumptions: (i) a mean-field assumption and (ii) a time-averaging assumption. Specifically these two assumptions are needed to calculate the approximate time-dependent probability distribution solution of protein fluctuations but if one is interested in steady-state then only the first assumption is needed. The mean-field assumption essentially equates with assuming small protein fluctuations compared to the mean number of proteins when the promoter is unbound, a reasonable assumption given that protein numbers are typically much larger than one. The time-averaging assumption implies that the probability distribution at time T of a linear network with a time-dependent parameter $\alpha(t)$ is approximately given by solving the master equation assuming the parameter is a time-independent constant to obtain the solution at time T and subsequently replacing the parameter (in the latter solution) by the time-averaged value of $\alpha(t)$ over the period $[0, T]$. This approximation was shown to correspond to the first term of the Magnus expansion of the time-dependent solution of the master equation.

We have verified that the LMA gives accurate probability distributions (compared to stochastic simulations and to direct numerical integration of the master equation) for feedback loops with and without cooperativity / bursting including those with time-dependent transcription. The accuracy was high independent of the type of feedback (positive or negative), of the modality of the distribution (unimodal or bimodal) and of the existence or lack of timescale separation. We found the accuracy of the LMA to be very high for short and long times and good at intermediate times. The likely reason is that to predict steady-state distributions the method needs only one assumption—the mean-field assumption whereas it needs in addition the time-averaging assumption for predictions in finite time (for short times the accuracy is necessarily high because of deterministic initial conditions). In all cases, the LMA well captures the changes in the shape of the distribution as a function of time, in particular, the transition from unimodal to bimodal behavior. The differences between the predicted and exact distribution are found to grow with the rate parameter controlling the nonlinear protein–promoter binding reaction and with the difference between the protein production rates of the two promoter states; however, these differences are typically barely noticeable to the

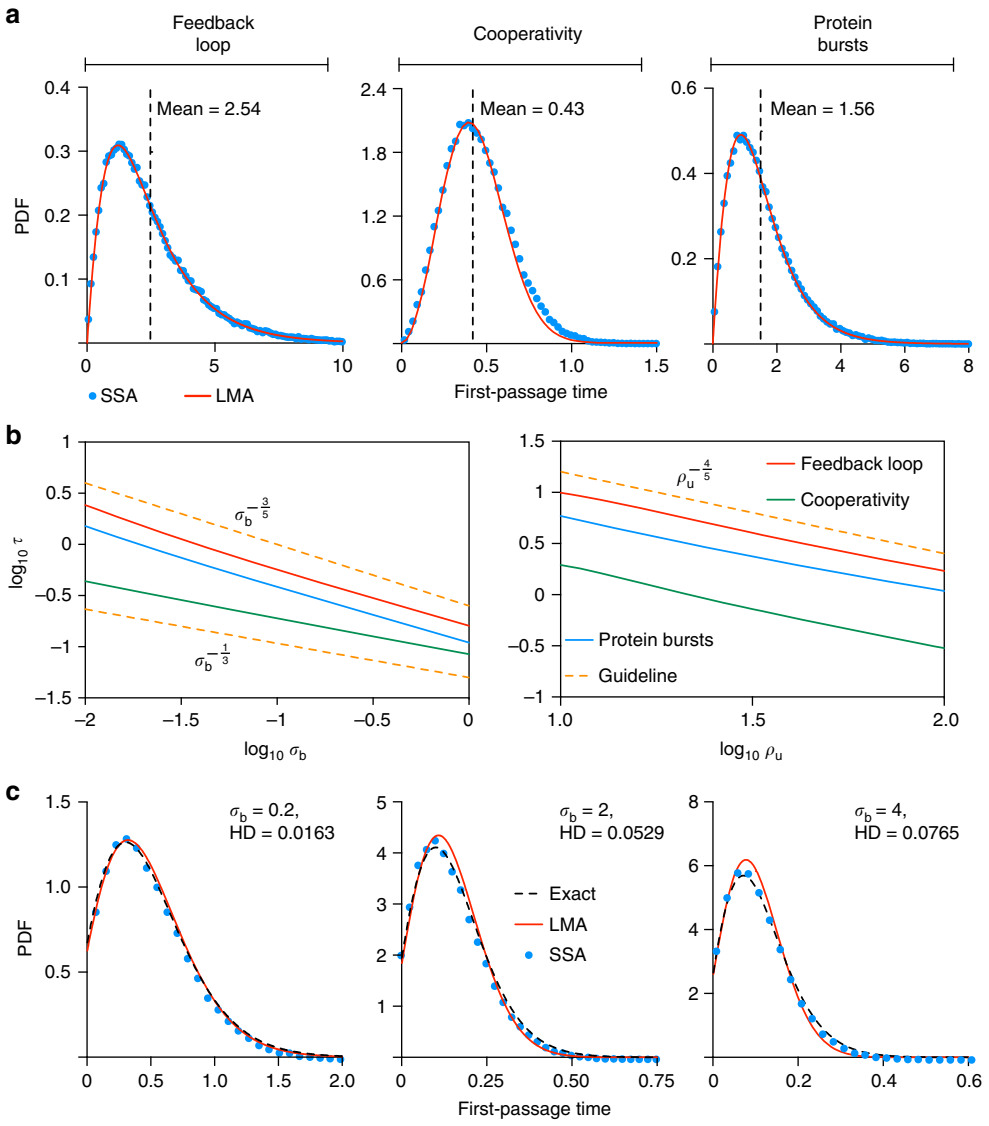


Fig. 7 First-passage time (FPT) analysis of the switching from promoter state G to G^*. **a** Compares the LMA prediction for the FPT probability density function (PDF) with zero proteins in state G initially (red solid line) with that obtained using the SSA (blue dots) for the three non-oscillatory feedback loops. Cooperativity and protein bursts generally reduce the mean FPT. The parameter values are: $\rho_u = 60$, $\sigma_b = 0.01$, the cooperativity order cp = 2 and the mean protein bursts size $b = 2$. Note that there is no dependence of the FPT distribution on σ_u and ρ_b since the first-passage time process stops when the state changes to G^*. **b** Shows that the mean FPT (τ) is a power law in two parameters σ_b and ρ_u. The solid lines are the LMA predictions, while the orange dashed lines are guidelines indicating the power law. **c** Compares the LMA prediction for the FPT distribution in steady-state conditions (red solid line) with that obtained using the SSA (blue dots) and with the exact solution (Eq. (18) in Supplementary Note 4) for the feedback loop with no cooperativity as a function of σ_b

naked eye, except for some intermediate times. We also found that for nonlinear GRNs with an external input oscillatory signal, the LMA's predictions are accurate for input frequencies far from the intrinsic resonant frequency of the GRN itself; this is due to the filtering properties of the time-averaging assumption. We also used the LMA to study how cooperativity and protein bursting affect the first-passage time distribution governing promoter switching, the sensitivity of the coefficient of variation squared to parameter perturbation and the stochastic bifurcation diagram. The extensive study over large swaths of parameter space was made possible by the fact that the LMA provides closed-form solutions for the protein distributions. This is a distinct computational advantage over the stochastic simulation algorithm and also over the finite-state projection algorithm (see Supplementary Note 3 and Supplementary Fig. 2 for details of the comparison of CPU time of the various algorithms).

The LMA, of course, cannot possibly solve the master equations of all gene regulatory networks encountered in nature. In particular when the nonlinear GRN has also bimolecular reactions that are not of the protein–promoter type, the LMA mapping does not lead to a linear GRN though it is still a simpler GRN than the original one. In such a case, it is typically difficult to solve exactly the master equation of the simplified GRN. Likely, progress can then be made by replacing the bimolecular reactions (not involving protein and promoter) by an effective first-order reaction/s such that one has again an effective linear GRN. For example, for GRNs, for which the protein is catalyzed by an enzyme, the catalysis can be effectively modeled by a first-order decay reaction for the protein with a Michaelis–Menten rate (as shown in one of our examples). This additional linearization might not always be possible or else even if possible it might still lead to unsolvable or very difficult to solve master equations; this

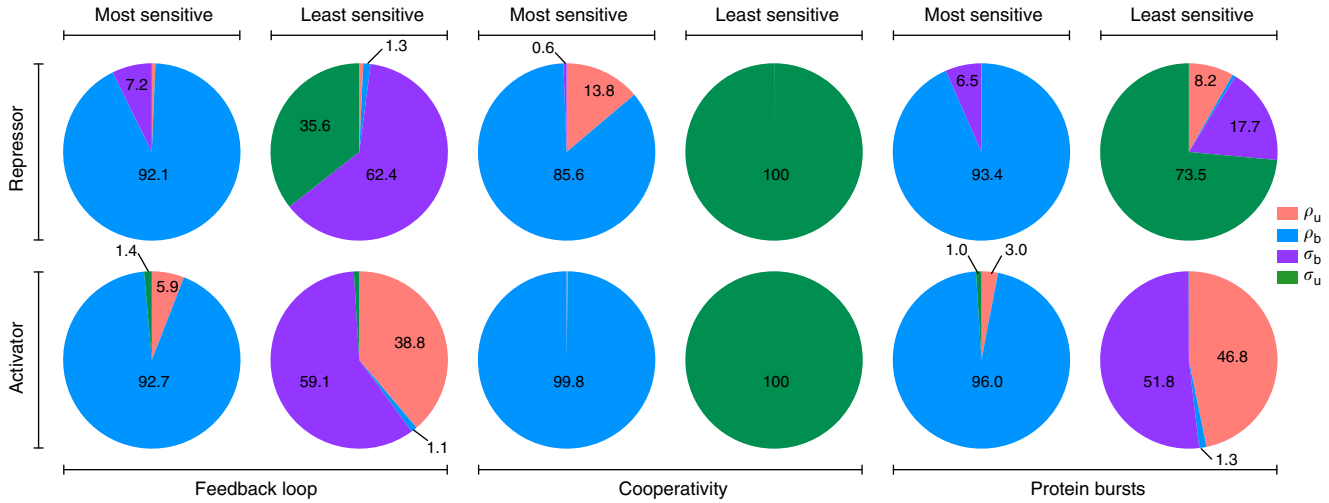


Fig. 8 Sensitivity of the coefficient of variation squared to parameter perturbation in steady-state conditions. The pie charts show the most and least sensitive parameters for the three types of non-oscillatory feedback loops, in activating (positive feedback) or repressing mode (negative feedback). Out of the four variables, ρ_b is the most sensitive parameter and occupies almost 90% in each of the six cases, whereas the least sensitive parameter is typically either σ_b or σ_u. The activator stands for $\rho_b > \rho_u$, whereas the repressor means $\rho_u > \rho_b$. The sensitivity results are obtained by using the LMA to calculate the logarithmic sensitivity of the coefficient of variation squared to the four parameters (σ_b, σ_u, ρ_b and ρ_u) on a regular lattice over the space: ρ_b, $\rho_u \in [1, 10^2]$ and σ_b, $\sigma_u \in [10^{-2}, 1]$. The lattice spacing is 5 for ρ_b, ρ_u and 0.05 for σ_b, σ_u. The cooperativity order was cp = 2 for the feedback loop with cooperativity and the mean protein bursts size was $b = 2$ for the feedback loop with bursting

has to be ascertained on a case-by-case basis and no general statements can be made in this regard.

We finish by noting that the LMA has significant advantages over current methods in the literature. Unlike the linear-noise approximation, it does not assume the distribution is Gaussian and that the means are well described by the deterministic rate equations. It does not assume timescale separation, a common assumption in the literature. It is also superior to moment-closure methods[38–40] since there is no underlying assumption of a distribution of any kind and also since it does not just give the moments but also the distribution itself (a detailed study of the accuracy of the moments provided by the LMA and comparison with common moment-closure methods is under investigation). The LMA also provides distributions in analytical or semi-analytical form for all times, a clear advantage over methods based on distribution reconstruction using maximum entropy[23] or finite-state projection[21]. Hence concluding, the newly devised LMA method provides a new tool for the systematic exploration of the stochastic properties of nonlinear GRNs in systems and synthetic biology.

Methods

Master and moment equations. Consider a chemical reaction network involving N distinct species interacting with each other in a well-stirred volume Ω via a set of R reactions $\sum_{i=1}^{N} s_{ir}X_i \xrightarrow{k_r} \sum_{i=1}^{N} p_{ir}X_i$, where X_i stands for species i ($i = 1, 2, \ldots, N$), $r = 1, 2, \ldots, R$ and the stoichiometric coefficients s_{ir} and p_{ir} are nonnegative integers specifying, the molecule numbers of reactants and products involved in reaction r, respectively. k_r is the rate constant of reaction r. The associated CME can be written as:

$$\partial_t P(\mathbf{n}, t) = \sum_{r=1}^{R} f_r(\mathbf{n} - \mathbf{S}_r)P(\mathbf{n} - \mathbf{S}_r, t) - \sum_{r=1}^{R} f_r(\mathbf{n})P(\mathbf{n}, t), \quad (8)$$

where $\mathbf{n} = [n_1, n_2, \ldots, n_N]^\top$ is the state vector of species molecule numbers, $P(\mathbf{n}, t)$ is the probability of the system being in state $\mathbf{n}$ at time t[4]. The i-th entry of the vector $\mathbf{S}_r$ is given by $p_{ir} - s_{ir}$, and $f_r(\mathbf{n})$ is the propensity function. The propensity function for the r^{th} reaction assuming mass-action kinetics is then given by[19]:

$$f_r(\mathbf{n}) = k_r \Omega \prod_{i=1}^{N} \frac{n_i!}{(n_i - s_{ir})!\Omega^{s_{ir}}}. \quad (9)$$

Furthermore, we shall absorb powers of Ω in to k_r so that the latter has units of inverse time for all reaction types.

The moment equations quantify the time evolution of moments. They can be derived directly from the CME, and have the compact form:

$$\partial_t \langle n_i \ldots n_l \rangle = \sum_{r=1}^{R} \langle (n_i + S_{ir}) \ldots (n_l + S_{lr}) f_r(\mathbf{n}) \rangle - \sum_{r=1}^{R} \langle n_i \ldots n_l f_r(\mathbf{n}) \rangle. \quad (10)$$

where $\langle n_i \ldots n_l \rangle = \sum_{n_i=0}^{\infty} \ldots \sum_{n_l=0}^{\infty} n_i \ldots n_l P(\mathbf{n}, t)$, the angled brackets denote the expectation operator and $S_{ij} = p_{ij} - s_{ij}$.

LMA for feedback loops with and without cooperativity. Here we provide details of the LMA for the nonlinear GRNs involving a feedback loop (Fig. 1a upper) and the feedback loop with cooperativity (Fig. 2b). For the latter, we shall here consider in detail the case of two proteins binding cooperatively to the promoter (cp = 2) and then briefly show how the procedure can be easily extended to the case of any number of proteins binding cooperatively.

Let the number of proteins, the unbound promoter state G and the bound promoter state G^* be denoted by n_p, $n_g = 1$ and $n_g = 0$, respectively. By the LMA, both types of nonlinear GRNs map onto the same linear GRN (Fig. 1a lower) whose stochastic dynamics is described by a master equation of the type given by Eq. (8). This is an equation for the time-evolution of $P(n_p, n_g, t)$. For convenience, since n_g can be in only two states, we write $P_{1-n_g}(n_p, t) = P(n_p, n_g, t)$ meaning that $P_0(n_p, t)$ is the probability that the promoter is in state G and there are n_p proteins at time t and $P_1(n_p, t)$ is the probability that the promoter is in state G^* and there are n_p proteins at time t. Thus, it follows that we can write the master equation as a set of two coupled master equations, one for each state:

$$\begin{aligned}
\partial_t P_0(n_p) &= \rho_u \left[P_0(n_p - 1) - P_0(n_p) \right] + (n_p + 1) P_0(n_p + 1) \\
&\quad - n_p P_0(n_p) + \sigma_u P_1(n_p) - \bar{\sigma}_b P_0(n_p), \\
\partial_t P_1(n_p) &= \rho_b \left[P_1(n_p - 1) - P_1(n_p) \right] + (n_p + 1) P_1(n_p + 1) \\
&\quad - n_p P_1(n_p) - \sigma_u P_1(n_p) + \bar{\sigma}_b P_0(n_p).
\end{aligned} \quad (11)$$

Note that the argument t is suppressed for notation simplicity. Note also that time and the parameters are dimensionless since we divided by the rate of protein degradation (as done in ref. [7]). In ref. [5], an exact time-dependent solution of the same master equations was obtained for the special case $\rho_u = 0$. It is straightforward to use the same method to treat the more general case of non-zero ρ_u, which leads to the exact solution for the probability distribution of the number of proteins at time t:

$$P(n_p, t) = \frac{1}{n_p!} \frac{d^{n_p}}{dw^{n_p}} (G_0(w, t) + G_1(w, t))|_{w=-1}, \quad (12)$$

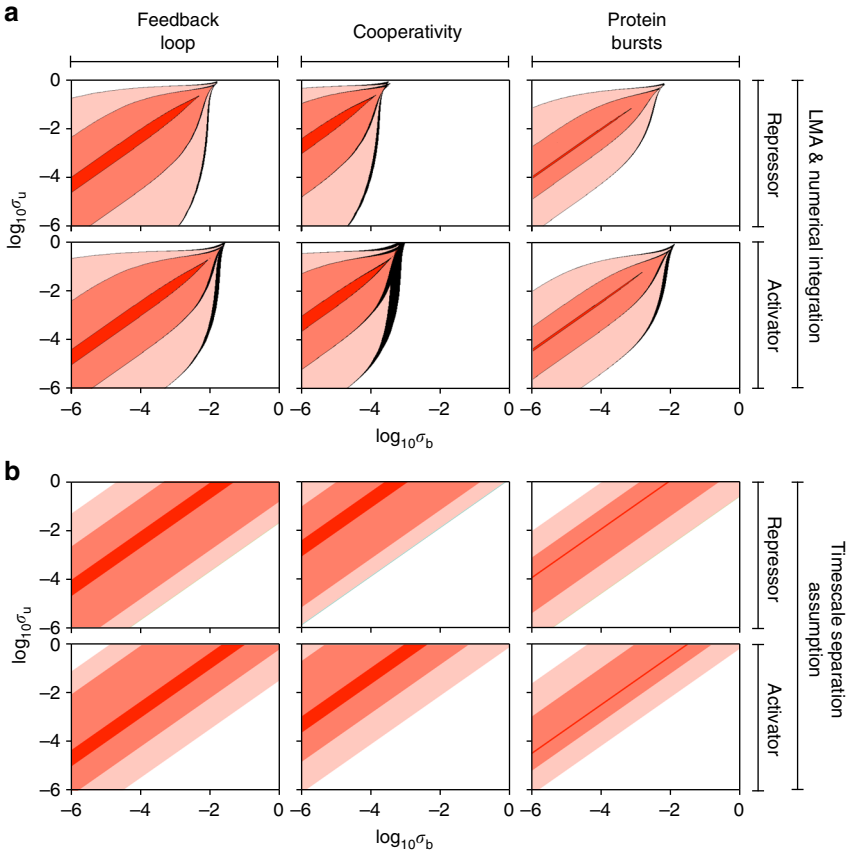


Fig. 9 Stochastic bifurcation diagram for non-oscillatory feedback loops. **a** The red shaded areas denote the regions of parameter space where the steady-state distribution of protein numbers is bimodal, the white areas shows the regions where the distribution is unimodal and the black areas show regions where the LMA and direct numerical integration of the master equations disagree on the number of modes of the distribution. The black regions are small and thus verify the accuracy of the LMA. The shade of red indicates the difference between gene expression in the two promoter states, where a lighter shade indicates a larger difference. Specifically from inside to outside, the dark red corresponds to $\rho_u = 40$, $\rho_b = 25$ (repressor) and $\rho_u = 25$, $\rho_b = 40$ (activator), the medium red corresponds to $\rho_u = 50$, $\rho_b = 25$ (repressor) and $\rho_u = 25$, $\rho_b = 50$ (activator), and the light red corresponds to $\rho_u = 60$, $\rho_b = 25$ (repressor) and $\rho_u = 25$, $\rho_b = 60$ (activator). The cooperativity order was cp = 2 for the feedback loop with cooperativity and the mean protein bursts size was $b = 2$ for the feedback loop with bursting. Generally the region of parameter space where bimodality is present is decreased by cooperativity and by bursting but is almost unaffected by the type of feedback (activating or repressing). **b** This is the same as **a** except that the number of modes of the steady-state distribution of protein numbers is calculated using a method in the literature which assumes timescale separation, i.e. slow promoter switching[12]. While this method captures the salient features of the bifurcation diagrams in **a** it also significantly over-estimates the extent of bimodality and thus illustrates the advantage of the LMA over timescale separation methods

where,

$$G_0(w,t) = \exp(\rho_b w)\Big[f(we^{-t})(-\rho_\Delta w)^{1-\Sigma} M\big(1-\bar{\sigma}_b, 2-\Sigma, -\rho_\Delta w\big) + g(we^{-t}) M\big(1+\sigma_u, \Sigma, -\rho_\Delta w\big)\Big], \quad (13)$$

$$G_1(w,t) = \sigma_u^{-1}\exp(\rho_b w)\Big[-\sigma_u f(we^{-t})(-\rho_\Delta w)^{1-\Sigma} M\big(-\bar{\sigma}_b, 2-\Sigma, -\rho_\Delta w\big) + \bar{\sigma}_b g(we^{-t}) M\big(\sigma_u, \Sigma, -\rho_\Delta w\big)\Big], \quad (14)$$

where $w = z - 1$ and the generating functions are defined as $G_i(z,t) = \sum_{n_p=0}^{\infty} z^{n_p} P_i\big(n_p, t\big)$. The function M(·,·,·) represents the Kummer function and we have also used the definitions $\rho_\Delta = \rho_b - \rho_u$, $\Sigma = \sigma_u + \bar{\sigma}_b + 1$, $f(w) = \frac{\bar{\sigma}_b}{\Sigma-1}(-\rho_\Delta w)^{\Sigma-1} e^{-\rho_u w} M\big(\sigma_u, \Sigma, -\rho_\Delta w\big)$ and $g(w) = \frac{\sigma_u}{\Sigma-1} e^{-\rho_u w} M\big(-\bar{\sigma}_b, 2-\Sigma, -\rho_\Delta w\big)$. We have assumed the initial conditions to be zero protein in state G which translates in to the conditions: $P_0(0,0) = 1$, $P_0(n_p, 0) = 0$ for $n_p > 0$ and $P_1(n_p, 0) = 0$ for all n_p.

