On the impact of discreteness and abstractions on modelling noise in gene regulatory networks

Chiara Bodei[a], Luca Bortolussi[b,g,h], Davide Chiarugi[g,c,*], Maria Luisa Guerriero[d], Alberto Policriti[e], Alessandro Romanel[f]

[a] *Dip. di Informatica, Università di Pisa, Italy*
[b] *Dip. di Matematica e Geoscienze, Università di Trieste, Italy*
[c] *Max Planck Institut of Colloids and Interfaces, Potsdam, Germany*
[d] *Systems Biology Ireland, University College Dublin, Ireland*
[e] *Dip. di Matematica e Informatica, Università di Udine, Udine, Italy*
[f] *Centre for Integrative Biology (CIBIO), University of Trento, Italy*
[g] *CNR-ISTI, Pisa*
[h] *Modelling and Simulation Group, University of Saarland, Campus E 1 3 Saarbruecken, Germany*

Abstract

In this paper we explore the impact of different forms of model abstraction and the role of discreteness on the dynamical behaviour of a simple model of gene regulation where a transcriptional repressor negatively regulates its own expression. We first investigate the relation between a minimal set of parameters and the system dynamics in a purely discrete stochastic framework, with the twofold purpose of providing an intuitive explanation of the different behavioural patterns exhibited and of identifying the main sources of noise. Then, we explore the effect of combining hybrid approaches and quasi-steady state approximations on model behaviour (and simulation time), to understand to what extent dynamics and quantitative features such as noise intensity can be preserved.

Keywords: gene regulatory networks, discrete modelling, hybrid system, quasi-steady state approximation, stochastic noise

1. Introduction

Regulating gene expression is a complex work of orchestration, where the instruments play with improvised variations without a fixed music sheet. Under this regard, the regulation process, in which DNA drives the synthesis of cell products such as RNA, and proteins, can be thought of as a stochastic process.

*Corresponding author
Email addresses: chiara@di.unipi.it (Chiara Bodei), luca@dmi.units.it (Luca Bortolussi), davide.chiarugi@mpikg.mpg.de (Davide Chiarugi), maria.guerriero@astrazeneca.com (Maria Luisa Guerriero), policriti@dimi.uniud.it (Alberto Policriti), romanel@science.unitn.it (Alessandro Romanel)

The amount of RNA and proteins in living cells must be thoroughly tuned, both to manage effectively housekeeping functions and to respond promptly to upcoming needs (e.g. to adapt to environmental changes). To this end, gene expression is equipped with several control mechanisms and strategies that grant both reliability and flexibility in terms of throughput. Nevertheless, when observed at the single cell level, the amount of molecules involved in gene expression and its regulation fluctuates randomly [1]. This stochastic effect at the molecular level turns out to play important roles in conditioning cell-scale phenomena, e.g. cellular fate decision making, incomplete penetrance or enhanced fitness through phenotypes variability [1].

Pioneering works [2, 3] showed that gene expression in populations of genotypically identical cells (i.e. with the same genetic constitution) is highly variable even when epigenetic conditions (i.e. the ones that result from external rather than genetic influence) are kept constant. In [4], the authors identified such a population variability and decomposed the extrinsic and intrinsic contributions therein. Also, in [5], it is shown that this variability it is controllable.

Recent developments in experimental techniques (see [1] for a review) have made it possible to detect and count individual molecules and, therefore, to measure the amount of mRNA and proteins in single cells. These measurements have clearly shown that the number of mRNA and proteins can vary significantly from cell to cell. This variability is caused by the fundamentally stochastic nature of the biochemical events involved in gene expression [1] and is studied, e.g., in [6, 7, 8], where population-level mathematical frameworks are introduced and applied.

As a consequence, the phenotypical variability (i.e. the variability resulting from the interaction of the genotype with the environment) exhibited by populations of identical organisms can be directly caused by stochasticity at the single cell level. Thus, it is becoming clear that noise and stochasticity underlie critical events in cell's life such as differentiation and decision making [9]. Moreover, some authors suggest that random phenotypic switching can represent an efficient mechanism for adapting to fluctuating environments (see, e.g., [9, 10]). These findings have raised new interest in analysing the role of noise in gene expression.