In steady-state conditions, the solution simplifies to:

$$P\big(n_p\big) = \frac{1}{n_p!}\frac{\mathrm{d}^{n_p}}{\mathrm{d}w^{n_p}} G(w)|_{w=-1}, \quad (15)$$

where

$$G(w) = \frac{\exp(\rho_b w)\sigma_u}{\sigma_u + \bar{\sigma}_b} M\big(1+\sigma_u, \Sigma, -\rho_\Delta w\big) + \frac{\exp(\rho_b w)\bar{\sigma}_b}{\sigma_u + \bar{\sigma}_b} M\big(\sigma_u, \Sigma, -\rho_\Delta w\big). \quad (16)$$

Next we have to determine the value of the effective parameter $\bar{\sigma}_b$ using the LMA's mean-field assumption. From Eq. (10), one can obtain the moment equations of the linear GRN up to the third order:

$$\mathcal{M}_{\mathrm{FL}}(\bar{\sigma}_b): \begin{cases} \partial_t\langle n_p\rangle = \rho_u\langle n_g\rangle + \rho_b\big(1-\langle n_g\rangle\big) - \langle n_p\rangle, \\ \partial_t\langle n_g\rangle = -\bar{\sigma}_b\langle n_g\rangle + \sigma_u\big(1-\langle n_g\rangle\big), \\ \partial_t\langle n_p^2\rangle = 2(\rho_u-\rho_b)\langle n_p n_g\rangle + (2\rho_b+1)\langle n_p\rangle - 2\langle n_p^2\rangle + (\rho_u-\rho_b)\langle n_g\rangle + \rho_b, \\ \partial_t\langle n_p n_g\rangle = \rho_u\langle n_g\rangle + \sigma_u\langle n_p\rangle - (1+\bar{\sigma}_b+\sigma_u)\langle n_p n_g\rangle, \\ \partial_t\langle n_p^2 n_g\rangle = (2\rho_u+1)\langle n_p n_g\rangle - (\sigma_u+\bar{\sigma}_b+2)\langle n_p^2 n_g\rangle + \rho_u\langle n_g\rangle + \sigma_u\langle n_p^2\rangle. \end{cases} \quad (17)$$

Note that $\mathcal{M}_{\mathrm{FL}}(\bar{\sigma}_b)$ is a set of closed ODEs given that $\bar{\sigma}_b$ is specified. The proposed LMA rests upon the idea of parametrizing $\bar{\sigma}_b$ by means of the conditional mean. Specifically, for the nonlinear feedback loop without cooperativity, since the nonlinear reaction is $P + G \to G^*$, then $\bar{\sigma}_b = \sigma_b\langle n_p | n_g = 1\rangle = \sigma_b\langle n_p n_g\rangle / \langle n_g\rangle$, where we used the fact that n_g is a Boolean variable. Similarly, for the nonlinear

feedback loop with cooperative order cp = 2, since the nonlinear reaction is $2P + G \to G^*$ then $\bar{\sigma}_b = \sigma_b \left\langle n_p \left(n_p - 1 \right) | n_g = 1 \right\rangle = \sigma_b \left\langle n_p \left(n_p - 1 \right) n_g \right\rangle / \left\langle n_g \right\rangle$.

Subsequently, we solve the set of differential equations with initial conditions $\left\langle n_p \right\rangle = \left\langle n_p^2 \right\rangle = \left\langle n_p n_g \right\rangle = \left\langle n_p^2 n_g \right\rangle = 0, \left\langle n_g \right\rangle = 1$ and with the aforementioned $\bar{\sigma}_b$ parameterization, i.e., solving $\mathcal{M}_{\rm FL}(\bar{\sigma}_b = \sigma_b \left\langle n_p n_g \right\rangle / \left\langle n_g \right\rangle)$ for feedback loop or $\mathcal{M}_{\rm FL}\left(\bar{\sigma}_b = \sigma_b \left(\left\langle n_p^2 n_g \right\rangle - \left\langle n_p n_g \right\rangle \right) / \left\langle n_g \right\rangle \right)$ for cooperative network with cp = 2 on the time interval [0, t]. We denote the solved moments of interest at time t' as $\left\langle n_g \right\rangle_{t'}$, $\left\langle n_p n_g \right\rangle_{t'}$ and $\left\langle n_p^2 n_g \right\rangle_{t'}$, where $0 \le t' \le t$. Hence, the effective time-dependent constants in the linear GRN are given by:

$$\bar{\sigma}_b(t') = \sigma_b \frac{\left\langle n_p n_g \right\rangle_{t'}}{\left\langle n_g \right\rangle_{t'}}, \bar{\sigma}_b(t') = \sigma_b \frac{\left\langle n_p^2 n_g \right\rangle_{t'} - \left\langle n_p n_g \right\rangle_{t'}}{\left\langle n_g \right\rangle_{t'}}, \tag{18}$$

for the noncooperative and cooperative feedback loops, respectively.

From these, we can compute the the time-averaged value of the effective parameter $\bar{\sigma}_b$ at time t:

$$\bar{\sigma}_b^* = \frac{\sigma_b}{t} \int_0^t \frac{\left\langle n_p n_g \right\rangle_{t'}}{\left\langle n_g \right\rangle_{t'}} {\rm d}t' \text{ and } \bar{\sigma}_b^{**} = \frac{\sigma_b}{t} \int_0^t \frac{\left\langle n_p^2 n_g \right\rangle_{t'} - \left\langle n_p n_g \right\rangle_{t'}}{\left\langle n_g \right\rangle_{t'}} {\rm d}t',$$

for non-cooperative and cooperative loops, respectively. This is the time-averaging assumption of the LMA.

Finally the approximate probability distribution at time t of the nonlinear GRN without cooperativity is given by Eqs. (12–14) with $\bar{\sigma}_b$ replaced by $\bar{\sigma}_b^*$ and for cooperativity the distribution is given by Eqs. (12–14) with $\bar{\sigma}_b$ replaced by $\bar{\sigma}_b^{**}$.

In steady-state, the solution is simpler. The moment equations can be solved with the time-derivative set to zero, i.e., $\mathcal{M}_{\rm FL}\left(\bar{\sigma}_b = \sigma_b \left\langle n_p n_g \right\rangle / \left\langle n_g \right\rangle\right) = 0$, leading to explicit expressions for $\left\langle n_g \right\rangle$ and $\left\langle n_p n_g \right\rangle$ from which one can calculate the effective parameter $\bar{\sigma}_b^* = \sigma_b \left\langle n_p n_g \right\rangle / \left\langle n_g \right\rangle$ of the non-cooperative feedback loop:

$$\bar{\sigma}_b^* = \frac{-1 + \rho_b \sigma_b - \sigma_u + \sqrt{\left(1 - \rho_b \sigma_b + \sigma_u\right)^2 + 4\rho_u \sigma_b (1 + \sigma_u)}}{2}. \tag{19}$$

For the cooperative feedback loop, the effective parameter $\bar{\sigma}_b^{**} = \sigma_b \left(\left\langle n_p^2 n_g \right\rangle - \left\langle n_p n_g \right\rangle \right) / \left\langle n_g \right\rangle$ can be obtained in a similar way. It is found to be the solution of a third-order polynomial given by:

$$\bar{\sigma}_b^{**} = \sigma_b \frac{\rho_b^2 \bar{\sigma}_b^{**} \left(1 + \bar{\sigma}_b^{**}\right) + 2\rho_b \rho_u \bar{\sigma}_b^{**} (1 + \sigma_u) + \rho_u^2 (1 + \sigma_u)(2 + \sigma_u)}{\left(1 + \bar{\sigma}_b^{**} + \sigma_u\right)\left(2 + \bar{\sigma}_b^{**} + \sigma_u\right)}. \tag{20}$$

Finally, the approximate steady-state probability distribution of the nonlinear GRN without cooperativity is given by Eqs. (15) and (16) with $\bar{\sigma}_b$ replaced by $\bar{\sigma}_b^*$ in Eq. (19) and for cooperativity the distribution is given by Eqs. (15) and (16) with $\bar{\sigma}_b$ replaced by $\bar{\sigma}_b^{**}$ as obtained from solving Eq. (20). See the Supplementary Note 3 for details of an efficient numerical implementation of the LMA of the feedback loop using Mathematica.

The procedure can also be easily extended for the case of general number of proteins cp = n binding cooperatively to the promoter. In this case by the mean-field approximation, the effective parameter is:

$$\bar{\sigma}_b^{***} = \sigma_b \frac{\left\langle \prod_{i=0}^{n-1} \left(n_p - i \right) n_g \right\rangle}{\left\langle n_g \right\rangle}.$$

By substituting in the moment equations up to the order of $\left\langle n_p^n n_g \right\rangle$ and solving in steady-state, one finds that the effective rate constant is the solution of the following implicit function:

$$\bar{\sigma}_b^{***} = \sigma_b \frac{\sum_{i=0}^{n} C_i^n \rho_u^i \rho_b^{n-i} \prod_{j=1}^{i} (j + \sigma_u) \prod_{j=0}^{n-1-i} \left(j + \bar{\sigma}_b^{***}\right)}{\prod_{i=1}^{n} \left(\bar{\sigma}_b^{***} + \sigma_u + i\right)}.$$

Finally, the approximate steady-state probability distribution of the nonlinear GRN with cooperativity order n is given by Eqs. (15) and (16) with $\bar{\sigma}_b$ replaced by $\bar{\sigma}_b^{***}$ as obtained from solving the implicit equation above. The construction of the time-dependent solution parallels that previously shown for the special case of $n = 2$.

LMA for feedback loop with protein bursts. The feedback loop with protein bursting (Fig. 2a) is mapped onto the linear GRN described by:

$$G \xrightarrow{\rho_u} G + mP, G^* \xrightarrow{\rho_b} G^* + mP, G \xrightarrow{\bar{\sigma}_b} G^*, G^* \xrightarrow{\sigma_u} G, P \xrightarrow{1} \emptyset, \tag{21}$$

where m is a discrete random variable sampled from the geometric distribution $\psi(m) = b^m/(1 + b)^{m+1}$ (see main text for justification of the choice of distribution). The mean burst size of gene expression is given by b. As in the previous example of noncooperative and cooperative feedback loops, we can derive coupled master equations of the linear GRN above:

$$\begin{aligned} \partial_t P_0\left(n_p\right) = {} & \rho_u \sum_{i=0}^{\infty} \psi(i) P_0\left(n_p - i\right) - \rho_u P_0\left(n_p\right) \\ & + \left(n_p + 1\right) P_0\left(n_p + 1\right) - n_p P_0\left(n_p\right) \\ & + \sigma_u P_1\left(n_p\right) - \bar{\sigma}_b P_0\left(n_p\right), \end{aligned} \tag{22}$$

$$\begin{aligned} \partial_t P_1\left(n_p\right) = {} & \rho_b \sum_{i=0}^{\infty} \psi(i) P_1\left(n_p - i\right) - \rho_b P_1\left(n_p\right) \\ & + \left(n_p + 1\right) P_1\left(n_p + 1\right) - n_p P_1\left(n_p\right) \\ & - \sigma_u P_1\left(n_p\right) + \bar{\sigma}_b P_0\left(n_p\right). \end{aligned} \tag{23}$$

A time-dependent solution for these master equations is presently missing from the literature and we provide one in the Supplementary Note 1. Here we shall simply refer to the exact steady-state and time-dependent solutions as $\mathcal{S}_{\rm PBSS}(\bar{\sigma}_b)$ and $\mathcal{S}_{\rm PBTD}(\bar{\sigma}_b)$, respectively.

Using Eq. (10), one can obtain the corresponding moment equations up to the second order:

$$\mathcal{M}_{\rm PB}(\bar{\sigma}_b): \begin{cases} \partial_t \left\langle n_p \right\rangle = \rho_u b \left\langle n_g \right\rangle + \rho_b b \left(1 - \left\langle n_g \right\rangle\right) - \left\langle n_p \right\rangle, \\ \partial_t \left\langle n_g \right\rangle = \sigma_u \left(1 - \left\langle n_g \right\rangle\right) - \bar{\sigma}_b \left\langle n_g \right\rangle, \\ \partial_t \left\langle n_p n_g \right\rangle = \rho_u b \left\langle n_g \right\rangle + \sigma_u \left\langle n_p \right\rangle - (1 + \bar{\sigma}_b + \sigma_u) \left\langle n_p n_g \right\rangle \end{cases}$$

Since the only nonlinear reaction in the nonlinear GRN is $P + G \to G^*$ then by the LMA's mean-field assumption $\bar{\sigma}_b = \sigma_b \left\langle n_p | n_g = 1 \right\rangle = \sigma_b \left\langle n_p n_g \right\rangle / \left\langle n_g \right\rangle$. Solving the set of differential equations $\mathcal{M}_{\rm PB}\left(\sigma_b \left\langle n_p | n_g = 1 \right\rangle\right)$ on the time interval [0, t], we obtain the moments of interest at time t' as $\left\langle n_g \right\rangle_{t'}$ and $\left\langle n_p n_g \right\rangle_{t'}$. The time-average $\bar{\sigma}_b^*$ is then constructed as before. The approximate solution for the probability distribution at time t of the nonlinear GRN with protein bursts is given by $\mathcal{S}_{\rm PBTD}\left(\bar{\sigma}_b^*\right)$.

For steady-state conditions, the moment equations can be solved explicitly (as before for the noncooperative and cooperative feedback loops) yielding:

$$\bar{\sigma}_b^* = \frac{1}{2}\left[-1 + \rho_b b \sigma_b - \sigma_u + \sqrt{\left(1 - \rho_b b \sigma_b + \sigma_u\right)^2 + 4\rho_u b \sigma_b (1 + \sigma_u)}\right].$$

The approximate steady-state solution for the probability distribution of the nonlinear GRN with protein bursts is then given by $\mathcal{S}_{\rm PBSS}\left(\bar{\sigma}_b^*\right)$.

LMA for feedback loop with oscillatory transcription. The feedback loop with oscillatory transcription (Fig. 2c) is mapped onto the same linear GRN used for cooperative and noncooperative feedback loops, namely that shown in Fig. 1a lower except for the parameters ρ_u and ρ_b which become $\rho_u \ell_t$ and $\rho_b \ell_t$, where $\ell_t = 1 + {\rm Am}\cos(k\pi t)$ is an oscillatory function, where Am is the amplitude and k is the frequency. Note that $0 < {\rm Am} < 1$ such that the protein production rate in each promoter state is positive at all times. The master equations and moment equations are thus given by Eqs. (11) and (17), respectively, with the aforementioned substitutions. An explicit time-dependent probability distribution solution of the master equations can be found in the Supplementary Note 2. The approximate time-dependent distributions of the nonlinear GRN can then be obtained by the same LMA procedure as for the noncooperative feedback loop.

The time-averaging assumption in the LMA. The chemical master equation can be compactly written as:

$$\frac{{\rm d}\mathbf{P}(t)}{{\rm d}t} = \mathbf{A}_{\rm L}(t)\mathbf{P}(t), \tag{24}$$

where $\mathbf{P}(t) = [P_0(t), P_1(t), \ldots]^\top$ and $P_i(t)$ is the probability that the system is in state i at time t. Each entry of the transition matrix $\mathbf{A}_{\rm L}(t)$ is defined by the propensity function governing the transition from one state to another. By means of the Magnus expansion of linear differential equations[41–43], the solution of the

master equation at time $t = T$ can be written as:

$$\mathbf{P}(T) = \exp(\mathbf{\Omega}(T))\mathbf{P}_0,$$

and $\mathbf{\Omega}(T) = \sum_{i=1}^{\infty} \mathbf{\Omega}_i(T)$. The first two terms of this expansion are:

$$\mathbf{\Omega}_1(T) = \int_0^T \mathbf{A}_{\mathrm{L}}(t)\mathrm{d}t, \tag{25}$$

$$\mathbf{\Omega}_2(T) = \frac{1}{2}\int_0^T \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}t_2 [\mathbf{A}_{\mathrm{L}}(t_1), \mathbf{A}_{\mathrm{L}}(t_2)]. \tag{26}$$

The convergence of this expansion has been extensively studied (for a review of known results see Section 2.7 in ref. [42]). The time-averaging assumption corresponds to the first term of the Magnus expansion, i.e., truncating the expansion to include only $\mathbf{\Omega}_1(T)$. This is since this term is the same as if we had to first solve Eq. (24) assuming a time-independent transition matrix, leading to $\mathbf{P}(T) = \exp(\mathbf{A}_{\mathrm{L}}T)\mathbf{P}_0$ and then replace the time-independent transition matrix $\mathbf{A}_{\mathrm{L}}$ in this result by the time-averaged matrix $\int_0^T \mathbf{A}_{\mathrm{L}}(t)\mathrm{d}t/T$. This first term of the Magnus series of the master equation can be shown to give a well-defined probability vector and hence is physically meaningful to consider the expansion to this order only (see Supplementary Note 8 for a proof). Hence the approximation error of our time-averaging assumption is given by the rest of the terms in the expansion. What follows is an analysis of this error, in particular we prove that the error is small for all times in the limit of small protein–promoter binding rate, σ_{b}.

First of all, we note that for nonlinear GRNs with constant rates, the time dependence of $\mathbf{A}_{\mathrm{L}}$(t) with the LMA arises from the mapping to a linear GRN with a time-dependent protein–promoter binding rate $\bar{\sigma}_{\mathrm{b}}(t)$ (for example see Eq. (18) for the case of feedback loops with and without cooperativity). We now look at the time-dependence of $\bar{\sigma}_{\mathrm{b}}(t)$. Since there are zero proteins initially, $\bar{\sigma}_{\mathrm{b}}(0) = 0$ while for long times $\bar{\sigma}_{\mathrm{b}}(t)$ approaches a constant steady-state value determined by the steady-state values of the moments, e.g. the value of $\langle n_{\mathrm{p}} n_{\mathrm{g}} \rangle$ and $\langle n_{\mathrm{g}} \rangle$ for the non-cooperative feedback loop. Furthermore the approach to steady-state occurs exponentially (see Supplementary Note 8 for a proof).

Since the time dependence of $\mathbf{A}_{\mathrm{L}}(t)$ stems from $\bar{\sigma}_{\mathrm{b}}(t)$, it also follows that $\mathbf{A}_{\mathrm{L}}(t)$ converges to $\mathbf{A}_{\mathrm{L}}(\infty)$ exponentially. This implies the following two statements. There exist some positive real numbers C_1 and δ_1 such that

$$\|\mathbf{A}_{\mathrm{L}}(t) - \mathbf{A}_{\mathrm{L}}(\infty)\| \le C_1 \mathrm{e}^{-\delta_1 t}$$

for any t and the matrix $\mathbf{A}_{\mathrm{L}}(t)$ can be expressed in terms of the steady-state matrix, i.e.,

$$\mathbf{A}_{\mathrm{L}}(t) = \mathbf{A}_{\mathrm{L}}(\infty) + \sigma_{\mathrm{b}}\mathbf{D}(t),$$

where $\mathbf{D}(t)$ is a discrepancy matrix and $\|\mathbf{D}(t)\| \le C_2 \mathrm{e}^{-\delta_2 t}$ for some positive real numbers δ_2 and C_2. Note that all norms are matrix 2-norms. Thus, for the Lie bracket, we have:

$$\begin{aligned}\|[\mathbf{A}_{\mathrm{L}}(t_1), \mathbf{A}_{\mathrm{L}}(t_2)]\| = \; & \|[\mathbf{A}_{\mathrm{L}}(\infty) + \sigma_{\mathrm{b}}\mathbf{D}(t_1), \mathbf{A}_{\mathrm{L}}(\infty) + \sigma_{\mathrm{b}}\mathbf{D}(t_2)]\| \\ \le \; & \sigma_{\mathrm{b}}\|\mathbf{A}_{\mathrm{L}}(\infty)\|\|\mathbf{D}(t_1)\| \\ & + \sigma_{\mathrm{b}}\|\mathbf{A}_{\mathrm{L}}(\infty)\|\|\mathbf{D}(t_2)\| \\ & + \sigma_{\mathrm{b}}^2\|\mathbf{D}(t_1)\|\|\mathbf{D}(t_2)\| \\ \le \; & \sigma_{\mathrm{b}} C_3 \mathrm{e}^{-\delta_2 t_2}\end{aligned}$$

for some positive real number C_3 and for any $t_2 \le t_1$. Therefore, the matrix norm $\mathbf{\Omega}_2(T)$ is upper bounded by:

$$\|\mathbf{\Omega}_2(T)\| \le \sigma_{\mathrm{b}} \frac{C_3}{2} \int_0^T \int_0^{t_1} \mathrm{e}^{-\delta_2 t_2} \mathrm{d}t_2 = \frac{\sigma_{\mathrm{b}} C_3}{2\delta_2^2}\left(\delta_2 T + \mathrm{e}^{-\delta_2 T} - 1\right) \sim \sigma_{\mathrm{b}}\mathcal{O}(T),$$

On the other hand, it is known that:

$$\lim_{T\to\infty} \frac{\|\mathbf{\Omega}_1(T)\|}{T} = \|\mathbf{A}_{\mathrm{L}}(\infty)\|.$$

Thus, we have:

$$\lim_{T\to\infty} \frac{\|\mathbf{\Omega}_2(T)\|}{\|\mathbf{\Omega}_1(T)\|} \sim \sigma_{\mathrm{b}}\mathcal{O}(1).$$

This result indicates that the approximation error is bounded in time and first order in σ_{b}. Applying similar arguments, it can be shown that all higher-order terms in the Magnus expansion ($\mathbf{\Omega}_i$, where $i \ge 3$) are bounded in time and have higher orders in σ_{b} than the first two terms. In other words, one can conclude that the time-averaging assumption of the LMA is uniformly valid in time and accurate provided the protein–promoter binding rate σ_{b} is small.

The first-passage time distribution of promoter switching. For our purposes, this is the distribution of the time it takes for the promoter to switch from G to G^*, given there are n proteins initially. We shall show this for the cooperative and noncooperative feedback loops (similar can be done for the bursty loop). Given the process, the only relevant reactions for this calculation are:

$$G \xrightarrow{\rho_{\mathrm{u}}} G + P, \mathrm{cp}\, P + G \xrightarrow{\sigma_{\mathrm{b}}} G^*, P \xrightarrow{1} \emptyset, \tag{27}$$

where $\mathrm{cp} = 1$ for the noncooperative feedback loop, and $\mathrm{cp} > 1$ for the cooperative feedback loop. The LMA maps this onto the simpler linear GRN:

$$G \xrightarrow{\rho_{\mathrm{u}}} G + P, \; G \xrightarrow{\bar{\sigma}_{\mathrm{b}}} G^*, P \xrightarrow{1} \emptyset \tag{28}$$

Using the same recipe as before, one writes the moment equations, applies the mean-field assumption, obtains the relevant moments and then calculates the effective time-dependent constants $\bar{\sigma}_{\mathrm{b}}(t')$. This implies the solution of the moment equations Eq. (17) with the constants ρ_{b} and σ_{u} set to zero (since the associated reactions are irrelevant to the first-passage time process as described above) and initial conditions $\langle n_{\mathrm{p}} \rangle = n$, $\langle n_{\mathrm{p}}^2 \rangle = n^2$, $\langle n_{\mathrm{p}} n_{\mathrm{g}} \rangle = n$, $\langle n_{\mathrm{p}}^2 n_{\mathrm{g}} \rangle = n^2$, $\langle n_{\mathrm{g}} \rangle = 1$.

The first-passage time distribution to switch from G to G^* is given by $P(t_{\mathrm{FP}} = t) = -\partial_t P_0(t)$[44], where $P_0(t)$ is the probability that the system is still in state G at time t given that initially it is in this state. Since the LMA maps the second-order reaction onto the linear reaction $G \to G^*$ with effective rate $\bar{\sigma}_{\mathrm{b}}(t)$, it follows from elementary probability arguments that:

$$P_0(t) = \exp\left(-\int_0^t \bar{\sigma}_{\mathrm{b}}(t')\mathrm{d}t'\right).$$

Hence, the final expression for the first-passage time distribution in the LMA is given by:

$$P(t_{\mathrm{FP}} = t) = \bar{\sigma}_{\mathrm{b}}(t) \exp\left(-\int_0^t \bar{\sigma}_{\mathrm{b}}(t')\mathrm{d}t'\right).$$

Code availability. The Mathematica code solving the LMA for the simple non-linear feedback loop (schematically shown in Fig. 1a (upper)) can be found at https://github.com/edwardcao3026/Linear-mapping-approximation. The details are provided in Supplementary Note 3.

Data availability. All relevant data are available from the authors.

Received: 8 January 2018 Accepted: 20 July 2018
Published online: 17 August 2018

References

1. Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. *Systems Biology in Practice: Concepts, Implementation and Application* (Wiley, Weinheim, 2008).
2. Grima, R. & Schnell, S. Modelling reaction kinetics inside cells. *Essays Biochem.* **45**, 41–56 (2008).
3. McAdams, H. H. & Arkin, A. It's a noisy business! genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69 (1999).
4. Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).
5. Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* **79**, 031911 (2009).
6. Pendar, H., Platini, T. & Kulkarni, R. V. Exact protein distributions for stochastic models of gene expression using partitioning of poisson processes. *Phys. Rev. E* **87**, 042720 (2013).
7. Grima, R., Schmidt, D. R. & Newman, T. J. Steady-state fluctuations of a genetic feedback loop: An exact solution. *J. Chem. Phys.* **137**, 035104 (2012).
8. Kumar, N., Platini, T. & Kulkarni, R. V. Exact distributions for stochastic gene expression models with bursting and feedback. *Phys. Rev. Lett.* **113**, 268105 (2014).
9. Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, Vol. 1 (Elsevier, Amsterdam, 1992).
10. Elf, J. & Ehrenberg, M. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.* **13**, 2475–2484 (2003).
11. Hayot, F. & Jayaprakash, C. The linear noise approximation for molecular fluctuations within cells. *Phys. Biol.* **1**, 205 (2004).
12. Thomas, P., Popović, N. & Grima, R. Phenotypic switching in gene regulatory networks. *Proc. Natl Acad. Sci. USA* **111**, 6994–6999 (2014).

13. Mélykúti, B., Hespanha, J. P. & Khammash, M. Equilibrium distributions of simple biochemical reaction systems for time-scale separation in stochastic reaction networks. *J. R. Soc. Interface* **11**, 20140054 (2014).
14. Roussel, M. R. & Zhu, R. Reducing a chemical master equation by invariant manifold methods. *J. Chem. Phys.* **121**, 8716–8730 (2004).
15. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc. Natl Acad. Sci. USA* **105**, 17256–17261 (2008).
16. Duncan, A., Liao, S., Vejchodský, T., Erban, R. & Grima, R. Noise-induced multistability in chemical systems: Discrete versus continuum modeling. *Phys. Rev. E* **91**, 042111 (2015).
17. Walczak, A. M., Sasai, M. & Wolynes, P. G. Self-consistent proteomic field theory of stochastic gene switches. *Biophys. J.* **88**, 828–850 (2005).
18. Thomas, P. & Grima, R. Approximate probability distributions of the master equation. *Phys. Rev. E* **92**, 012120 (2015).
19. Schnoerr, D., Sanguinetti, G. & Grima, R. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *J. Phys. A* **50**, 093001 (2017).
20. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
21. Munsky, B. & Khammash, M. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* **124**, 044104 (2006).
22. Andreychenko, A., Bortolussi, L., Grima, R., Thomas, P. & Wolf, V. In: *Modeling Cellular Systems* (eds Graw, F. et al.) 39–66 (Springer, Cham, Switzerland, 2017).
23. Smadbeck, P. & Kaznessis, Y. N. A closure scheme for chemical master equations. *Proc. Natl Acad. Sci. USA* **110**, 14261–14265 (2013).
24. Bar-Even, A. et al. Noise in protein expression scales with natural protein abundance. *Nat. Gen.* **38**, 636 (2006).
25. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–362 (2006).
26. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in escherichia coli. *Nature* **403**, 339–342 (2000).
27. Dodd, A. N. et al. Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* **309**, 630–633 (2005).
28. Rao, C. V. & Arkin, A. P. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.* **118**, 4999–5010 (2003).
29. Thomas, P., Straube, A. V. & Grima, R. The slow-scale linear noise approximation: an accurate, reduced stochastic description of biochemical networks under timescale separation conditions. *BMC Syst. Biol.* **6**, 39 (2012).
30. Albayrak, C. et al. Digital quantification of proteins and mRNA in single mammalian cells. *Mol. Cell* **61**, 914–924 (2016).
31. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
32. Schnoerr, D., Cseke, B., Grima, R. & Sanguinetti, G. Efficient low-order approximation of first-passage time distributions. *Phys. Rev. Lett.* **119**, 210601 (2017).
33. Rao, C. V., Wolf, D. M. & Arkin, A. P. Control, exploitation and tolerance of intracellular noise. *Nature* **420**, 231–237 (2002).
34. Fell, D. & Cornish-Bowden, A. *Understanding the Control of Metabolism*, Vol. 2 (Portland Press, London, 1997).
35. Qian, H., Shi, P.-Z. & Xing, J. Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. *Phys. Chem. Chem. Phys.* **11**, 4861–4870 (2009).
36. Bressloff, P. C. Stochastic switching in biology: from genotype to phenotype. *J. Phys. A* **50**, 133001 (2017).
37. Ochab-Marcinek, A. & Tabaka, M. Bimodal gene expression in noncooperative regulatory systems. *Proc. Natl Acad. Sci. USA* **107**, 22096–22101 (2010).
38. Grima, R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.* **136**, 154105 (2012).
39. Soltani, M., Vargas-Garcia, C. A. & Singh, A. Conditional moment closure schemes for studying stochastic dynamics of genetic circuits. *IEEE Trans. Biomed. Circuits Syst.* **9**, 518–526 (2015).
40. Lakatos, E., Ale, A., Kirk, P. D. & Stumpf, M. P. Multivariate moment closure techniques for stochastic kinetic models. *J. Chem. Phys.* **143**, 094107 (2015).
41. Magnus, W. On the exponential solution of differential equations for a linear operator. *Commun. Pure Appl. Math.* **7**, 649–673 (1954).
42. Blanes, S., Casas, F., Oteo, J. A. & Ros, J. The Magnus expansion and some of its applications. *Phys. Rep.* **470**, 151–238 (2009).
43. Iserles, A. & MacNamara, S. Magnus expansions and pseudospectra of master equations. Preprint at https://arxiv.org/abs/1701.02522 (2017).
44. Redner, S. *A Guide to First-passage Processes* (Cambridge Univ. Press, Cambridge, 2001).

Acknowledgements

This work was supported by a grant from the BBSRC (BB/M018040/1).

Author contributions

Z.C. and R.G. designed research, carried out research and wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05822-0.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells

Zhixing Cao[a,b] and Ramon Grima[b,1]

[a]The Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, People's Republic of China; and [b]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Edited by Charles S. Peskin, New York University, New York, NY, and approved January 21, 2020 (received for review June 25, 2019)

The stochasticity of gene expression presents significant challenges to the modeling of genetic networks. A two-state model describing promoter switching, transcription, and messenger RNA (mRNA) decay is the standard model of stochastic mRNA dynamics in eukaryotic cells. Here, we extend this model to include mRNA maturation, cell division, gene replication, dosage compensation, and growth-dependent transcription. We derive expressions for the time-dependent distributions of nascent mRNA and mature mRNA numbers, provided two assumptions hold: 1) nascent mRNA dynamics are much faster than those of mature mRNA; and 2) gene-inactivation events occur far more frequently than gene-activation events. We confirm that thousands of eukaryotic genes satisfy these assumptions by using data from yeast, mouse, and human cells. We use the expressions to perform a sensitivity analysis of the coefficient of variation of mRNA fluctuations averaged over the cell cycle, for a large number of genes in mouse embryonic stem cells, identifying degradation and gene-activation rates as the most sensitive parameters. Furthermore, it is shown that, despite the model's complexity, the time-dependent distributions predicted by our model are generally well approximated by the negative binomial distribution. Finally, we extend our model to include translation, protein decay, and auto-regulatory feedback, and derive expressions for the approximate time-dependent protein-number distributions, assuming slow protein decay. Our expressions enable us to study how complex biological processes contribute to the fluctuations of gene products in eukaryotic cells, as well as allowing a detailed quantitative comparison with experimental data via maximum-likelihood methods.

stochastic gene expression | master equation | perturbation theory

In the past two decades, advances in the real-time measurement of single-cell dynamics have revealed the stochastic nature of gene expression (1) and spurred a huge interest in the construction, simulation, and analytic solution of stochastic models of intracellular processes (2–4). Many experiments report the measurement of messenger RNA (mRNA), and, hence, there is a general need for stochastic models which can realistically predict the temporally varying distribution of mRNA molecule numbers in single cells. The word "realistically" is key because while there are a number of stochastic models of mRNA fluctuations in the literature, nevertheless, because of the complexity of the mRNA life cycle, currently very few of these models incorporate some of the detailed biological knowledge gleaned from single-cell experiments—this is particularly true for eukaryotic cells, where the transcription process is more complex than in prokaryotes and where compartmentation plays an important role (5).

The simplest stochastic model of mRNA fluctuations assumes that the gene is continuously ON, producing mRNA at some constant rate, followed by mRNA decay or its dilution due to cell division (often referred to as a constitutive expression model). If all these processes are approximated by effective first-order reactions, then the model is easy to solve and predicts a Poisson distribution of mRNA molecule numbers in cells (6). However, there is a large body of experimental evidence showing that the distribution of molecule-number fluctuations is typically non-Poisson (7–10) [except for housekeeping genes (11)], and, hence, modifications of this model are clearly needed. Adding an intermediate state which can either represent nascent mRNA or nuclear mRNA leads to the same Poisson distribution (12) (M1 in Fig. 1*A*). In contrast, assuming that a gene can switch between an ON and an OFF state (M2 in Fig. 1*A*) does lead to non-Poissonian mRNA fluctuations. The model has also been solved exactly analytically (13), and, in certain limits, it predicts bursty mRNA expression (8), a phenomenon which has been measured experimentally (14). This model, commonly called the two-state or telegraph model, has thus been widely adopted in the literature as the standard model of stochastic mRNA dynamics in eukaryotic cells (15–17). A recent application of the model is its use to infer the promoter-switching rates and the transcription rate of thousands of genes in mouse and human fibroblasts (10) from single-cell RNA-sequencing data. However, it is clear that this model is still far from including well-known processes, such as mRNA maturation, cell division, gene replication,

Significance

The random nature of gene expression is well established experimentally. Mathematical modeling provides a means of understanding the factors leading to the observed stochasticity. In this article, we extend the classical two-state model of stochastic mRNA dynamics to include a considerable number of salient features of single-cell biology, such as cell division, replication, mRNA maturation, dosage compensation, and growth-dependent transcription. By means of biologically relevant approximations, we obtain expressions for the time-dependent distributions of mRNA and protein numbers. These provide insight into how fluctuations are modified and controlled by complex intracellular processes.

Author contributions: Z.C. and R.G. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

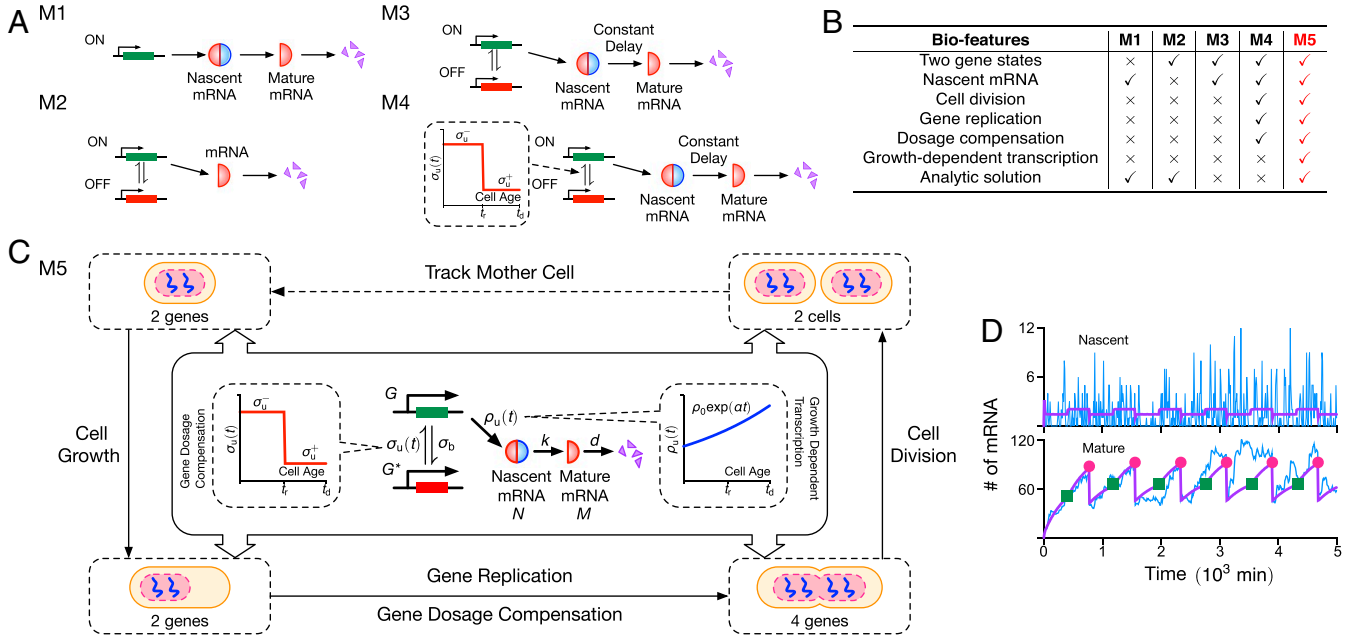
The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: ramon.grima@ed.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910888117/-/DCSupplemental.

First published February 18, 2020.



Bio-features	M1	M2	M3	M4	M5
Two gene states	×	✓	✓	✓	✓
Nascent mRNA	✓	×	✓	✓	✓
Cell division	×	×	×	✓	✓
Gene replication	×	×	×	✓	✓
Dosage compensation	×	×	×	✓	✓
Growth-dependent transcription	×	×	×	×	✓
Analytic solution	✓	✓	×	×	✓

Fig. 1. (*A*) Illustration of four models of stochastic mRNA dynamics in the literature. Note that nascent mRNA is shown as joined blue and red semicircles, illustrating its unspliced nature (blue for introns and red for exons), while mature mRNA being composed of only exons is shown as red semicircles. The models describe transcription (constitutive, e.g., M1; or intermittent, e.g., M2, M3, and M4), mRNA maturation (M1, M3, and M4), and details of the cell cycle (M4); their biological features are compared in *B* and discussed in the main text. Only the simplest of these two models (M1 and M2) have been analytically solved. The model (M5) proposed in this article is illustrated in *C*: It builds upon model M4 by adding growth-dependent transcription, describes maturation as a stochastic process, and has the major advantage of being analytically solvable. The model is composed of nonreactive components (dosage compensation, replication, cell division, and growth-dependent transcription) and reactive components; the latter are shown in the central boxes where G, G^* denote genes in the ON, OFF states; N is nascent mRNA; and M is mature mRNA. (*D*) We show stochastic simulations of M5 using the SSA (25), where the purple lines denote the mean, and a typical time series is shown in blue. The green squares and red dots indicate the gene-replication time (t_r) and cell-division time (t_d), respectively. We use parameters measured for *Nanog* in mouse embryonic stem cells from ref. 26: $\rho_u = 2.11$ min^{-1}, $d = 0.00245$ min^{-1}, $\sigma_b = 0.609$ min^{-1}, and $\sigma_u^- = 0.0282$ min^{-1}. The gene-replication time $t_r = 400$ min, cell-division time $t_d = 780$ min, maturation rate $k = 0.1299$ min^{-1}, and gene-dosage parameters $\sigma_u^+ = 0.71\sigma_u^-$ are reported in ref. 19 for the same type of cells. Note that we set $\alpha = 0$, meaning that there is no growth-dependent transcription. Each realization is initiated with zero nascent and mature mRNAs and the gene in the ON state.

dosage compensation, and growth-dependent transcription, all of which have been shown to have significant effects on the mRNA molecule numbers inside cells.

There are few studies which have modified the standard model of stochastic mRNA dynamics to include some of the mentioned biological processes: Senecal et al. (18) modified the standard model by including the conversion of nascent mRNA to mature mRNA after a constant time delay (model M3 in Fig. 1*A*), while Skinner et al. (19) extended this model further by including cell division, gene replication, and dosage compensation (model M4 in Fig. 1*A*). Note that there are also models which explicitly account for cell division (20, 21) or for replication (22), but have no description for promoter switching or other detailed features of the mRNA life cycle.

The disadvantage of the last two models (M3 and M4) compared to the previous two (M1 and M2) is that there is no known analytic solution to them. Hence, they have been explored purely via stochastic simulations. In Fig. 1*B*, we summarize the features of the models that we have discussed. In this paper, we develop an analytically tractable stochastic model of mRNA dynamics in diploid cells, which includes all of the processes in the most advanced simulation model (M4) developed thus far, while also additionally including growth-dependent transcription (the scaling of transcription with cellular volume) (23), a mechanism maintaining mRNA concentration homeostasis (24) irrespective of changes in cellular volume and DNA content. We also extend the model to include protein dynamics. The advantage of our theory is that it leads to intuitive insight that cannot be easily obtained from stochastic simulations. In addition, our model provides a framework which can be easily extended to include more intricate biological phenomena.

Results

Model Specification. The dynamic biochemical processes described by our model are illustrated in Fig. 1*C* and are as follows:

1) In the prereplication stage of the cell cycle, two gene copies can independently switch from ON (G) to OFF state (G^*) with rate σ_b and from OFF to ON with rate σ_u^-. Gene-copy independence has been shown experimentally (19).
2) In the ON state, there is the production of nascent (unspliced mRNA) denoted by N which subsequently becomes mature (spliced; denoted by M) with rate k. We assume that the maturation time is exponentially distributed rather than a deterministic time delay as assumed in previous studies (18, 19). This assumption is used because it makes the model analytically tractable (since then the model is Markov) and also since it has been shown (28) that the distribution of nascent mRNA numbers is insensitive to whether the delay is deterministic or exponentially distributed with identical mean maturation time, provided that $\rho_u \gg \sigma_b, \sigma_u$, which is the case experimentally (Table 1).
3) The mature mRNA decays with rate d via first-order kinetics, which is a common assumption supported by experiments (29).
4) Growth-dependent transcription whereby the transcription rate is proportional to the volume of the cell—this models the mRNA-concentration homeostasis mechanism reported in ref. 23. We assume that the cell volume increases exponentially with cell age t

Table 1. Kinetic parameters reported in various experimental papers

Cell type (gene)	σ_u(min^{-1})	σ_b(min^{-1})	ρ_u (min^{-1})	d (min^{-1})	Burst size (ρ_u/σ_b)	Fraction ON time ($\sigma_u/(\sigma_u+\sigma_b)$)	Timescale ratio δ	Reference
Yeast (POL1)	0.0700	0.680	2.00	0.0693	2.9	0.093	9.71	(11)
Yeast (PDR5)	0.3000	5.300	11.30	0.0495	2.1	0.054	17.67	(11)
Mouse embryonic stem cells (Oct4)	0.0092	0.018	1.90	0.0023	105.6	0.338	1.96	(19)
Mouse embryonic stem cells (Nanog)	0.0019	0.007	0.80	0.0022	115.9	0.216	3.63	(19)
Mouse embryonic stem cells (Nanog)	0.0282	0.609	2.11	0.0025	3.5	0.044	21.60	(26)
Human osteosarcoma (c-Fos)	0.1075	0.313	7.30	0.0462	23.4	0.256	2.91	(18)
Mouse hepatocytes (Acly)	0.0010	0.002	0.25	0.0004	129.3	0.337	1.97	(27)
Mouse hepatocytes (Actb)	0.0013	0.036	2.52	0.0004	70.1	0.034	28.77	(27)
Mouse hepatocytes (Srebf1)	0.0015	0.013	1.80	0.0022	137.7	0.102	8.82	(27)
Mouse hepatocytes (Insr1)	1.6×10^{-5}	3.5×10^{-4}	0.03	3.3×10^{-5}	81.0	0.045	21.00	(27)
Mouse fibroblasts (3,575 genes)	0.0022	0.236	0.69	0.0035	6.9	0.092	101.73	(10)
Human fibroblasts (1,609 genes)	0.2173 (d)	6.752 (d)	272.11 (d)	N/A	134.5	0.102	34.86	(10)
Mouse fibroblasts (16 genes)	0.0136	0.167	2.48	0.0109	17.2	0.074	20.11	(14)

The transcription rate (ρ_u), the mRNA degradation rate (d), the rate at which the gene switches from ON to OFF (σ_b), and the rate at which it switches from OFF to ON (σ_u) have been estimated from experimental data by using various models of stochastic mRNA dynamics (mostly using the standard model M2 in Fig. 1*A*). The data reveal that gene expression is bursty (gene is ON for short times, and, in that time, a large burst of mRNA produced). The large values of the ratio $\delta=\sigma_b/\sigma_u$ show that gene-inactivation events occur far more frequently than gene-activation events. Note that 1) estimates for the last three rows represent averages over genes; 2) human fibroblast estimates for σ_u, σ_b, ρ_u are in terms of d, and the latter was not measured; and 3) we do not report the value of σ_u before and after replication because the vast majority of studies do not take into account replication and, hence, report a single value. See *SI Appendix*, Section 2 for details. N/A, not applicable.

(where $0\leq t\leq t_d$ and t_d is the cellular interdivision time); this assumption is supported by experimental evidence for a variety of mammalian cells (30). Hence, the effective transcription rate follows the equation $\rho_u(t)=\rho_0\exp(\alpha t)$, where ρ_0 is the transcription rate at the start of the cell cycle, $\alpha=(1/t_d)\ln(V_f/V_0)$; V_0 is the cell volume at the beginning of the cell cycle; and V_f is the cell volume just before the cell divides. If there is no growth-dependent transcription, then $\rho_u(t)$ is a time-independent constant and corresponds to setting $\alpha=0$.

5) Replication results in a doubling of the gene copies from two to four at cell age t_r (replication time). We assume that this occurs instantaneously—i.e., replication occurs over a period which is much shorter than the length of the cell cycle. We shall refer to the gene which is replicated as the mother copy and the duplicated gene as the daughter copy. The daughter copy can either inherit the gene state from the mother copy (31), or else all copies (mother and daughter) are reset to the OFF state upon replication. One plausible explanation for the latter case is that to avoid the potential risk of conflict between replication and transcription (and the resulting DNA damage), it is highly likely that in the region where replication is actively ongoing or just completed, there is no transcription, indicating an OFF state (figure 2C in ref. 32).
6) Dosage compensation is modeled as a change in the value of the rate at which the gene switches from OFF to ON upon replication, specifically $\sigma_u=\sigma_u^-$ if $0\leq t<t_r$ and $\sigma_u=\sigma_u^+$ if $t_r\leq t\leq t_d$. This assumption can explain experimental data (19). Note that dosage compensation is another mechanism (besides growth-dependent transcription) which results in approximate mRNA concentration homeostasis (24) over the duration of the cell cycle (33).
7) Binomial partitioning of nascent and mature mRNA at cell division. We here assume that nascent and mature mRNA segregate independently of each other with a probability $1/2$ of ending up in one of the two daughter cells. The time between successive cell divisions is assumed to be fixed. This is a simplification, since a number of experiments show interdivision time variability (34, 35). The assumption of a fixed cell-cycle length is made to make the mathematical analysis of the model tractable. We will also show later how the theory can be modified to describe the effect of cell-cycle-length variability.

Approximate Solution of the Model. A master equation can be written which describes the exact stochastic dynamics of the above model with the replication and cell-division processes modeled via appropriate boundary conditions (see *SI Appendix*, section 1 for details). Given the myriad complex biological functions described by the model, it should come as no surprise that we were unable to find an exact solution to this master equation (indeed, much simpler models often cannot be solved exactly; see ref. 4 for a review of the state of the art in solutions of chemical master equations). Our approach will consist of breaking the model into submodels, where each considers only a subset of bioprocesses, followed by solving each submodel approximately and then integrating the results to obtain a solution to the full model.

We note from Table 1 that the vast majority of eukaryotic genes are characterized by a large value of the ratio $\delta=\sigma_b/\sigma_u$ (gene-inactivation rate divided by the gene-activation rate), i.e., genes spend most of their time in the OFF state. For the moment, we ignore the processes of cell division and replication and focus on nascent mRNA dynamics due to promoter switching, growth-dependent transcription, and maturation for a single gene copy. As we show in *SI Appendix*, Section 3.1, in this case for large δ, the generating function corresponding to the time-dependent marginal distribution of nascent mRNA numbers of a single gene $P(n_N,t)$ can be written (by a slight abuse of notation) as:

$$G(u,t)=\sum_{n_N=0}^{\infty}(1+u)^{n_N}P(n_N,t) \qquad [1]$$

$$=g(ue^{-kt})\left(\frac{\rho_0e^{-kt}u-\sigma_b}{\rho_0e^{\alpha t}u-\sigma_b}\right)^{\frac{\sigma_u}{\alpha+k}}.$$

Here, we have assumed that the initial marginal distribution of i nascent mRNA molecules is $P(i)=p_i$, which implies $g(u)=\sum_i p_i(u+1)^i$. It can be shown that Eq. **1** implies that for large δ, the stochastic reaction dynamics stemming from the combined processes of promoter switching and nascent mRNA production with a time-dependent transcription rate to account for growth-dependent transcription can be described by a simpler system of one effective reaction. In this reaction, nascent mRNA is produced at rate σ_u in bursts whose size are distributed according to a negative binomial distribution with a time-dependent mean burst size $(\rho_0/\sigma_b)\exp(\alpha t)$ (*SI Appendix*, Section 3.2). The major advantage of this effective reaction description is that it dispenses with an explicit gene-state description which considerably simplifies the calculations to follow. In particular, the issue of how to choose the gene state at the beginning of replication is circumvented—intuitively, this is possible because since the gene spends most of its time in the OFF state, the two mechanisms described in point 5 in Model Specification cannot be distinguished in practice.

$$G_A(u,t)=\begin{cases}\left(\dfrac{\rho_0 u e^{-kt}-\sigma_b}{\rho_0 u e^{\alpha t}-\sigma_b}\right)^{\frac{\sigma_u^-}{\alpha+k}} & t\in[0,t_r),\\ \left(\dfrac{\rho_0 u e^{-kt}-\sigma_b}{\rho_0 u e^{-k(t-t_r)+\alpha t_r}-\sigma_b}\right)^{\frac{\sigma_u^-}{\alpha+k}}\left(\dfrac{\rho_1 u e^{-k(t-t_r)}-\sigma_b}{\rho_1 u e^{\alpha(t-t_r)}-\sigma_b}\right)^{\frac{\sigma_u^+}{\alpha+k}} & t\in[t_r,t_d),\end{cases} \quad \textbf{[2A]}$$

$$G_B(u,t)=\begin{cases}1 & t\in[0,t_r),\\ \left(\dfrac{\rho_1 u e^{-k(t-t_r)}-\sigma_b}{\rho_1 u e^{\alpha(t-t_r)}-\sigma_b}\right)^{\frac{\sigma_u^+}{\alpha+k}} & t\in[t_r,t_d).\end{cases} \quad \textbf{[2B]}$$

Next, we include the processes of replication and dosage compensation. Specifically, let $G_A(u,t)$ and $G_B(u,t)$ be the generating functions describing the dynamics of nascent mRNAs born within a cell cycle for a single mother and daughter copy, respectively. By using Eq. **1**, it follows that the generating functions of nascent mRNA produced by mother and daughter copies are piece-wisely defined and given by Eqs. **2A** and **2B**. The first part $t\in[0,t_r)$ of $G_A(u,t)$ describes the stochastic dynamics of nascent mRNA born in the prereplication time. Note that the initial condition is zero, and $\sigma_u=\sigma_u^-$. The second part $t\in[t_r,t_d)$ of $G_A(u,t)$ describes the stochastic dynamics of nascent mRNA born in the postreplication time. This is given by Eq. **1**, with g replaced by the initial condition which is the generating function at replication time (from the expression for $t\in[0,t_r)$); also note that $\sigma_u=\sigma_u^+$ (due to dosage compensation), ρ_0 is replaced by $\rho_1=\rho_0 e^{\alpha t_r}$ since this is the transcription rate at replication time, and time t is replaced by $t-t_r$. Note that, intuitively, the second part $t\in[t_r,t_d)$ of $G_A(u,t)$ (which describes postreplication dynamics) can be written as a product of two factors because there is independence between nascent mRNA inherited from the prereplication stage and the nascent mRNA born in the postreplication stage. Since there is no transcription activity in the prereplication time for the daughter copy, the generating function $G_B(u,t)$ is trivially equal to 1 for $t\in[0,t_r)$. The second part $t\in[t_r,t_d)$ of $G_B(u,t)$ can be found similarly as for $G_A(u,t)$. Note that the individual factors in the generating-function expressions can be written as a product of the generating functions for the binomial and negative binomial distributions (*SI Appendix*, section 3).

Finally, we add the details of cell division and the associated binomial partitioning. There are two processes contributing to the number of mRNAs at a particular cell age t of a given cell cycle n: 1) the decay of mRNAs inherited from the previous cycle, and 2) the production of new mRNAs in cell cycle n. These processes are independent from each other when the gene spends most of its time in one state (as in our case), and, hence, it follows that we can write:

$$\begin{aligned}P_n(n_N,t)&=\frac{1}{n_N!}\frac{d^{n_N}G_n(u,t)}{du^{n_N}}\bigg|_{u=-1},\\ G_n(u,t)&=\underbrace{G_n(ue^{-kt},0)}_{\text{death process}}\underbrace{G_A^2(u,t)G_B^2(u,t)}_{\text{new born mRNA}}\ \forall t\in[0,t_d),\end{aligned} \quad \textbf{[3]}$$

where $P_n(n_N,t)$ is the marginal distribution of nascent mRNA numbers at time t in cell cycle n. Note that the power of 2 on the right-hand side of Eq. **3** arises from the diploidy of gene copies and from assuming that they are independent of each other (as mentioned in point 1 of Model Specification). Binomial partitioning at cell age t_d leads to a relationship between the initial conditions for the generating function of the n^{th} cycle and the generating function of the $n-1^{\text{th}}$ cycle at time t_d, which can be shown (*SI Appendix*, section 3.3) to lead to the condition:

$$G_n(u,0)=G_{n-1}(\eta u,0)G_A^2\left(\frac{u}{2},t_d\right)G_B^2\left(\frac{u}{2},t_d\right),\ n>1 \quad \textbf{[4]}$$

where $\eta=e^{-kt_d}/2$. Note that $G_1(u,0)$ is the initial condition at the beginning of the first cycle, and this is a user-input condition; for all results in this paper, we assumed $G_1(u,0)=1$, implying no nascent mRNA initially.

Hence, summarizing Eqs. **3** and **4** together with Eqs. **2A** and **2B** provides an approximate time-dependent solution of the marginal distribution of nascent mRNA numbers valid for all cell ages and cellular generations; the assumption behind our derivation was that the gene spends most of its time in the OFF state. Note that while the derivation assumed a growth-dependent transcription rate as described in point 4 of Model Specification, nevertheless, it is straightforward to derive similar results for a general time-dependent transcription rate (*SI Appendix*, section 10); this may be useful to describe synthetic gene-regulatory networks where the transcription rate can be arbitrarily regulated over time.

In *SI Appendix*, section 4, we show that an approximate time-dependent solution of the marginal distribution of mature mRNA numbers can be similarly derived, provided that one further assumes that the timescales of nascent mRNA are much shorter than those of mature mRNA. The timescales of these two types of mRNA are determined by k and d, respectively (their elimination

rates), and there is strong evidence showing $k \gg d$. For the gene c-Fos in human osteosarcoma cells, Senecal et al. (18) estimated a value of $k = 1.25\ \text{min}^{-1}$, $d = 0.0462\ \text{min}^{-1}$, while for Nanog and Oct4 in mouse embryonic stem cells, Skinner et al. (19) estimated $k = 0.13\ \text{min}^{-1}$, $d = 0.0022\ \text{min}^{-1}$, and $k = 0.29\ \text{min}^{-1}$, $d = 0.0023\ \text{min}^{-1}$, respectively. Hence, k is considerably larger than d generally.

Specifically, we show using singular perturbation theory that if the gene spends most of its time in the OFF state ($\delta \gg 1$) and if nascent mRNA maturation is fast ($k \gg d$), then the marginal distribution of mature mRNA numbers within a single cell is approximately given by Eqs. **3** and **4** together with Eqs. **2A** and **2B**, with k replaced by d and with the change of variable $n_N \to n_M$. Note that under the fast-maturation assumption, the dynamics of nascent mRNA do not affect the dynamics of mature mRNA; for a detailed discussion, see *SI Appendix*, section 4.

Numerical Evaluation of the Accuracy of the Approximate Distributions. A main assumption behind our derivation is that $\delta \gg 1$ holds for many genes, but clearly this is not the case for all (Table 1). It is unclear how large δ has to be for our approximation to be accurate. Hence, we next test the accuracy of our theory by fixing $k \gg d$ (this assumption holds for all genes that we could find data for; *SI Appendix*, section 2) and varying parameters ρ_u and σ_b (relative to the rest which are fixed) to vary δ at constant mean burst size ρ_u/σ_b for the case of no growth-dependent transcription ($\alpha = 0$). We then quantify the accuracy of our approximation by calculating the Hellinger distance (HD) between the approximate probability distribution of nascent and mature mRNA numbers and the exact numerical solution of the chemical master equation as a function of δ, which is varied over two orders of magnitude from 2 to 200. Note that the HD has the properties of being symmetric and satisfies the triangle inequality, thus implying that it is a distance metric on the space of probability distributions (unlike, e.g., the commonly used Kullback–Leibler divergence); it is also conveniently a fraction, which makes for easy interpretation.

Our approximate theory was derived by using the assumption that $\delta \gg 1$, and, hence, we expect the accuracy of the theory to increase with δ. This is verified in the heatmap shown in Fig. 2*A*, where it is shown that the error in our approximate theory is inversely proportional to δ; is a weak function of absolute time, i.e., independent of cell age and generation; and is generally small (HD $\ll 1$) for both nascent (Fig. 2 *A*, *Upper*) and mature (Fig. 2 *A*, *Lower*) mRNA. The exact probability distributions for two time points are compared with the approximate distributions in Fig. 2*B* for three different values of δ. Note that the match between approximate analytic solution and the exact solution (using stochastic simulation algorithm [SSA]) is excellent for $\delta = 20$ and acceptable for $\delta = 5$ for all times; for smaller values of δ, the nascent mRNA distribution still is well approximated, but the same cannot be said for the mature mRNA distribution. Hence, our theory is accurate for the majority of genes reported in Table 1.

In Fig. 2*C*, we compare the time-dependent mean and variance predicted from the approximate theory with the exact result (from SSA), showing the accuracy of the theory to capture the cyclic behavior of the moments due to the dynamic processes of replication and cell division. The theory's accuracy remains high, even when growth-dependent transcription is turned on $\alpha > 0$ (*SI Appendix*, Fig. S2).

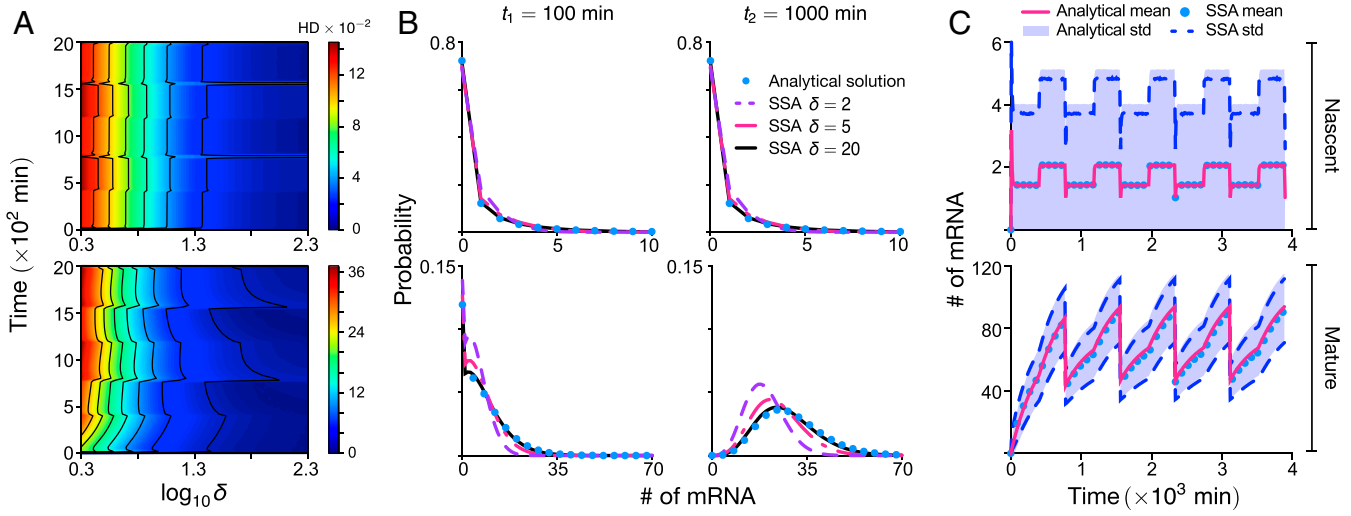


Fig. 2. Accuracy of the approximate analytic solution to the stochastic model of eukaryotic gene expression. The approximate probability-distribution solution for nascent mRNA numbers is given by Eqs. **3** and **4** together with Eqs. **2A** and **2B**; the distribution for mature mRNA numbers is given by the same equations with *k* replaced by *d* and n_N replaced by n_M. Here, we investigate the accuracy of this solution relative to the exact solution, which is numerically computed by using the Finite State Projection algorithm (FSP) or the SSA; note that these simulations are of the full model (without any approximation) as described in Model Specification and *SI Appendix*, section 1. *Upper* and *Lower* show information for nascent and mature mRNA, respectively. (*A*) The heatmap for the HD (which quantifies the distance between the approximate and exact marginal distributions) for both nascent and mature mRNA as a function of absolute time and the dimensionless timescale ratio δ. Note that $\delta = \sigma_b/\sigma_u^-$ is here varied by changing σ_b while keeping $\sigma_u^- = 0.0282\ \text{min}^{-1}$ and the mean burst size $\rho_u/\sigma_b = 3.5$ constant; the rest of the parameter values are the same as those in Fig. 1. Due to the computational demand of producing a heatplot, the exact solution is here computed by using FSP; computations using the SSA at random points in the heatplot were indistinguishable from those using FSP. (*B*) Marginal distributions of the nascent and mature mRNA for two time points and three values of δ compared to those obtained from stochastic simulations using the SSA. (*C*) Plots of the mean and SD (std) versus time as predicted by the approximate analytical solution and exact stochastic simulations for $\delta = 21.60$ using the SSA (which also agree with those using FSP). All plots show that the accuracy of the distributions and moments predicted by the approximate analytic solution is high provided δ is not too small (larger than about five). Note that $\alpha = 0$ in all panels, meaning that there is no growth-dependent transcription.

Effect of Cell-Cycle-Length Variability on mRNA Distributions. Our theory assumes a fixed cell-cycle length and synchronized cell cycles among cells. This is the case when cells are subjected to certain environmental conditions (36, 37), when the circadian clock gates the cell cycle (38, 39), and during certain phases of morphogenesis (40). However, variation in the cell-cycle length is likely common (e.g., in Fig. 3*A*, we show experimental data for mouse fibroblasts), which leads to asynchronous behavior. We modified the SSA such that the cell-cycle times t_d are assumed to be random variables independently drawn from an Erlang distribution (which well approximates the experimental data in Fig. 3*A*) and the replication time is exactly in the middle of each cycle. Each trajectory of the algorithm corresponds to a forward lineage, i.e., either of the two daughter cells is followed with equal probability. The distribution of the number of mature mRNAs is constructed from an ensemble of these single-cell trajectories; this distribution is shown as blue dots in Fig. 3*B*, where the parameters are those measured for two mouse genes. The mature mRNA distribution can also be predicted by modifying our theory to take into account asynchronous cell cycles, but keeping the assumption of fixed cell-cycle length (*SI Appendix*, section 6); this prediction is shown as a red curve in Fig. 3*B*. Excellent agreement between theory and the SSA is found for 16 mouse genes (two are shown in Fig. 3*B* and the rest in *SI Appendix*, Fig. S5). As shown in *SI Appendix*, section 6, the implicit reason for the accuracy of the modified theory is the fact that the mRNA lifetime in mouse fibroblasts is typically much less than the average cell-cycle length; i.e., rapid degradation averages out timing fluctuations, which is in line with other recent studies (41).

Sensitivity Analysis of the Coefficient of Variation of mRNA Fluctuations. An important use of the analytic results is that we can efficiently calculate the sensitivity of the coefficient of variation of mature mRNA (SD divided by the mean) to small perturbations in the eight parameters of the model. To this end, we first used our approximate theory to calculate closed-form expressions for the cyclo-stationary mean and variance of mature mRNA, which we denote as $\overline{\langle n_M \rangle}_t$ and $\overline{\sigma}^2_{n_M,t}$, respectively (*SI Appendix*, section 5). Note that the cyclo-stationary conditions ensue in the limit of biological steady-state growth (20), which is achieved when the probability that a cell of age t has a given number of molecules of certain species is independent of which generation it belongs to—i.e., setting $G_{n+1}(u,t) = G_n(u,t)$.

The cyclo-stationary coefficient of variation of mature mRNA noise averaged over the cell cycle is then given by $\overline{\mathrm{CV}} = t_d^{-1} \int_{t=0}^{t_d} \sqrt{\overline{\sigma}^2_{n_M,t}} / \overline{\langle n_M \rangle}_t \, dt$ (which is computed numerically over 100 discrete time points evenly distributed within a cell cycle). The relative sensitivity of $\overline{\mathrm{CV}}$ to a parameter r is then given by $\Lambda_r = (r/\overline{\mathrm{CV}}) \partial \overline{\mathrm{CV}} / \partial r$ (43), meaning that a 1% change in the value of parameter r leads to Λ_r% change in $\overline{\mathrm{CV}}$. We next computed the relative sensitivities for eight of the rate parameters for a large

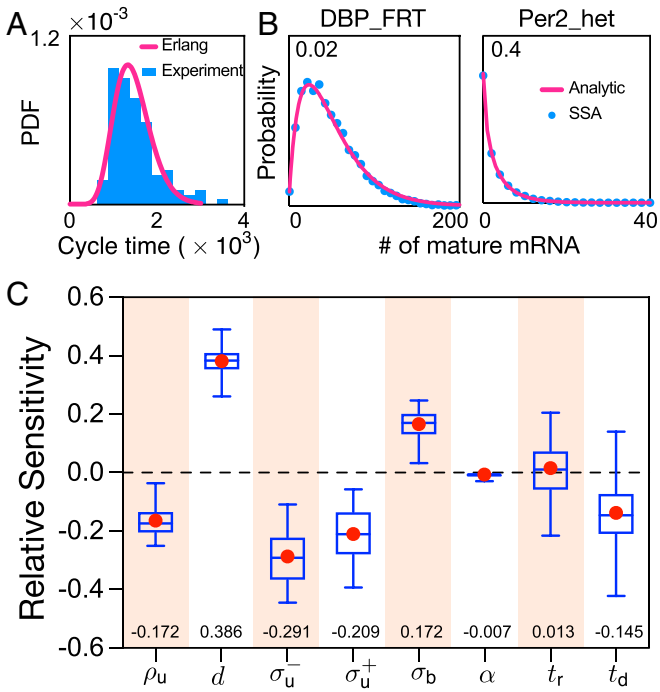


Fig. 3. (*A* and *B*) Cell-cycle-length variability and its effect on mature mRNA distributions. (*A*) The Erlang distribution provides a good fit to the experimentally measured cell-cycle time distribution of NIH 3T3 cells in ref. 42. PDF, probability distribution function. (*B*) The SSA modified such that the cell-cycle times are random variables independently drawn from an Erlang distribution is used to obtain the mature mRNA distributions (for the genes DBP_FRT and Per2_het reported in *SI Appendix*, Table S1) measured across an ensemble of cells (blue dots; see main text for details). The mature mRNA distributions are accurately predicted by modifying our theory (red solid lines; *SI Appendix*, section 6, Eq. 26) to take into account the asynchronicity of cell cycles across the population. Note that replication always occurs in the middle of a cell cycle. (*C*) Relative sensitivity of the cyclo-stationary coefficient of variation of mature mRNA noise averaged over the cell cycle to eight parameters. Box plots indicate the median and the 25% and 75% quantiles, with the mean marked as red dots. The median relative sensitivities are also shown as numbers at the bottom of the plot. The sensitivity analysis is carried out on 567,540 parametric combinations estimated for 1,051 genes of CAST allele data of mouse embryonic stem cells. Our results show that the degradation rate d is the most sensitive parameter, followed by σ_u^-, while α is the least sensitive one. Note that there is no dependence of the coefficient of variation of mature mRNA on k in the approximate theory, and, hence, k is not a relevant parameter. See *SI Appendix*, section 11 for the specific range of parameter choice and their justification.

number of genes in mouse embryonic stem cells (10). Note that the 1,051 genes selected for this analysis are characterized by a timescale ratio $\delta \geq 5$, a value which is large enough to guarantee the accuracy of our approximate analytic solution (Fig. 2), which is used to calculate the coefficient of variation. The results illustrated by using box plots in Fig. 3*C* show that the most sensitive parameters were the mRNA degradation rate d and the dosage-compensation parameters σ_u^-, σ_u^+, while the least sensitive parameters were the growth-dependent transcription parameter α and the replication time t_τ.

Effective Negative Binomial Approximation of Mature mRNA Distributions. While our theory gives approximate distributions of nascent and mRNA molecule numbers, these distributions are complex and cannot be easily written in terms of known simple distributions. It has been frequently observed that many measured number distributions can be easily fit by the negative binomial distribution (or its continuous analog the gamma distribution). Indeed, this is one of the major reasons why the two-state model of mRNA dynamics (model M2 in Fig. 1) has become widely adopted, since in the bursty limit, the probability distribution of molecule numbers is negative binomial. Hence, we next investigate whether our approximate distributions can also be well fit by negative binomial distributions.

Assuming a negative binomial distribution (NB) for the number of molecules of mature mRNA, $P(n_M) \sim \mathrm{NB}(r, p)$ with parameters r and p, its mean and variance are given by $\langle n_M \rangle_e = rp/(1-p)$, $\sigma^2_{n_M,e} = rp/(1-p)^2$. We equate these two moments to the cyclo-stationary moments from our theory: Specifically, we set $\langle n_M \rangle_e = \overline{\langle n_M \rangle}_t/2$ and $\sigma^2_{n_M,e} = \overline{\sigma}^2_{n_M,t}/2$, where the factor of $1/2$ accounts for the two independent gene copies in our model. One can then find simple expressions for r and p:

$$r = \frac{1}{2}\frac{\overline{\langle n_M \rangle}_t^2}{\overline{\sigma}^2_{n_M,t} - \overline{\langle n_M \rangle}_t}, \quad p = 1 - \frac{\overline{\langle n_M \rangle}_t}{\overline{\sigma}^2_{n_M,t}}.$$

Hence, we have constructed a negative binomial approximation $\mathrm{NB}(r, p)$ to our model's predicted distribution of the mature mRNA number distribution at cell age t in the cyclo-stationary limit. We test the accuracy of this approximation in Fig. 4. In Fig. 4*A*, we compute the HD distance between the distribution solution of our model (as computed by using Eqs. **2A**, **2B**, **3**, and **4** with k changed to d and n_N changed to n_M) and the negative binomial approximation for 21 time points in the cell cycle using parameters for 1,051 genes in mouse embryonic stem cells (same as used for sensitivity analysis). Remarkably, we find the HD to be much <1 for all genes and all cell ages, implying that the negative binomial approximation is an excellent one in practice. Given the well-known fact that the negative binomial is a good approximation to the steady-state solution of the standard two-state model of mRNA dynamics in bursty conditions (8), in an indirect sense, our results in Fig. 4*A* also show that the cyclo-stationary distribution of our model at a particular cell age can be well approximated by the steady-state distribution of the two-state model (for a particular choice of effective parameters). This argument is further reinforced in Fig. 4*B*, where we show that the HD averaged over one cycle for a particular gene is roughly linearly dependent with Θ in log space, where Θ is the absolute difference between the uncentered third moments of the two-state model and its negative binomial approximation. Note that Θ is defined in terms of the parameters ρ_u, σ_u^-, and d specific to a particular gene:

$$\Theta = \frac{2\hat{\rho}_u^3 \hat{\sigma}_b \hat{\sigma}_u^- (1+\hat{\sigma}_u^-)}{(\hat{\sigma}_b + \hat{\sigma}_u^-)^2 (1+\hat{\sigma}_b+\hat{\sigma}_u^-)^2 (2+\hat{\sigma}_b+\hat{\sigma}_u^-)}, \qquad [5]$$

where $\hat{\rho}_u = \rho_u/\hat{d}$, $\hat{\sigma}_u^- = \sigma_u^-/\hat{d}$, $\hat{\sigma}_b = \sigma_b/\hat{d}$, and $\hat{d} = d + \ln 2/t_d$. For a derivation of this expression, see *SI Appendix*, section 7. Eq. **5** shows that the error in the negative binomial approximation is inversely proportional to the rate of promoter switching and directly proportional to the transcription rate. In Fig. 4*C*, we show the full versus effective negative binomial distributions as a function of cell age for the gene with the largest HD (Ndufa4)—even in this extreme case, the two distributions cannot be distinguished by eye, thus showing the high accuracy of the negative binomial approximation to our model for a large number of genes in mouse embryonic stem cells.

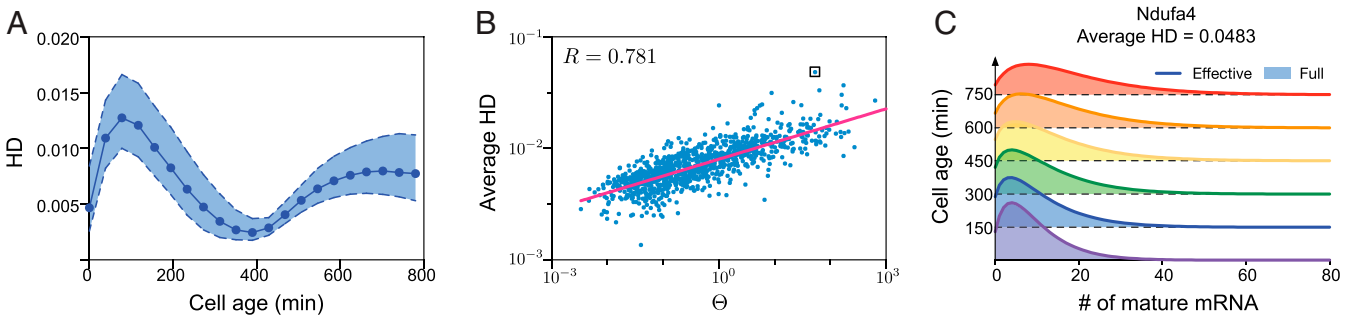


Fig. 4. Effective negative binomial approximation for the mature mRNA number distribution of our model. The approximation is obtained by matching the cyclo-stationary mean and variance of our model to those of an effective negative binomial distribution. (*A*) We calculate the HD between the effective distribution and the full distribution for 21 equidistant time points through a whole cell cycle. Each point corresponds to the HD median for 1,051 genes in mouse embryonic stem cells (same data used for Fig. 3*C*), whereas the broken lines show the 25% and 75% quantiles. Note that the four parameters t_d = 780 min, $\alpha = 0$, t_r = 400 min, and $\sigma_u^+ = 0.7\sigma_u^-$ are the same for all genes. (*B*) A plot of the HD for each gene averaged over 21 time points in the cell cycle versus the index Θ; the two quantities are linearly correlated in log space with Pearson correlation coefficient $R = 0.781$. (*C*) The matching of the effective and full distributions in time for the gene Ndufa4, which has the largest HD in *B* (shown as a boxed point).

Including Protein Dynamics. Thus far, the model only describes the stochastic mRNA dynamics. Given the increasing number of single-cell measurements of protein expression in eukaryotic cells (44–46), we next extend our theory to provide expressions for the protein distributions.

$$G_{\mathrm{A}}(u,t)=\begin{cases}\left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}t}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{-}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}}, & t\in[0,t_{\mathrm{r}}),\\ \left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}t}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}(t-t_{\mathrm{r}})}-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{-}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}}\left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}(t-t_{\mathrm{r}})}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{+}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}} & t\in[t_{\mathrm{r}},t_{\mathrm{d}}),\end{cases} \quad [6A]$$

$$G_{\mathrm{B}}(u,t)=\begin{cases}1 & t\in[0,t_{\mathrm{r}}),\\ \left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}(t-t_{\mathrm{r}})}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{+}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}} & t\in[t_{\mathrm{r}},t_{\mathrm{d}}).\end{cases} \quad [6B]$$

The model has the same seven features as described at the beginning of Results, but with an additional two features: 1) mRNA is translated into protein at rate λ; and 2) protein decay occurs with rate d_{p}. Both reactions are assumed to obey first-order kinetics (6). Again, we need to make some approximation to proceed further: 1) We assume that the timescale ratio δ is large; and 2) we assume that the timescales of nascent and mature mRNA dynamics are much shorter than those of protein. The first assumption we know is satisfied for a large number of genes. The second assumption can be justified as follows. The timescales of nascent mRNA, mature mRNA, and protein are approximately given by the inverse of the elimination rates of each, i.e., k, d, and d_{p}, respectively. Now, as we saw earlier, $k\gg d$. Also, Schwanhäusser et al. (47) report that the median mRNA decay rate d is approximately five times larger than the median protein decay rate d_{p} for NIH 3T3 mouse fibroblasts (calculated over 4,200 genes); the cumulative distribution of the ratio of the two decay rates is shown in *SI Appendix*, Fig. S7. Similarly, for 1,962 genes in budding yeast, the median of the ratio of the mRNA decay rate to protein decay rate is approximately three (3). Hence, for a substantial number of genes, the assumption $k\gg d\gg d_{\mathrm{p}}$ holds, and that implies that protein dynamics occurs over a much slower timescale than both nascent and mature mRNA dynamics. Given assumptions 1 and 2, we can show using perturbation theory applied to the master equation of the model (*SI Appendix*, section 8) that the temporal protein distribution is given by Eqs. **3** and **4** with the replacements $k\to d_{\mathrm{p}}$ and $n_{\mathrm{N}}\to n_{\mathrm{P}}$ together with the generating functions given by Eqs. **6A** and **6B**, where $b=\lambda/d$ is the translational burst size quantifying the mean number of protein produced during the lifetime of mature mRNA. Note that this derivation is for the case of no growth-dependent transcription, i.e., $\alpha=0$.

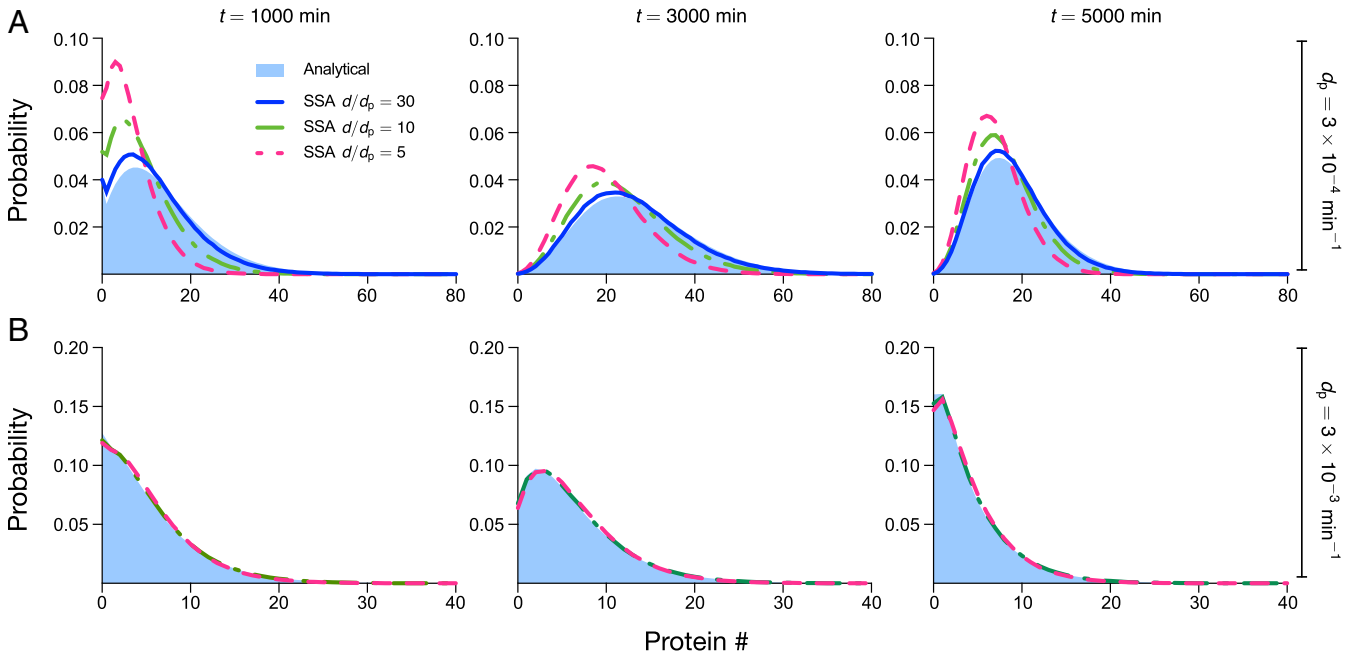


Fig. 5. Comparison of protein distributions predicted by theory (under the assumption of slow protein dynamics) and stochastic simulations using the SSA. (*A*) Using parameters typical of mammalian cells, we find that our theory agrees well with simulations for $d/d_{\mathrm{p}}=30$, is acceptable for $d/d_{\mathrm{p}}=10$, and performs poorly for $d/d_{\mathrm{p}}=5$; agreement tends to be better with increasing time, but accuracy is mostly determined by d/d_{p}. Note that the protein lifetime is $\ln 2/d_p$ = 2,310 min, the cell-cycle duration is t_d = 1,560 min, and the mRNA lifetime is $\ln 2/d=462$, 231, and 77 min for $d/d_{\mathrm{p}}=5$, 10, 30, respectively. Given the value of t_{d}, we have that t = 1,000; 3,000; 5,000 min corresponds to cells in generations one, two, and three, respectively. (*B*) Parameter values as in *A*, except that the protein, mRNA decay, and translation rates are multiplied by 10. Note that the ratio d/d_{p} is unchanged from *A*, but both protein and mRNA lifetimes are now 10 times smaller, meaning that they are significantly less than the cell-cycle time. This condition leads to significantly improved agreement between theory and simulations, such that they are indistinguishable for $d/d_{\mathrm{p}}=5$, 10, 30. See *SI Appendix*, section 12 for the choice of parameters.

The accuracy of the approximation is tested via stochastic simulations in Fig. 5*A* using parameters typical of mammalian cells (47). The protein distribution obtained from theory (shown as solid blue) is compared with SSA results for the protein distribution at three different times and three different values of the ratio of mRNA to protein-decay rates d/d_p. The discrepancy between theory and simulations decreases as d/d_p increases (as expected) and is particularly good for $d/d_p > 10$. There is also a small increase in accuracy of the theory with increasing absolute time, though the major determinant is the decay ratio. Simulations show that the accuracy of the theory increases if we increase the protein and mRNA decay rates while keeping their ratio constant—in particular, whenever the mRNA lifetime is much less than the protein lifetime and when the latter is less than the cell-cycle length, then the agreement with theory is excellent for a wide range of values of the ratio d/d_p (compare Fig. 5 *A* and *B*). Similarly, the accuracy of the theory also increases if we increase the cell-cycle length at constant ratio d/d_p [note that the limit of infinite cell-cycle length corresponds to the conventional case where partitioning due to cell division is not explicitly taken into account (3)]. Hence, summarizing, our expressions for the approximate protein distributions are accurate whenever $d/d_p \gg 1$ and $d_p t_d > 1$. From *SI Appendix*, Fig. S7, it can be deduced that about 20% of proteins in mammalian cells satisfy $d/d_p > 10$; also, analysis of the dataset in ref. 47 shows that only 30% of all proteins in mammalian cells have decay lifetimes less than the cell-cycle length. Hence, while timescale separation between mRNA and protein has played an important role in the development of reduced stochastic models of gene expression in bacteria and yeast (3), it appears that the same technique should be used with care when developing reduced models of stochastic gene expression in mammalian cells.

The model can also be further extended to include bimolecular gene–protein interactions, which are common in nature (48). Specifically, we consider the case of negative feedback mediated by an auto-regulatory motif, whereby the transition from the ON to OFF state of all gene copies (i.e., the two copies prereplication and the four copies postreplication) is mediated by protein binding to the gene. Using a recently developed technique, the linear mapping approximation (LMA) (49), we show in *SI Appendix*, section 9 how the distributions of mRNA and protein for the model with no feedback (derived earlier) can be used to construct approximate distributions for the model with feedback.

The LMA is based on two assumptions: 1) a conditional mean-field approximation which equates to assuming small protein fluctuations compared to the mean number of proteins when the promoter is unbound; and 2) a time-averaging assumption which corresponds to the first term of the Magnus expansion of the time-dependent solution of the master equation and which is uniformly valid in time, provided the protein–gene binding rates are not too large. The approximation error tends to be dominated by assumption 2. This, however, is typically small, as we show in Fig. 6, where it is clear that the LMA accurately captures the effect of negative feedback on the mature mRNA and protein distributions for both prereplication and postreplication times in the cell cycle.

Discussion

In this article, we have developed a model of gene expression in eukaryotic cells which includes a high level of biological detail compared to previous models in the literature, while remaining analytically tractable. Specifically, our model takes into account gene

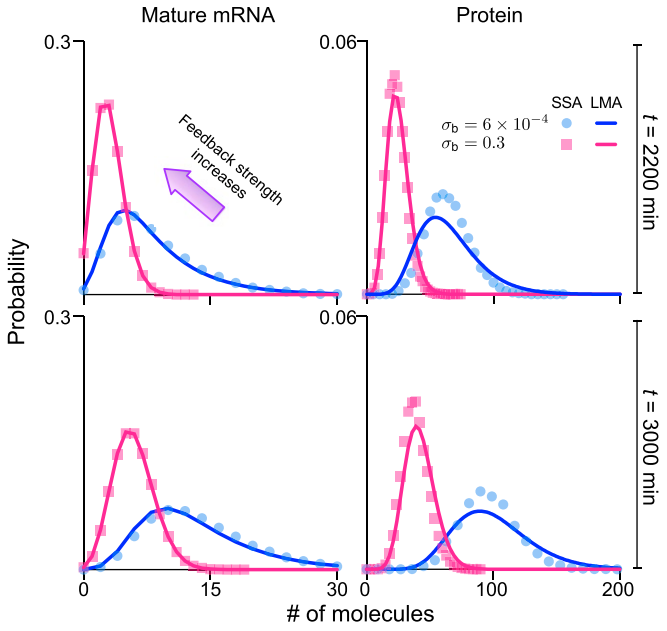


Fig. 6. LMA solution of the stochastic model with a negative-feedback auto-regulatory mechanism. The predicted distributions for mature mRNAs and proteins for time t = 2,200 min (prereplication) and t = 3,000 min (postreplication) of generation 2 are calculated from the equations for the generating functions in *SI Appendix*, section 9. LMA predictions for mature mRNA distributions agree with SSA results with remarkable accuracy, whereas the predictions for protein distributions agree with the SSA to an acceptable accuracy. Note that the ratio of decay rates d/d_p is five (the median reported in *SI Appendix*, Fig. S7), so that the timescale separation assumption is marginally satisfied. The graphs show that increasing feedback strength σ_b from 6×10^{-4} to 0.3 (500-fold change) substantially reduces the number of mature mRNAs and proteins. Specifically, the cell-cycle duration t_d = 1,560 min is selected to be 26 h, close to the data reported for NIH 3T3 in the supplementary information of ref. 47, the gene replication $t_r = 800$ min occurs roughly in the middle of the cell cycle. The decay rates of mature mRNA and protein are 0.005 and 0.001 min^{-1}, respectively (the half-lives are 2.3 and 11.5 h, respectively) and, hence, within the range for the same cell line reported in figure 2C of ref. 47. The other kinetic parameters are chosen as $\rho_u = 0.15$ min^{-1}, $\rho_b = 0.0075$ min^{-1}, $k = 10$ min^{-1}, $\sigma_u^- = 0.003$ min^{-1}, $\sigma_u^+ = 0.71\sigma_u^-$, $\lambda = 2d$, and $\alpha = 0$. See *SI Appendix*, section 9 for details.

replication, binomial partitioning due to cell division, dosage compensation, growth-dependent transcription, promoter switching, and the translation of mature mRNA into proteins. The model also provides a description of both nascent and mature mRNA distributions, which is necessary to make sense of high-resolution experimental data (11). We have shown that by breaking this complex model into a set of simpler submodels, solving each submodel approximately using timescale-separation methods (50), and then integrating the results together, it is possible to derive closed-form expressions for the time-dependent distribution of the numbers of nascent mRNA, mature mRNA, and protein inside a single cell. Specifically, we have made use of two assumptions: 1) Nascent mRNA dynamics is much faster than mRNA dynamics, which itself is much faster than protein dynamics; and 2) the gene-inactivation rate is much larger than the gene-activation rate. We have provided experimental evidence that these assumptions are reasonable for a large number of genes in several different types of eukaryotic cells grown under different conditions, and, hence, our model provides a detailed quantitative model of eukaryotic gene expression. A major advantage of our analytic approach is that, despite the biological complexity described by the model, it leads to simple distributions (negative binomial) for the molecule numbers for all cell ages and generations. It also provides a quantitative description for both nascent and mature mRNA dynamics, both of which are measurable observables. A description in terms of the two is advantageous, since nascent mRNA closely reflects the kinetics of transcription, while mature mRNA reflects additional processes downstream of transcription.

Numerical evaluation of these distributions is far more computationally efficient than direct simulation using the SSA. This implies that the model's behavior can be easily predicted across vast swaths of parameter space and that usually prohibitive tasks, such as stochastic sensitivity analysis, can be straightforwardly performed. Indeed, using our theory, we calculated the sensitivity of the coefficient of variation of noise (averaged over the cell cycle and in the cyclo-stationary limit) to small changes in the parameter values measured for mammalian cells. The parameters ordered according to the magnitude of their sensitivities are mRNA degradation rate, the rate of gene activation (before and after replication), rate of gene inactivation, transcription rate, cell-division time, replication time, and the parameter determining the coupling between transcription rate and cell growth. This suggests that variations in the values of the mRNA degradation rates and of the promoter-switching rates across cells are among the most significant sources of variability in gene expression across a population of cells (what is often termed extrinsic noise). Another major advantage of our closed-form expressions for the time-dependent distributions, and, in particular, their approximation by negative binomial distributions, is that they can be used to obtain the likelihood of a set of experimental observations of the molecule numbers—the likelihood can then be used within a Markov chain Monte Carlo algorithm to obtain the posterior distributions of parameters (51, 52).

Our model cannot resolve the effects of polymerase and transcription-factor fluctuations on mRNA and protein dynamics (53). It is also the case that our model cannot take into account the effect of cell-cycle-length variability on the distribution of protein numbers because the low protein degradation rates do not average out timing fluctuations (54). Lastly, the sharing of resources can potentially modify many cellular processes (55). Future work will involve extensions of the model to include these and other salient features of single-cell biology.

Data Availability. The data used in the paper are described in *SI Appendix*, section 2. The simulation code is available from the corresponding author upon request.

ACKNOWLEDGMENTS. Z.C. was supported by the UK Research Councils' Synthetic Biology for Growth program, the Biotechnology and Biological Sciences Research Council (BBSRC), the Engineering and Physical Sciences Research Council, and Medical Research Council Grant BB/M018040/1. R.G. was supported by BBSRC Grant BB/M025551/1. R.G. thanks Sara Buonomo and Peter Swain for useful discussions and insightful feedback.

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
2. J. Paulsson, Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).
3. V. Shahrezaei, P. S. Swain, Analytical distributions for stochastic gene expression. *Proc. Acad. Natl. Sci. U.S.A.* **105**, 17256–17261 (2008).
4. D. Schnoerr, G. Sanguinetti, R. Grima, Approximation and inference methods for stochastic biochemical kinetics: A tutorial review. *J. Phys. A* **50**, 093001 (2017).
5. N. Battich, T. Stoeger, L. Pelkmans, Control of transcript variability in single mammalian cells. *Cell* **163**, 1596–1610 (2015).
6. N. Maheshri, E. K. O'Shea, Living with noisy genes: How cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 413–434 (2007).
7. I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036 (2005).
8. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
9. R. D. Dar *et al.*, Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Acad. Natl. Sci. U.S.A.* **109**, 17454–17459 (2012).
10. A. J. M. Larsson *et al.*, Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
11. D. Zenklusen, D. R. Larson, R. H. Singer, Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* **15**, 1263–1271 (2008).
12. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
13. J. Peccoud, B. Ycart, Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
14. D. M. Suter *et al.*, Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
15. N. Molina *et al.*, Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Acad. Natl. Sci. U.S.A.* **110**, 20563–20568 (2013).
16. L.-h. So *et al.*, General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* **43**, 554–560 (2011).
17. M. Kaern, T. C. Elston, W. J. Blake, J. J. Collins, Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
18. A. Senecal *et al.*, Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep.* **8**, 75–83 (2014).
19. S. O. Skinner *et al.*, Single-cell analysis of transcription kinetics across the cell cycle. *Elife* **5**, e12175 (2016).
20. O. G. Berg, A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **71**, 587–603 (1978).
21. D. Huh, J. Paulsson, Random partitioning of molecules at cell division. *Proc. Acad. Natl. Sci. U.S.A.* **108**, 15004–15009 (2011).
22. J. R. Peterson, J. A. Cole, J. Fei, T. Ha, Z. A. Luthey-Schulten, Effects of DNA replication on mRNA noise. *Proc. Acad. Natl. Sci. U.S.A.* **112**, 15886–15891 (2015).
23. O. Padovan-Merhar *et al.*, Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
24. J. Lin, A. Amir, Homeostasis of protein and mRNA concentrations in growing cells. *Nat. Commun.* **9**, 4496 (2018).
25. D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
26. H. Ochiai, T. Sugawara, T. Sakuma, T. Yamamoto, Stochastic promoter activation affects NANOG expression variability in mouse embryonic stem cells. *Sci. Rep.* **4**, 7125 (2014).
27. K. B. Halpern *et al.*, Bursty gene expression in the intact mammalian liver. *Mol. Cell* **58**, 147–156 (2015).
28. H. Xu, S. O. Skinner, A. M. Sokac, I. Golding, Stochastic kinetics of nascent RNA. *Phys. Rev. Lett.* **117**, 128101 (2016).
29. Y. Wang *et al.*, Precision and functional specificity in mRNA decay. *Proc. Acad. Natl. Sci. U.S.A.* **99**, 5860–5865 (2002).
30. C. Cadart *et al.*, Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. *Nat. Commun.* **9**, 3275 (2018).
31. N. Reverón-Gómez *et al.*, Accurate recycling of parental histones reproduces the histone modification landscape during DNA replication. *Mol. Cell* **72**, 239–249 (2018).
32. S. Hamperl, K. A. Cimprich, Conflict resolution in the genome: How transcription and replication make it work. *Cell* **167**, 1455–1467 (2016).
33. C. A. Vargas-Garcia, K. R. Ghusinga, A. Singh, Cell size control and gene expression homeostasis in single-cells. *Curr. Opin. Struct. Biol.* **8**, 109–116 (2018).

34. O. Sandler *et al.*, Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature* **519**, 468–471 (2015).
35. A. Golubev, Applications and implications of the exponentially modified gamma distribution as a model for time variabilities related to cell proliferation and gene expression. *J. Theor. Biol.* **393**, 203–217 (2016).
36. G. M. Walker, Synchronization of yeast cell populations. *Methods Cell Sci.* **21**, 87–93 (1999).
37. Y. Tian, C. Luo, Y. Lu, C. Tang, Q. Ouyang, Cell cycle synchronization by nutrient modulation. *Integr. Biol.* **4**, 328–334 (2012).
38. C. Gérard, A. Goldbeter, Entrainment of the mammalian cell cycle by the circadian clock: Modeling two coupled cellular rhythms. *PLoS Comput. Biol.* **8**, e1002516 (2012).
39. E. Kowalska *et al.*, NONO couples the circadian clock to the cell cycle. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1592–1599 (2013).
40. P. H. O'Farrell, J. Stumpff, T. T. Su, Embryonic cleavage cycles: How is a mouse like a fly?*Curr. Biol.* **14**, R35–R45 (2004).
41. P. Thomas, Making sense of snapshot data: Ergodic principle for clonal cell populations. *J. R. Soc. Interface* **14**, 20170467 (2017).
42. C. A. Yates, M. J. Ford, R. L. Mort, A multi-stage representation of cell proliferation as a Markov process. *Bull. Math. Biol.* **79**, 2905–2928 (2017).
43. B. Ingalls, Sensitivity analysis: From model parameters to system behaviour. *Essays Biochem.* **45**, 177–194 (2008).
44. A. A. Cohen *et al.*, Protein dynamics in individual human cells: Experiment and theory. *PLoS One* **4**, e4901 (2009).
45. A. J. Hughes *et al.*, Single-cell Western blotting. *Nat. Methods* **11**, 749–755 (2014).
46. C. Albayrak *et al.*, Digital quantification of proteins and mRNA in single mammalian cells. *Mol. Cell* **61**, 914–924 (2016).
47. B. Schwanhäusser *et al.*, Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
48. R. Grima, D. R. Schmidt, T. J. Newman, Steady-state fluctuations of a genetic feedback loop: An exact solution. *J. Chem. Phys.* **137**, 035104 (2012).
49. Z. Cao, R. Grima, Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **9**, 3305 (2018).
50. D. Cappelletti, C. Wiuf, Elimination of intermediate species in multiscale stochastic reaction networks. *Ann. Appl. Probab.* **26**, 2915–2958 (2016).
51. V. Stathopoulos, M. A. Girolami, Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philos. Trans. Roy. Soc. A* **371**, 20110541 (2013).
52. Z. Cao, R. Grima, Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *J. R. Soc. Interface* **16**, 20180967 (2019).
53. C. R. Bartman *et al.*, Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Mol. Cell* **73**, 519–532 (2019).
54. M. Soltani, C. A. Vargas-Garcia, D. Antunes, A. Singh, Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes. *PLoS Comput. Biol.* **12**, e1004972 (2016).
55. A. Y. Weiße, D. A. Oyarzún, V. Danos, P. S. Swain, Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1038–E1047 (2015).

I&EC research
Industrial & Engineering Chemistry Research

Review

Cite This: Ind. Eng. Chem. Res. 2019, 58, 17164−17173

pubs.acs.org/IECR

110th Anniversary: An Overview on Learning-Based Model Predictive Control for Batch Processes

Jingyi Lu,[†] Zhixing Cao,[*,‡,∥] Chunhui Zhao,[§] and Furong Gao[*,†]

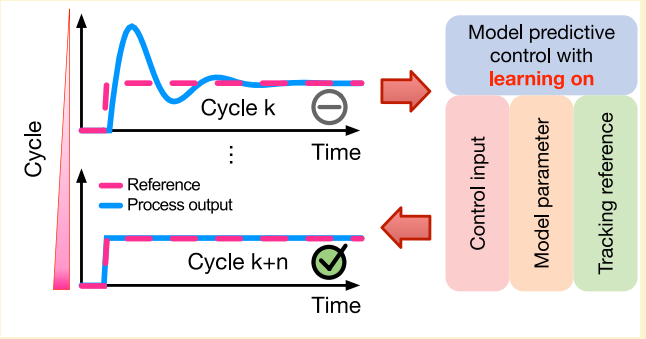
[†]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong

[‡]Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

[§]School of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

[∥]Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China

ABSTRACT: Batch processes repeatedly execute a given set of tasks over a finite duration, whose versatility and ability to adapt to rapidly changing markets make it prevalent in a multitude of industrial fields, particularly in the era of "smart manufacturing". Nevertheless, frequent switching and wide-ranging facility operations incur significant nonlinearity and time variability in process dynamics, both of which constitute remarkable challenges toward the regulation of batch processes. Among the various regulatory schemes, the integration of model predictive control and iterative learning schemes stands out, because of its inheritance of the merits of both: (i) ease of handling physical constraints and (ii) utilizing the repetitive operation pattern to adjust control input, process variables, and reference to improve control performance, consequently enhancing product quality. This review intends to account the recent technical advancements during the past two decades, from the perspective of the three different levels of learning mechanisms: control input, model parameter, and tracking reference. We conclude by providing insights into future research.



1. INTRODUCTION

Operations in industrial processes generally consist of two types: continuous and discrete. Continuous processes are typically employed in a wide range of high-volume industries, such as the petroleum industry, whereas discrete operations are essential features of batch processes that repetitively perform a given set of sequential tasks within a finite duration for the purpose of mass production.[1] A standard example of a batch process is the injection molding process, whose workflow is shown in Figure 1. Products of batch processes range vastly from specialty chemicals, such as detergents, flocculants, adhesives, antifreeze to semiconductors, and polymer products, constituting a trillion-dollar-scale business.[1−3] The prevalence of batch processes principally stems from its economic and technical merits:

(i) Operational discontinuities endow batch processes with remarkable flexibility in revising manufacturing recipes to meet the rapidly changing demands of market, potentially avoiding overproduction and reducing storage cost.

(ii) The discontinuity also renders batch processes more robust against malfunction or facility failure, because of simpler maintenance.

Notably, massive numbers of biological and biomedical processes present batch-process-like behavior. For instance, blood glucose variation arising from a regular daily meal intake was tracked for the purpose of attaining an optimal setting for an artificial pancreas controller to minimize the hyperglycemic and hypoglycemic risk of diabetic patients.[4,5] Another example is cellular gene expression in actively proliferating cells. Despite a multitude of noisy factors observed in experiments of gene expression,[6−9] there are clear signs of repeatable patterns in mRNA and protein molecules induced by the regular occurrence of gene replication and cell division.[10,11] In summary, batch processes are of great interest, not only in industrial manufacturing but also academically for studying biological and biomedical processes, highlighting the necessity for research efforts that provide quantitative insights into dynamic analysis or controller synthesis for underlying processes.

On the other hand, batch processes are notorious for their ubiquitous nonlinearity and time-varying dynamics, being underpinned by a combination of frequent switching and

Received: April 30, 2019
Revised: August 16, 2019
Accepted: August 25, 2019
Published: August 26, 2019

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

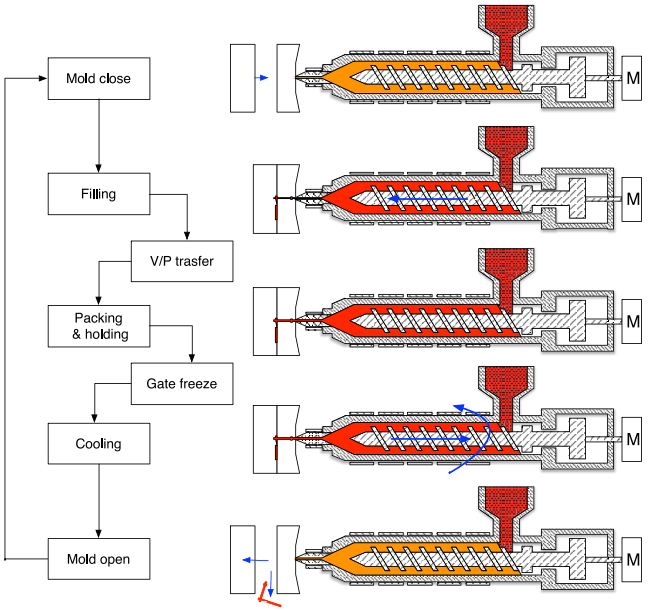


Figure 1. Workflow of injection molding. An entire working cycle consists of three phases: filling, packing, and holding & cooling. A cycle is initiated with closing the mold. Following that, the filling phase starts with the screw moving forward, pushing polymer melt into the mold cavity through runner and gate. In the meantime, the pressure in mold cavity increases gradually until triggering a procedure called V/P transfer and marking the initiation of the packing and holding phase, during which a small amount of additional polymer melt is injected to offset the shrinkage induced by polymer cooling and solidification. Once the gate freezes, the packing and cooling phase terminates and the cooling phase starts, which continues until the polymer has fully solidified and products are ejected. Accompanying that, the screw rotates in high speed to melt polymer granules preparing to start the next cycle. All of the procedures then are repeated again. Similar operational patterns have also been noted in other batch processes, such as batch reactors.

wide ranging facility operations, consequently posing an outstanding challenge to batch process control. Early attempts to resolve the problem include the direct application of continuous batch process control techniques or under minor tailoring,[12−17] usually either failing to yield impressive control performance or contending with overwhelming computational cost, motivating the development of novel techniques for batch processes. A seminal breakthrough was made by Arimoto[18] in 1984, with the proposal of *iterative learning control (ILC)*. Although it was initially proposed for controlling robotic arms, the underlying idea of exploiting the innate repeatable operation pattern of the process to adjust current control input according to trial error readily known from previous cycles was truly innovative. In fact, it imitates the human learning process, resulting in control performance improvement from cycle to cycle, an idea that is now appreciated by numerous fields, particularly stimulating a huge wave of interest in the community of batch process control. ILC as a feedforward control technique was found to suffer from poor robustness against irreproducible noise or disturbance, highlighting a high risk for occurrence of rapid performance deterioration. To resolve such a problem, there appeared a burst of ILC variants integrated with various classic feedback control strategies including state feedback,[19,20] internal model control,[21] optimal control,[22,23] and robust control,[24−26] so that any undesired nonrepetitive factors are attenuated or even almost perfectly eliminated by feedback mechanism. Excellent reviews on more general ILC are given in refs 27−30. Amid the arsenal of ILC variants, the integration of ILC with *model predictive control (MPC)*[31−33] is identified as being good at (i) handling physical constraints on control inputs and process outputs straightforwardly and (ii) utilizing repeatable operation patterns to attenuate any undesired repeatable factors to a minimum by iteratively updating ("learning") a controller in a single or multiple level(s), thus becoming a competitive branch in engineering practice. To the best of our knowledge, a proper survey of this topic is still absent from the literature. This review attempts to fill this gap in the literature and differs from the aforementioned general ILC reviews[27−30] in several respects:

(i) we collectively review the progress of MPC for batch processes under the umbrella of learning-based MPC, wherein we emphasize that the learning mechanism can occur on one of the levels: control input, process parameter, tracking reference, or several of them, rather than purely on control input (which is conventionally referred as *iterative learning model predictive control (ILMPC)*);

(ii) we first present the fundamental principles of each category and then summarize the merits and drawbacks of each one to serve the purpose of a brief tutorial; and

(iii) more technical details pertaining to stability analysis are also provided to readers with more advanced knowledge.

The rest of the overview is organized as follows. Section 2 quantitatively formulates the batch process control problem in the framework of MPC. Sections 3−5 review the state-of-the-art of learning-based MPC in a categorizing manner. Discussions and outlook for open topics are presented in Section 6.

2. PRELIMINARIES FOR BATCH PROCESS CONTROL

Without a loss of generality, the dynamics of a batch process can be described by a nonlinear state space model in discrete time:

$$\begin{aligned} x_{k,t+1} &= f(x_{k,t}, u_{k,t}, \theta) \\ y_{k,t} &= g(x_{k,t}) \end{aligned} \tag{1}$$

where $t \in \mathbb{I}_{[1,N]}$ is an integer between 1 and N, denoting the time instant with N being the duration of a cycle, and the cycle number is represented by $k \in \mathbb{I}_{[1,\infty)}$. The variables $x_{k,t}$, $u_{k,t}$, and $y_{k,t}$ represent the system state, input, and output at time t and cycle k, respectively, while θ parametrizes the dynamics of a batch process, which can be a set of either constants or real functions in time. The exact knowledge of functions f, g, and parameter θ is rarely known due to either the presence of complicated underlying mechanisms, technical limitations, or a mix of both. Given the difficulty to obtain exact knowledge, an approximation is always possible, which is usually referred as a process model. The model is a must in learning-based MPC, as suggested by the name, and can be generally described by

$$\begin{aligned} x_{k,t+1|t} &= \bar{f}(x_{k,t}, u_{k,t}, \bar{\theta}) \\ y_{k,t+1|t} &= \bar{g}(x_{k,t+1|t}) \end{aligned} \tag{2}$$

Here, $x_{k,t+1|t}$ and $y_{k,t+1|t}$ denote the one-step prediction of process state and output in time t. The overbars indicate the

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

approximation (modeling) to f, g, and θ, and emphasize the existence of topological or parametric disparities thereof, which play a key role in determining performance of a batch process controller.

Now the batch process control problem is readily presented as an optimization problem (akin to the classic framework of MPC):

Problem 1

$$\min_{\mathbf{u}_{k,t:t+p_n-1|t}} J(\mathbf{r}_{t+1:t+p_n}, \mathbf{y}_{k,t+1:t+p_n|t}, \mathbf{u}_{k,t:t+p_n-1|t})$$

subject to

$$x_{k,t+i|t} = \bar{f}(x_{k,t+i-1|t}, u_{k,t+i-1|t}, \bar{\theta})$$
$$y_{k,t+i|t} = \bar{g}(x_{k,t+i|t})$$
$$x_{k,t|t} = x_{k,t}$$
$$u_{k,t+i-1|t} \in \mathcal{U}$$
$$x_{k,t+i|t} \in \mathcal{X}$$
$$y_{k,t+i|t} \in \mathcal{Y}$$

for any integer i between 1 and prediction horizon p_n in any time t and cycle k. The vectors $\mathbf{r}_{t+1:t+p_n}$, $\mathbf{y}_{k,t+1:t+p_n|t}$, and $\mathbf{u}_{k,t:t+p_n-1|t}$ collect the given references, the predictions on process output from time $t+1$ to $t+p_n$ and control input from time t to $t+p_n-1$, respectively, i.e., $\mathbf{y}_{k,t+1:t+p_n|t} = [y_{k,t+1|t}^{\mathrm{T}}, \ldots, y_{k,t+p_n|t}^{\mathrm{T}}]^{\mathrm{T}}$. The feasible domains on control input, process state, and process output are denoted by script letters $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$, respectively. The fundamental idea underpinning Problem 1 is to recursively minimize a given functional J over a finite prediction horizon while satisfying some physical constraints to ensure the normal functioning of equipment and process safety.

Indeed, the functional J can be tailored to serve different customized demands in reality. For instance, it can be (i) some metrics quantifying the discrepancy between a given reference $\mathbf{r}$ and process output $\mathbf{y}$ over a prediction horizon or a subset of time points thereof when perfect tracking may not be attainable,[34,35] (ii) some indicators of economic interest to be maximized,[2,36,37] (iii) a Dirac function on the terminal process output $y_{k,N|t}$ focusing on final product quality, (iv) reaction time requiring minimization,[38] or (v) distributional metrics in order to find the most desirable shape of crystal size distribution.[15]

Learning can be intuitively interpreted as improvement from past trials, enabling us to define it mathematically. For example, learning is allowed to occur on (i) control input $\mathbf{u}_k$ that becomes a function of past control input, i.e., $\mathcal{F}(\mathbf{u}_{k-1})$; (ii) process parameters $\bar{\theta}_k$, which is formulated as a function of past estimates, i.e., $\mathcal{F}(\bar{\theta}_{k-1})$; and (iii) reference $\mathbf{r}_k$, which combines the knowledge from past references, i.e., $\mathcal{F}(\mathbf{r}_{k-1})$. The occurrence on the three different levels constitutes the categorizing system presented in this overview, as well as the following three sections.

3. LEARNING ON CONTROL INPUT

The first category of learning-based MPC is exemplified by the classic ILMPC, where the control input $u_{k,t}$ is a function of previous control input, i.e., $\mathcal{F}(u_{k-1,t})$, reflecting control input updating (learning) based on information fed from past trials. One of the plausible formulations of $\mathcal{F}$ is

$$u_{k,t} = \mathcal{F}(u_{k-1,t}) = u_{k-1,t} + u_{\mathrm{ff}} + u_{\mathrm{fb}}$$

representing iterative revisions on control input in the previous cycle by incorporating trial error information in the past (u_{ff}), as well as real-time feedback information u_{fb}. The former intends to successively adjust control input to compensate the inevitable discrepancy between a real process (f and g) and its process model ($\bar{f}$ and $\bar{g}$), whereas the latter is useful to enhance robustness against noise and unrepeatable disturbances. Hence, to successfully implement the ILMPC controller, one only needs to determine u_{ff} and u_{fb} optimally. One of the classic determinations is presented in the following subsection.

3.1. Basic Idea. The model of a batch process most commonly used is the linear time invariant state space model:

$$x_{k,t+1|t} = \bar{A}x_{k,t} + \bar{B}u_{k,t|t}$$
$$y_{k,t+1|t} = Cx_{k,t+1|t} \tag{3}$$

whose convenient development arises from the availability of a wide range of data-driven modeling methods.[39] As mentioned in the previous section, the presence of strong nonlinearity of batch process indicates significant plant-model mismatch between eqs 1 and 3, which is typically defined as

$$\delta_{k,t} = g \circ f(x_{k,t}, u_{k,t}, \theta_t) - C\bar{A}x_{k,t} - C\bar{B}u_{k,t}$$

such that the process output at time t and cycle k can be rewritten as

$$y_{k,t} = y_{k,t+1|t} + \delta_{k,t}$$

If functions f and g are continuously differential and Lipschitz, which is generally the case in practice, then there exist some constants $\mathcal{L}^x$ and $\mathcal{L}^y$ such that the difference of mismatches of two successive cycles is upper bounded by

$$\|\delta_{k,t} - \delta_{k-1,t}\|_2 \le \mathcal{L}^x\|x_{k,t} - x_{k-1,t}\|_2 + \mathcal{L}^u\|u_{k,t} - u_{k-1,t}\|_2$$

In other words, if $x_{k,t}$ and $u_{k,t}$ are close to $x_{k-1,t}$ and $u_{k-1,t}$, then the plant−model mismatch in the previous cycle ($\delta_{k-1,t}$) is a fairly good approximation to that in the current cycle ($\delta_{k,t}$), also suggesting a more refined prediction of process output:

$$\begin{aligned}\hat{y}_{k,t+1|t} &= y_{k,t+1|t} + \delta_{k-1,t} \\ &= y_{k,t+1|t} - y_{k-1,t+1|t} + y_{k-1,t+1} \\ &= C\bar{A}\Delta x_{k,t} + C\bar{B}\Delta u_{k,t} + y_{k-1,t+1} \\ &\approx y_{k,t+1}\end{aligned} \tag{4}$$

Here, Δ is the cycle-wise difference operator (i.e., $\Delta x_{k,t} = x_{k,t} - x_{k-1,t}$). The third equality can be easily deduced from eq 3. Putting eq 4 in the framework of Problem 1, the optimization problem of lLMPC then turns to be

Problem 2

$$\min_{\mathbf{u}_{k,t:t+p_n-1|t}} \left\|\mathbf{r}_{t+1:t+p_n} - \hat{\mathbf{y}}_{k,t+1:t+p_n|t}\right\|_Q^2 + \left\|\mathbf{u}_{k,t:t+p_n-1} - \mathbf{u}_{k-1,t:t+p_n-1|t}\right\|_R^2 \tag{5a}$$

subject to

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

$$\hat{y}_{k,t+i|t} = C\bar{A}^i \Delta x_{k,t} + y_{k-1,t+i} + \sum_{j=0}^{i-1} C\bar{A}^j \bar{B}(u_{k,t+j|t} - u_{k-1,t+j})$$

other constraints (5b)

where Q and R are positive definite weight matrices, and $\|\mathbf{x}\|_Q = \sqrt{\mathbf{x}^T Q \mathbf{x}}$. The second term in eq 5a is used to ensure that the control input of cycle k stays so close to that of cycle $k-1$ that the plant-model mismatch $\delta_{k-1,t}$ is a valid approximation to $\delta_{k,t}$. The constraint eq 5b is derived by iterating the third equality in eq 4 i times.

In the absence of any other constraint imposed, Problem 2 further reduces to a convex quadratic programming, and an explicit solution is possible to solve in the form of

$$u_{k,t|t} = u_{k-1,t} + \underbrace{\mathcal{K}\mathbf{e}_{k-1,t+1:t+p_n}}_{u_{ff}} - \underbrace{\mathcal{K}\bar{G}_1 \Delta x_{k,t}}_{u_{fb}} \quad (6)$$

with the gain $\mathcal{K}$ and the matrix $\bar{G}_1$ compactly defined as

$$\mathcal{K} = \mathcal{I}_{p_n}(\bar{G}_2^T Q \bar{G}_2 + R)^{-1} \bar{G}_2^T Q$$

$$\mathcal{I}_{p_n} = [1, \underbrace{0, ..., 0}_{p_n - 1}]$$

$$\bar{G}_1 = [(C\bar{A})^T, ..., (C\bar{A}^{p_n})^T]^T$$

$$\bar{G}_2 = \begin{bmatrix} C\bar{B} & 0 & ... & 0 \\ C\bar{A}\bar{B} & C\bar{B} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C\bar{A}^{p_n-1}\bar{B} & C\bar{A}^{p_n-2}\bar{B} & ... & C\bar{B} \end{bmatrix}$$

and $\mathbf{e}_{k,t:t+t_1} = \mathbf{r}_{t:t+t_1} - \mathbf{y}_{k,t:t+t_1}$. There are two components included in the control law described by eq 6: (i) the feedforward part u_{ff} is a linear combination of tracking error in the previous cycle, and (ii) the feedback part u_{fb} is proportional to the difference between process states, whose gains are determined optimally. By contrast, an explicit solution is generally impossible to have if some other physical constraints are imposed, but a numerical one may still be obtainable with affordable computation.[40] Although discussion thus far has focused on tracking problems, ILMPC can also be tailored to cater for other problems, such as point-to-point tracking problems[34] and economic optimization problems.[41]

3.2. Literature Review. The idea on modulating control input to refine predictive model by using historical information cyclewise was first reported in ref 42, wherein process disturbance was assumed to be a function only of cycle and independent of system states and inputs, and was estimated by averaging prediction error of previous cycles. Later, the idea[43] was extended to handle plant−model mismatch, which is also a function of control input as well and can be effectively attenuated at the cost of solving a high-dimensional optimization problem (Problem 2) over an entire cycle duration. Given the high computational cost thereof, Shi and co-workers[44] adapted ILMPC for online implementation by using a dynamic model and reducing prediction horizon to be smaller than cycle duration. Besides, there have been a large number of ILMPC variants proposed to improve robustness against (i) stochastic noises using Kalman filter[45] or gain adaptation,[46] (ii) nonrepetitive deterministic disturbances using two-stage optimization to optimally estimate model states and seek proper control actions,[47,48] and (iii) nonlinearity by feeding back a linear combination of prediction errors of multiple time points in the previous cycle with optimized weights.[49]

3.3. Stability Analysis. Stability issues play a crucial role in the design of ILMPC controllers; failing to address it properly may lead to significant degradation of controller performance and even jeopardize equipment safety. In this section, we summarize state-of-the-art ILMPC stability analysis and highlight the importance of this concept—two-dimensional recursive feasibility, which was traditionally overlooked when the optimization problem of ILMPC is constrained.

The most classic stability argument is to formulate a dynamic system in the state space model compactly into a lifted model,

$$\mathbf{y}_{k,2:N} = G\mathbf{u}_{k,1:N-1}$$

by stacking outputs and inputs of an entire cycle into supervectors $\mathbf{y}_{k,2:N}$ and $\mathbf{u}_{k,1:N-1}$, respectively. The matrix G herein is composed of matrices $\bar{A}$, $\bar{B}$, and C in eq 3. By defining a matrix M in the same sense as G, one can show the error dynamics in the form of

$$\mathbf{e}_{k,2:N} = M\mathbf{e}_{k-1,2:N} \quad (7)$$

Using contraction analysis, it is not difficult to conclude that eq 7 converges if and only if the absolute value of every eigenvalue of M is strictly <1. A quantitative condition on how to choose the weight matrix R in eq 5a (the most important parameter determining the gain $\mathcal{K}$ in control law eq 6) to achieve stability is nontrivial, except for a special case where the prediction horizon is 1: plant-model mismatch is absent and consequently matrix M becomes lower triangular with diagonal elements being always strictly <1 for any positive definite matrix R. Otherwise, the only intuitive condition that is helpful to achieve stability is one that heavily penalizes control input (large R) but at the expense of a slow convergence rate.[44,50]

Of note, this contraction-based argument starts to break when other physical constraints are introduced in eq 5b, since the availability of an explicit solution to eq 6 is compromised. To this end, Liu and Wang[51] proposed to enforce a constraint based on a two-dimensional Lyapunov function; unfortunately, one does not know if the optimization problem is guaranteed to be feasible. In other words, the approach can neither ensure stability, because there is a subtle problem—consistent feasibility, which is a concept of fundamental importance in the field of MPC for continuous processes[52−54] and should be recursively satisfied, i.e., feasibility at time t indicating feasibility at time $t+1$. How to extend this concept to ILMPC for batch processes has remained elusive hitherto, since one must address consistent feasibility on both time and cycle directions, which is substantially more challenging.

This outstanding problem of two-dimensional feasibility for ILMPC was first systematically discussed in ref 55, where an equality constraint on terminal predictions was implemented:

$$x_{k,t+p_n|t} = x_{k-1,t+p_n} \quad (8)$$

such that control inputs of previous cycle can constitute a set of feasible solutions for the ongoing cycle. Strikingly, this set of feasible solutions automatically grants the objective function to be nonincreasing cyclewise, and the strictly monotonic convergence on tracking error can be further shown provided that the optimization problem is convex. This idea was also

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

used to show the stability of an economic IMPC controller.[41] Note that the constraint described by eq 8 may be too stringent to admit any solution other than $u_{k,t+i} = u_{k-1,t+i}$, potentially leading to no improvement of control performance from cycle to cycle. Therefore, a less restrictive version to eq 8 was developed in ref 56, which is

$$x_{k,t+p_n|t} \in \mathcal{S}^k \tag{9}$$

with $\mathcal{S}^k$ being a cycle-dependent but time-invariant set containing $x_{k-1,t+p_n}$. The constraint described by eq 9 can be fulfilled by (i) driving states into an ellipsoidal $\mathcal{S}^k$ by an auxiliary controller,

$$\Delta u_{k,t+i} = K_1 \Delta x_{k,t+i|t} + K_2 e_{k-1,t+i+1}, \qquad i \in \mathbb{I}_{[p_n-1,N-1]}$$

where two terms ensure that the underlying system converge, both timewise and cyclewise, and whose gains K_1 and K_2 are solved from a coupled set of linear matrix inequalities,[57] or (ii) using geometric computation for a polytopic set $\mathcal{S}^k$ without need of an auxiliary controller.[58] Ensured monotonic convergence on the tracking error was reported in refs55, 57, and 58, whereas that on the upper bound of tracking error was achieved by a means developed in ref 49.

4. LEARNING ON MODEL PARAMETERS

By contrast, an alternative learning mechanism is to constantly modulate the process parameters to drive output of a parametrized model toward that of the real process of interest locally (in the proximity of a given tracking reference) rather than globally, the latter of which is generally not achievable, because of the presence of strong nonlinearity and time variation. Control input adapts in accordance with model updating. To be specific, the parametric estimate $\bar{\theta}_{k,t}$ in eq 2 (at time t and cycle k) turns to be a function of that in the previous cycle, i.e., $\mathcal{F}(\bar{\theta}_{k-1,t})$. Consequently, the problem of devising the learning mechanism is specified to be how to seek an appropriate form of $\mathcal{F}(\bar{\theta}_{k-1,t})$.

4.1. Basic Idea. In this section, we present the dynamics of real process in the form of

$$y_{k,t} = G_t(z^{-1})u_{k,t}$$

with G being the transfer function and its subscript t emphasizing the time variation of process dynamics. One can always find a model for G_t via classic system identification methods,[39] based on historical information within an ongoing cycle ($\hat{G}_{t|t-1}$). Specifically, the corresponding prediction of process output becomes

$$y_{k,t|t-1} = \hat{G}_{t|t-1}u_{k,t}$$

and prediction error ϵ_T (whose subscript T stands for prediction from the time direction) is immediately found as

$$\begin{aligned}\epsilon_T &= y_{k,t} - y_{k,t|t-1}\\ &= [G_t - \hat{G}_{t|t-1}]u_{k,t}\\ &= \underbrace{[G_{t-1} - \hat{G}_{t|t-1}]u_{k,t}}_{\text{Estimation error}} + \underbrace{[G_t - G_{t-1}]u_{k,t}}_{\text{Systematic error}}\end{aligned}$$

The last step shows the two components of prediction error—estimation error and systematic error. One can then proceed with the error decomposition, since $\hat{G}_{t|t-1}$ only uses information up to time $t-1$, thus being a good estimate of G_{t-1} rather than G_t. In view of law of large numbers, the estimation error therein can be eliminated as long as $\hat{G}_{t|t-1}$ is an unbiased estimator, whereas the systematic error persists in this scheme. In contrast, it is also possible to predict process output based on information across cycles (typically of previous cycles but at the same time), as

$$y_{k|k-1,t} = \hat{G}_{k|k-1,t}u_{k,t}$$

where $\hat{G}_{k|k-1,t}$ acts as an estimator to G_t, and we speculate that the estimator is unbiased, since it depicts the dynamics at time t. As a result, the prediction error (of using cyclewise predictions) is found to be

$$\begin{aligned}\epsilon_C &= y_{k,t} - y_{k|k-1,t}\\ &= [G_t - \hat{G}_{k|k-1,t}]u_{k,t}\end{aligned}$$

which indicates the possibility of elimination of the systematic error. Admittedly, the arguments are simple and intuitive but provide answers to why one should update the model (or equivalently $\bar{\theta}_{k,t}$) cyclewise, providing key insights into how to devise the learning mechanism.

To be specific, the process G_t is exemplified as an *autogressive with exogenous input (ARX)* model that can be compactly presented in a linear-regression form,

$$y_{k,t} = \phi_{k,t}^{\mathrm{T}}\theta_t$$

with $\phi_{k,t}$ denoting a collection of inputs and outputs

$$\phi_{k,t} = [-y_{k,t-1}, ..., -y_{k,t-n_a}, u_{k,t-1}, ..., u_{k,t-n_b}]$$

and n_a and n_b being the model orders on outputs and inputs, respectively. The vector θ_t is a set of process parameters, whose estimate $\bar{\theta}_{k,t}$ can be inferred by solving the following optimization problem,

$$\bar{\theta}_{k,t} = \arg\min_{\hat{\theta}} \sum_{i=1}^{k} \left\| y_{i,t} - \phi_{i,t}^{\mathrm{T}}\hat{\theta} \right\|_2^2 \tag{10}$$

and subsequently used to update control input. The problem in eq 10 allows efficient computation through recursive least-squares (one possible form of $\mathcal{F}$).[59] Notably, the objective function in eq 10 is defined across multiple cycles rather than across multiple times in the classic sense so that the systematic error can be possibly removed, as previously argued.

Having parameter estimates of the ongoing cycle been found, we can compute the control input at time t and cycle k within the framework akin to that of Problem 1:

Problem 3

$$\min_{\mathbf{u}_{k,t:t+p_n-1|t}} \left\| \mathbf{r}_{t+1:t+p_n} - \mathbf{y}_{k,t+1:t+p_n|t} \right\|_Q^2 + \left\| \mathbf{u}_{k,t:t+p_n-1|t} \right\|_R^2$$

subject to

$$y_{k+1,t+i|t} = \phi_{k+1,i|t}^{\mathrm{T}}\bar{\theta}_{k,t+i}$$

$$\phi_{k,t+i|t} = [-y_{k,t+i-1|t}, ..., -y_{k,t-n_a}, u_{k,t+i-1|t}, ..., u_{k,t-n_b}]^{\mathrm{T}}$$

other constraints

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

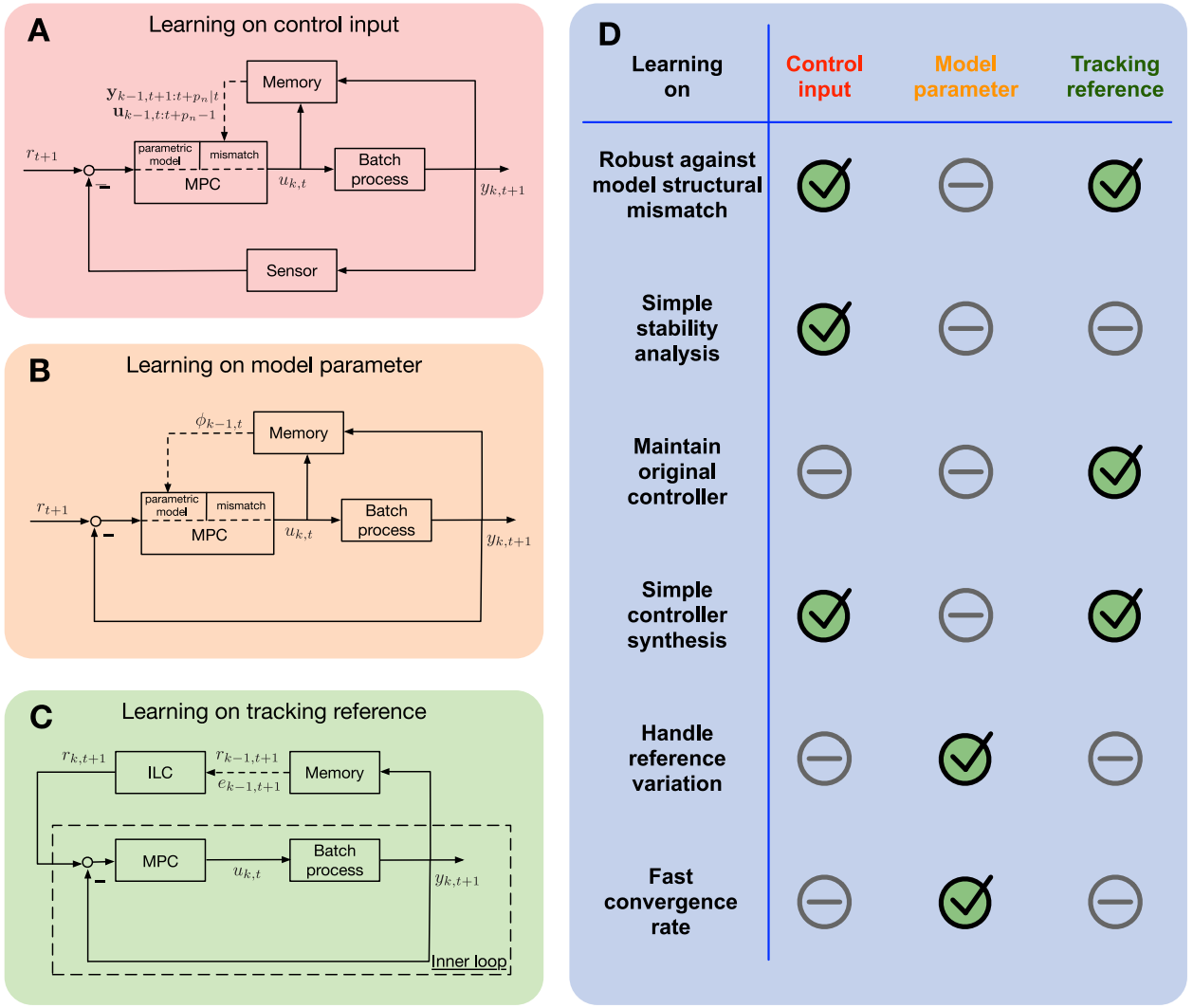


Figure 2. Three types of learning-based model predictive control: MPC with learning on control input, model parameter and tracking reference. (A–C) Schematics of the three types ((A) learning on control input, (B) learning on model parameter, and (C) learning on tracking reference). (D) Summary for merits and drawbacks of the three learning mechanisms.

where the process output, $y_{k+1,t+i\,|\,p}$ is predicted based on the cyclewise model parameter estimate $\bar{\boldsymbol{\theta}}_{k,t+i}$, instead of the timewise estimate, $\bar{\boldsymbol{\theta}}_{k+1,t+i-1}$ in the traditional sense.

4.2. Literature Review. The earliest study on MPC based on parameter learning can be traced back to ref 60, in which model parameters are updated with timewise historical data. Such a learning mechanism is only suitable for handling dynamics with slow time variation, thereby limiting wide applicability. Rather, estimates obtained from a cyclewise optimization problem (eq 10) are unbiased; however, they may not guarantee improvement on control performance as expected. Indeed, the removal of the systematic error is at the cost of losing the smoothing effect on estimates, consequently leading to undesirable fluctuations on parameter estimates, and ultimately propagating to process output, as first noted in ref 59 and later supported by numerical evidence in ref 61 (see Figure 5 therein). To circumvent the issue, one must refine parameter estimates, either (i) by using a priori process knowledge to formulate sensible constraints on position of poles and zeros,[62] closed-loop performance,[63] or range of time constants[61] into the optimization problem (eq 10), or (ii) by using averaging-based identification methods.[61]

Note that both MPCs based on control input learning and model parameter learning are essentially data-driven methods for refining model predictions, but from different angles: the former aims to capture plant–model mismatch (or disturbance) more effectively, while the latter intends to find a better nominal description to real-time dynamics. One may envision that better batch process control performance may be gained by orchestrating both learning mechanisms, which is the case reported in ref 65. The authors therein used a time- and cycle-dependent linear model to approximate a nonlinear batch process by means of the mean value theorem; interestingly, they found a control law in the form of eq 6 but with the gain $\mathcal{K}$ being a function of time and cycle rather than a constant.

4.3. Stability Analysis. Within this learning mechanism, the convergence of identification algorithm is a key player in stabilizing the process of interest and deciding closed-loop control performance. Given its importance, it was systematically studied in refs 59, 61, 62, and 64, wherein identifications rely on the condition of persistent excitation. Nevertheless, the general conclusions of closed-loop stability of parameter-learning based MPC still remains elusive but are available only in few special cases: conditions were found for unconstrained

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164–17173

MPC, depending on a special assumption that the sign of each element of the parameter vector $\boldsymbol{\theta}_{k,t}$ is known and postulated to be invariant. The recent finding on robust MPC with a recursive model update,[66] which uses a combination of online set-membership system identification and homothetic prediction tubes, may provide clues to developing sufficient conditions to ensure closed-loop stability of parameter-learning based MPC, particularly when the optimization problem is constrained.

5. LEARNING ON TRACKING REFERENCE

Finally, we will discuss the learning on tracking reference, which is also known as indirect iterative learning control.

5.1. Basic Idea. The learning mechanism on tracking reference automatically admits the inclusion of an upper-level controller modulating the reference fed to a local controller, hence constituting a cascade control structure, as shown in Figure 2C. Specifically, the reference at cycle k ($r_{k,t}$) can either be a constant or a function of time, whose update is determined by a functional on its precursor $\mathcal{F}(r_{k-1,t})$. A simple but plausible implementation of the functional $\mathcal{F}$ is

$$r_{k,t} = r_{k-1,t} + \Delta r$$

where the increment $\Delta \mathbf{r}$ can be decided based on a broad range of criteria, such as optimizing some indices of clinical or economic interest.[4] Among all, the simplest determination may be making $\Delta \mathbf{r}$ be proportional to the tracking error in the past:

$$r_{k,t} = r_{k-1,t} + Ke_{k-1,t}$$

which leads to the P-type indirect ILC.[67] Here, we consider the most general case by putting the subscript t back to emphasize the time variation. Then, the reference seamless fits into the framework of Problem 1, yielding an optimization problem for indirect ILMPC.

Problem 4

$$\min_{\mathbf{u}_{k,t:t+p_n-1|t}} \left\| \mathbf{r}_{k,t+1:t+p_n} - \mathbf{y}_{k,t+1:t+p_n|t} \right\|_Q^2 + \left\| \mathbf{u}_{k,t:t+p_n-1|t} \right\|_R^2$$

subject to

$$y_{k,t+i|t} = C\bar{A}^i x_{k,t} + \sum_{j=0}^{i-1} C\bar{A}^j \bar{B} u_{k,t+j|t}$$

other constraints

Here, $\mathbf{r}_{k,t+1:t+p_n}$ is a vector comprising the reference $r_{k,t}$ from time $t+1$ to $t+p_n$ of cycle k. The absence of other constraints allows one to find an explicit solution to Problem 4, which is

$$u_{k,t|t} = u_{k-1,t} + \mathcal{K}_1[\mathcal{K}_3 \mathbf{e}_{k-1,t+1:t+p_n} - \bar{G}_1 \Delta x_{k,t}] \quad (11)$$

with $\mathcal{K}_3 = \mathrm{diag}(K, ..., K)$. Surprisingly, despite reference-learning based MPC and ILMPC being motivated differently, the control law of the former (eq 11) reduces to that of the latter (eq 6) when $K = I$.

5.2. Literature Review. Indirect ILC was initially developed within a framework of adaptive control to simplify stability analysis,[68,69] and was later shaped to integrate with various types of feedback control, such as PID control and robust control for different application purposes.[70−73] The first integration with MPC was determined in ref 67, which presents technical details on how to design the learning increment $\Delta \mathbf{r}$ and determine the learning gain K in eq 11. The corresponding stability analysis using two-dimensional system theory was given in ref 74, and a successful clinical application was reported in ref 75. Instead of using indirect ILMPC for reference tracking, it was also applied in the problem of point-to-point tracking.

5.3. Stability Analysis. In view of the resemblance between the indirect ILMPC and ILMPC presented in Section 5.1, this incentivizes us to study the stability of the indirect ILMPC along a similar idea as ILMPC when constraints are absent. Typical examples include refs 67 and 74. By contrast, the stability analysis for the system using constrained MPC (the inner controller) is not clear, motivating more-careful inspections. The difficulty mainly arises from the cycle-to-cycle variation of reference signal fed to the inner MPC controller, while the analogous problem of studying the stability for classic MPC with time-varying tracking reference has not been fully addressed yet and is only reported for constant or piecewise constant references.[76,77]

6. CONCLUSION AND OUTLOOK

6.1. Summary. The study of batch processes has experienced a rapid growth in the past two decades, because of its excellent adaptability to the diverse manufacturing environment of today. Precise control on key variables in batch processes is an important albeit challenging goal, because of a multitude of factors: complex underlying mechanisms, highly nonlinear systems, time variation, restrictive constraints, etc. This review discussed three types of learning strategies integrated with model predictive control by (i) summarizing the state-of-the-art strategies, (ii) presenting fundamental principles of stability analysis, and (iii) discussing the advantages and disadvantages (see Figure 2D) of each type of analysis. Clearly, this is our own take on the state of the field, and, as such, we have been selective in the publications we have included in this review. We do not intend to devalue the many excellent results contributed by the batch process control community that are not included herein.

6.2. Future Research. In this section, we highlight three topics, which we think are either promising or are of primary importance for the field in the near future.

Stability analysis of each type of learning-based MPC will continue to be of primary importance. Although remarkable progress has been made with regard to stability analysis of the ILMPC, that of the other two types is far from well understood, thereby calling for more research efforts. These problems, albeit daunting, are potentially illuminated by recent development made for adaptive robust MPC[66,78,79] and dual MPC,[80,81] the latter of which uses a dual control scheme and optimizes identification and regulation objectives simultaneously.

One of the longstanding problems not only of learning-based MPC but also of ILC in general is how to improve robustness against nonrepetitive variations, including cycle-dependent references and disturbances, data dropouts in the networked context,[82] and variations on cycle duration.[83] These factors compromise the repetitive assumption underpinning ILC, making the synthesis of controllers capable of handling the variations a somewhat ill-defined problem. However, it may still be possible to achieve, given recent encouraging advancements: (i) the tube MPC method devising a tube enveloping nonrepetitive factors,[84] and (ii) the achievements

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

of machine learning techniques providing finer capture of complicated dynamics, such as high-order intercycle periodicity.[85]

The final emerging problem arises from the increasing intersystem heterogeneity of batch processes. For instance, it is commonly found in plastic factories that identical injection molding machines use different molds to manufacture various products, resulting in an ensemble of dynamics. It is even more pronounced in systems biology that isogenic cells exhibit different phenotypes principally stemming from the noisiness of multiple steps in gene expression.[6,7,9−11,86,87] Indeed, cellular gene expression can be subsumed into a batch process, given that cells are continuously proliferating; however, the cycle duration varies from cell to cell and from generation to generation. In view of the large number of similar processes, tuning a controller to each one of them manually constitutes an impossible mission. Promising solutions may include ensemble control[88,89] and model migration.[90−93]

AUTHOR INFORMATION

Corresponding Authors

*E-mail: zcao@ecust.edu.cn (Z. Cao).
*E-mail: kefgao@ust.hk (F. Gao).

ORCID

Zhixing Cao: 0000-0003-2600-5806
Chunhui Zhao: 0000-0002-0254-5763
Furong Gao: 0000-0002-5900-1353

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Z.C. gratefully acknowledges careful proofreading by James Holehouse. F.G. acknowledges support from the National Natural Science Foundation of China (Project No. 61433005) and Hong Kong Research Grant Council (Grant No. 16207717).

REFERENCES

(1) Reeve, A. Batch Control, the recipe for success. *Process Eng.* **1992**, *73*, 33−34.
(2) Bonvin, D. Control and optimization of batch processes. *Encycl. Syst. Control* **2015**, 133−138.
(3) Rippin, D. Simulation of single and multiproduct batch chemical plants for optimal design and operation. *Comput. Chem. Eng.* **1983**, *7*, 137−156.
(4) Cao, Z.; Gondhalekar, R.; Dassau, E.; Doyle, F. J. Extremum seeking control for personalized zone adaptation in model predictive control for type 1 diabetes. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 1859−1870.
(5) Cao, Z.; Dassau, E.; Gondhalekar, R.; Doyle, F. J., III Extremum seeking control based zone adaptation for zone model predictive control in type 1 diabetes. *IFAC-PapersOnLine* **2017**, *50*, 15074−15079.
(6) Cao, Z.; Grima, R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **2018**, *9*, 3305.
(7) Elowitz, M. B.; Levine, A. J.; Siggia, E. D.; Swain, P. S. Stochastic gene expression in a single cell. *Science* **2002**, *297*, 1183−1186.
(8) Ozbudak, E. M.; Thattai, M.; Kurtser, I.; Grossman, A. D.; Van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nat. Genet.* **2002**, *31*, 69.
(9) Cao, Z.; Grima, R. Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *J. R. Soc., Interface* **2019**, *16*, 20180967.
(10) Berg, O. G. A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **1978**, *71*, 587−603.
(11) Skinner, S. O.; Xu, H.; Nagarkar-Jaiswal, S.; Freire, P. R.; Zwaka, T. P.; Golding, I. Single-cell analysis of transcription kinetics across the cell cycle. *eLife* **2016**, *5*, No. e12175.
(12) Cluett, W.; Shah, S.; Fisher, D. Adaptive control of a batch reactor. *Chem. Eng. Commun.* **1985**, *38*, 67−78.
(13) Clarke-Pringle, T.; MacGregor, J. F. Nonlinear adaptive temperature control of multi-product, semi-batch polymerization reactors. *Comput. Chem. Eng.* **1997**, *21*, 1395−1409.
(14) Yang, Y.; Gao, F. Adaptive control of the filling velocity of thermoplastics injection molding. *Control Eng. Pract.* **2000**, *8*, 1285−1296.
(15) Nagy, Z. K.; Braatz, R. D. Robust nonlinear model predictive control of batch processes. *AIChE J.* **2003**, *49*, 1776−1786.
(16) Lucia, S.; Finkler, T.; Basak, D.; Engell, S. A new robust NMPC scheme and its application to a semi-batch reactor example. *IFAC-PapersOnLine* **2012**, *45*, 69−74.
(17) Lucia, S.; Finkler, T.; Engell, S. Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. *J. Process Control* **2013**, *23*, 1306−1319.
(18) Arimoto, S.; Kawamura, S.; Miyazaki, F. Bettering operation of robots by learning. *Journal of Robotic Systems* **1984**, *1*, 123−140.
(19) Jang, T.-J.; Choi, C.-H.; Ahn, H.-S. Iterative learning control in feedback systems. *Automatica* **1995**, *31*, 243−248.
(20) Wang, L.; Yu, J.; Zhang, R.; Li, P.; Gao, F. Iterative Learning Control for Multiphase Batch Processes With Asynchronous Switching. *IEEE Trans. Syst., Man Cybern., Syst.* **2019**, DOI: 10.1109/TSMC.2019.2916006.
(21) Liu, T.; Gao, F.; Wang, Y. IMC-based iterative learning control for batch processes with uncertain time delay. *J. Process Control* **2010**, *20*, 173−180.
(22) Amann, N.; Owens, D. H.; Rogers, E. Iterative learning control using optimal feedback and feedforward actions. *Int. J. Control* **1996**, *65*, 277−293.
(23) Luo, W.; Wang, L.; Zhang, R.; Gao, F. 2D Switched Model-based Infinite Horizon LQ Fault-tolerant Tracking Control for Batch Process. *Ind. Eng. Chem. Res.* **2019**, *58*, 9540.
(24) Liu, T.; Gao, F. Robust two-dimensional iterative learning control for batch processes with state delay and time-varying uncertainties. *Chem. Eng. Sci.* **2010**, *65*, 6134−6144.
(25) Shen, Y.; Wang, L.; Yu, J.; Zhang, R.; Gao, F. A hybrid 2D fault-tolerant controller design for multi-phase batch processes with time delay. *J. Process Control* **2018**, *69*, 138−157.
(26) Lu, J.; Cao, Z.; Zhang, R.; Gao, F. Nonlinear monotonically convergent iterative learning control for batch processes. *IEEE Trans. Ind. Electron.* **2018**, *65*, 5826−5836.
(27) Bristow, D. A.; Tharayil, M.; Alleyne, A. G. A survey of iterative learning control. *IEEE Control Syst.* **2006**, *26*, 96−114.
(28) Ahn, H.-S.; Chen, Y.; Moore, K. L. Iterative learning control: Brief survey and categorization. *IEEE Trans. Syst., Man, Cybern., Part C* **2007**, *37*, 1099−1121.
(29) Wang, Y.; Gao, F.; Doyle, F. J., III Survey on iterative learning control, repetitive control, and run-to-run control. *J. Process Control* **2009**, *19*, 1589−1600.
(30) Shen, D.; Wang, Y. Survey on stochastic iterative learning control. *J. Process Control* **2014**, *24*, 64−77.
(31) Morari, M.; Lee, J. H. Model predictive control: past, present and future. *Comput. Chem. Eng.* **1999**, *23*, 667−682.
(32) Qin, S. J.; Badgwell, T. A. A survey of industrial model predictive control technology. *Control Eng. Pract.* **2003**, *11*, 733−764.
(33) Grüne, L.; Pannek, J. *Nonlinear Model Predictive Control* **2017**, 45−69.
(34) Oh, S.-K.; Park, B. J.; Lee, J. M. Point-to-point iterative learning model predictive control. *Automatica* **2018**, *89*, 135−143.
(35) Chen, Y.; Chu, B.; Freeman, C. T. Point-to-point iterative learning control with optimal tracking time allocation. *IEEE Trans. Control Syst. Technol.* **2018**, *26*, 1685−1698.

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

(36) Srinivasan, B.; Palanki, S.; Bonvin, D. Dynamic optimization of batch processes: I. Characterization of the nominal solution. *Comput. Chem. Eng.* **2003**, *27*, 1−26.
(37) Aydin, E.; Bonvin, D.; Sundmacher, K. Dynamic optimization of constrained semibatch processes using Pontryagins minimum principleAn effective quasi-Newton approach. *Comput. Chem. Eng.* **2017**, *99*, 135−144.
(38) Jang, H.; Lee, J. H.; Biegler, L. T. A robust NMPC scheme for semi-batch polymerization reactors. *IFAC-PapersOnLine* **2016**, *49*, 37−42.
(39) Ljung, L. *System Identification: Theory for the User*; Pearson, 1999.
(40) Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press, 2004.
(41) Long, Y.; Xie, L.; Liu, S. Iterative Learning Economic Model Predictive Control. *arXiv Preprint* **2018**, arXiv:1801.02969.
(42) Bone, G. M. A novel iterative learning control formulation of generalized predictive control. *Automatica* **1995**, *31*, 1483−1487.
(43) Lee, K. S.; Chin, I.-S.; Lee, H. J.; Lee, J. H. Model predictive control technique combined with iterative learning for batch processes. *AIChE J.* **1999**, *45*, 2175−2187.
(44) Shi, J.; Gao, F.; Wu, T.-J. Single-cycle and multi-cycle generalized 2D model predictive iterative learning control (2D-GPILC) schemes for batch processes. *J. Process Control* **2007**, *17*, 715−727.
(45) Lee, J. H.; Lee, K. S.; Kim, W. C. Model-based iterative learning control with a quadratic criterion for time-varying linear systems. *Automatica* **2000**, *36*, 641−657.
(46) Shen, D.; Xu, J. An iterative learning control algorithm with gain adaptation for stochastic systems. *IEEE Trans. Autom. Control* **2019**, DOI: 10.1109/TAC.2019.2925495
(47) Chin, I.; Qin, S. J.; Lee, K. S.; Cho, M. A two-stage iterative learning control technique combined with real-time feedback for independent disturbance rejection. *Automatica* **2004**, *40*, 1913−1922.
(48) Lu, J.; Cao, Z.; Wang, Z.; Gao, F. A two-stage design of two-dimensional model predictive iterative learning control for non-repetitive disturbance attenuation. *Ind. Eng. Chem. Res.* **2015**, *54*, 5683−5689.
(49) Lu, J.; Cao, Z.; Gao, F. Multipoint iterative learning model predictive control. *IEEE Trans. Ind. Electron.* **2019**, *66*, 6230−6240.
(50) Lee, K. S.; Lee, J. H. Convergence of constrained model-based predictive control for batch processes. *IEEE Trans. Autom. Control* **2000**, *45*, 1928−1932.
(51) Liu, T.; Wang, Y. A synthetic approach for robust constrained iterative learning control of piecewise affine batch processes. *Automatica* **2012**, *48*, 2762−2775.
(52) Scokaert, P. O.; Mayne, D. Q.; Rawlings, J. B. Suboptimal model predictive control (feasibility implies stability). *IEEE Trans. Autom. Control* **1999**, *44*, 648−654.
(53) Mayne, D. Q.; Rawlings, J. B.; Rao, C. V.; Scokaert, P. O. Constrained model predictive control: Stability and optimality. *Automatica* **2000**, *36*, 789−814.
(54) Rawlings, J. B. Tutorial overview of model predictive control. *IEEE Control Syst.* **2000**, *20*, 38−52.
(55) Lu, J.; Cao, Z.; Gao, F. A stable two-time dimensional (2D) model predictive control with zero terminal state constraints for constrained batch processes. *IFACPapersOnLine* **2015**, *48*, 514−519.
(56) Rosolia, U.; Borrelli, F. Learning model predictive control for iterative tasks. a datadriven control framework. *IEEE Trans. Autom. Control* **2018**, *63*, 1883−1896.
(57) Lu, J.; Cao, Z.; Gao, F. Ellipsoid invariant set-based robust model predictive control for repetitive processes with constraints. *IET Control Theory Appl.* **2016**, *10*, 1018−1026.
(58) Lu, J.; Gao, F. A polytopic invariant set based iterative learning model predictive control. *IFAC-PapersOnLine* **2019**, *52*, 649−654.
(59) Cao, Z.; Yang, Y.; Lu, J.; Gao, F. Constrained two dimensional recursive least squares model identification for batch processes. *J. Process Control* **2014**, *24*, 871−879.
(60) Rho, H.-J.; Huh, Y.-J.; Rhee, H.-K. Application of adaptive model-predictive control to a batch MMA polymerization reactor. *Chem. Eng. Sci.* **1998**, *53*, 3729−3739.
(61) Cao, Z.; Yang, Y.; Yi, H.; Gao, F. Priori knowledge-based online batch-to-batch identification in a closed loop and an application to injection molding. *Ind. Eng. Chem. Res.* **2016**, *55*, 8818−8829.
(62) Cao, Z.; Zhang, R.; Lu, J.; Gao, F. Two-time dimensional recursive system identification incorporating priori pole and zero knowledge. *J. Process Control* **2016**, *39*, 100−110.
(63) Cao, Z.; Zhang, R.; Lu, J.; Gao, F. Online identification for batch processes in closed loop incorporating priori controller knowledge. *Comput. Chem. Eng.* **2016**, *90*, 222−233.
(64) Cao, Z.; Lu, J.; Zhang, R.; Gao, F. Online average-based system modelling method for batch process. *Comput. Chem. Eng.* **2018**, *108*, 128−138.
(65) Chi, R.; Hou, Z.; Jin, S.; Huang, B. An improved data-driven point-to-point ILC using additional on-line control inputs with experimental verification. *IEEE Trans. Syst., Man, Cybern., Syst.* **2019**, *49*, 687−696.
(66) Lorenzen, M.; Cannon, M.; Allgöwer, F. Robust MPC with recursive model update. *Automatica* **2019**, *103*, 461−471.
(67) Wang, Y.; Tuo, J.; Zhao, Z.; Gao, F. Optimal structure of learning-type set-point in various set-point-related indirect ILC algorithms. *Ind. Eng. Chem. Res.* **2011**, *50*, 13427−13434.
(68) Tayebi, A.; Chien, C.-J. A unified adaptive iterative learning control framework for uncertain nonlinear systems. *IEEE Trans. Autom. Control* **2007**, *52*, 1907−1913.
(69) Chien, C.-J.; Tayebi, A. Further results on adaptive iterative learning control of robot manipulators. *Automatica* **2008**, *44*, 830−837.
(70) Tan, K.; Zhao, S.; Xu, J.-X. Online automatic tuning of a proportional integral derivative controller based on an iterative learning control approach. *IET Control Theory Appl.* **2007**, *1*, 90−96.
(71) Wang, Y.; Liu, T.; Zhao, Z. Advanced PI control with simple learning set-point design: Application on batch processes and robust stability analysis. *Chem. Eng. Sci.* **2012**, *71*, 153−165.
(72) Liu, T.; Wang, X. Z.; Chen, J. Robust PID based indirect-type iterative learning control for batch processes with time-varying uncertainties. *J. Process Control* **2014**, *24*, 95−106.
(73) Hao, S.; Liu, T.; Gao, F. PI based indirect-type iterative learning control for batch processes with time-varying uncertainties: A 2D FM model based approach. *J. Process Control* **2019**, *78*, 57−67.
(74) Shi, J.; Zhou, H.; Cao, Z.; Jiang, Q. A design method for indirect iterative learning control based on two-dimensional generalized predictive control algorithm. *J. Process Control* **2014**, *24*, 1527−1537.
(75) Wang, Y.; Zhang, J.; Zeng, F.; Wang, N.; Chen, X.; Zhang, B.; Zhao, D.; Yang, W.; Cobelli, C. Learning can improve the blood glucose control performance for type 1 diabetes mellitus. *Diabetes Technol. Ther.* **2017**, *19*, 41−48.
(76) Betti, G.; Farina, M.; Scattolini, R. A robust MPC algorithm for offset-free tracking of constant reference signals. *IEEE Trans. Autom. Control* **2013**, *58*, 2394−2400.
(77) Limón, D.; Alvarado, I.; Alamo, T.; Camacho, E. F. MPC for tracking piecewise constant references for constrained linear systems. *Automatica* **2008**, *44*, 2382−2387.
(78) Tanaskovic, M.; Fagiano, L.; Smith, R.; Morari, M. Adaptive receding horizon control for constrained MIMO systems. *Automatica* **2014**, *50*, 3019−3029.
(79) Adetola, V.; Guay, M. Robust adaptive MPC for constrained uncertain nonlinear systems. *Int. J. Adapt. Control Signal Process.* **2011**, *25*, 155−167.
(80) Heirung, T. A. N.; Foss, B.; Ydstie, B. E. MPC-based dual control with online experiment design. *J. Process Control* **2015**, *32*, 64−76.
(81) Heirung, T. A. N.; Ydstie, B. E.; Foss, B. Dual adaptive model predictive control. *Automatica* **2017**, *80*, 340−348.

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173

(82) Shen, D.; Xu, J.-X. A novel Markov chain based ILC analysis for linear stochastic systems under general data dropouts environments. *IEEE Trans. Autom. Control* **2017**, *62*, 5850−5857.
(83) Shen, D.; Zhang, W.; Xu, J.-X. Iterative learning control for discrete nonlinear systems with randomly iteration varying lengths. *Syst. Control Lett.* **2016**, *96*, 81−87.
(84) Lu, J.; Cao, Z.; Zhang, R.; Bo, C.; Gao, F. A tube feedback iterative learning control for batch processes. *IFAC-PapersOnLine* **2018**, *51*, 785−790.
(85) Klenske, E. D.; Zeilinger, M. N.; Scholkopf, B.; Hennig, P. Gaussian process-based predictive control for periodic error correction. *IEEE Trans. Control Syst. Technol.* **2016**, *24*, 110−121.
(86) Cao, Z.; Filatova, T.; Oyarźun, D. A.; Grima, R. Multi-scale bursting in stochastic gene expression. *bioRxiv* **2019**, 717199.
(87) Holehouse, J.; Grima, R. Revisiting the reduction of stochastic models of genetic feedback loops with fast promoter switching. *bioRxiv* **2019**, 657718.
(88) Kuritz, K.; Zeng, S.; Allgöwer, F. Ensemble Controllability of Cellular Oscillators. *IEEE Control Syst. Lett.* **2019**, *3*, 296−301.
(89) Zeng, S.; Allgoewer, F. A moment-based approach to ensemble controllability of linear systems. *Syst. Control Lett.* **2016**, *98*, 49−56.
(90) Luo, L.; Yao, Y.; Gao, F. Cost-effective process modeling and optimization methodology assisted by robust migration techniques. *Ind. Eng. Chem. Res.* **2015**, *54*, 5736−5748.
(91) Luo, L.; Yao, Y.; Gao, F. Bayesian improved model migration methodology for fast process modeling by incorporating prior information. *Chem. Eng. Sci.* **2015**, *134*, 23−35.
(92) Lu, J.; Yao, Y.; Gao, F. Model migration for development of a new process model. *Ind. Eng. Chem. Res.* **2009**, *48*, 9603−9610.
(93) Lu, J.; Gao, F. Model migration with inclusive similarity for development of a new process model. *Ind. Eng. Chem. Res.* **2008**, *47*, 9508−9516.

DOI: 10.1021/acs.iecr.9b02370
Ind. Eng. Chem. Res. 2019, 58, 17164−17173