The regulation process includes multiple steps leading from gene transcription to the translation of the resulting mRNA to obtain the encoded protein. Each step represents a possible control point, where several biochemical mechanisms play a role (see [11] for a comprehensive review and [12] for an application to model selection). Characterising the contributions of each single control point in the regulation process of gene expression is a complex task. Identifying which strategies come into play in generating or dampening noisy behaviours is even more challenging. The extensively studied regulation paradigm represented by the feedback control strategy can be used to explain the mechanisms controlling gene transcription and translation. In such mechanisms, the global intensity of the feedback depends on parameters related to every single control point. Computational methods can significantly help investigating the synergistic mechanisms underlying the regulation of gene expression.

There are several modelling strategies that can lead to different kinds of computational models, depending on both the particular purposes and on the features of the available data. Models of biological systems proposed in the literature can vary in terms of the abstraction level used to represent molecular amounts (in this regard a model can be either *discrete* or *continuous*) and in terms of the underlying paradigm used for describing the temporal evolution of the system, which can be either *deterministic* or *stochastic*.

Ordinary Differential Equations (ODEs) have been extensively used over the years to describe the behaviour of biological processes. They provide modellers with powerful and well assessed analysis and simulation techniques. Nevertheless, ODE models implicitly assume continuous and deterministic change of concentrations, abstracting away noise and randomness due to stochastic fluctuations. This, for example, makes it difficult to capture qualitatively different outcomes arising from identical initial conditions (e.g., [13, 14, 3]).

One way to represent noise is to couple a Gaussian noise term to the model equations, obtaining a set of Stochastic Differential Equations (SDEs). This approach succeeded in gaining insights on the stochasticity of gene expression underlying circadian clocks [15] and genetic switches [16]. However, continuous methods still fail to properly describe various phenomena arising from stochastic fluctuations in systems involving small copy numbers of molecules [17], as in the case of bimodal mRNA distributions generated by long transcriptional bursts, during which mRNA level approaches a new steady state [1].

The copy numbers of molecules and individual entities in the cell space are discrete and the reactions in which they are involved are stochastic events. Consequently, approaches based on a discrete and stochastic formulation, such as the ones built upon Continuous-Time Markov Chains (CTMCs) [18, 19], have been successfully introduced to overcome the modelling limitations of continuous methods. Stochastic systems are formally represented through a chemical master equation and have also been studied, rather directly, in the form of autocatalytic reactions systems, with approaches such as the one described in [20].

However, since the analytic solution of the underlying equation is often infeasible for real size systems, these models are usually studied resorting to simulation approaches, mostly based on (variants of) Gillespie's stochastic simulation algorithm [21]. Unfortunately, in some cases, even numerical simulation can be computationally very expensive. A compromise between accuracy and efficiency can be obtained by combining discrete and continuous evolution in so-called hybrid approaches [22, 23, 24]. In this context, we recall also [25, 26], where hybrid approaches for stochastic simulation of gene networks have been developed.

In this work, we consider a simple model of gene regulation in a transcription/translation genetic network, where a transcriptional repressor negatively regulates its own expression. This model has been widely studied (e.g. [27, 28]), because it is a minimal system that explicitly describes the processes of transcription and translation and because it is a basic component of many complex biological systems. Despite its apparent simplicity, understanding its behaviour is not easy, because this is governed in a non-trivial way by several quantitative parameters. Actually, different parameter combinations lead to a range of qual-

itatively different dynamics. In particular, in [27], the intensity of noise in this system is analysed in terms of various parameters, with special emphasis on the strength of the negative feedback. That paper demonstrates how the application of engineering principles to the role of feedbacks in a biological context could be misleading. The authors show, indeed, that noise generally increases with feedback strength, in contrast to the common knowledge. They also relate the possible different dynamical regimes with the different regions of the parameter space. The overall behaviour emerges from the complex interaction between feedback strength and other parameters of the system, governing the dynamics of the protein and of the mRNA.

Starting from the analysis proposed in [27], we investigate the model with two goals in mind: (i) identifying the *parameters that play a key role in the regulation* process, by systematically studying the impact of parameters variations on the global dynamics of the system. In this way, we establish a link between the parameter space and the observed temporal patterns, i.e. the diverse behavioural phenotypes; (ii) quantifying the *impact of each reaction* of the modelled system on the overall dynamics and on noise patterns. This allows us to identify those reactions that, having a minor influence on the global noise pattern, can be safely approximated in a deterministic fashion.

At a higher level and on a longer term, we aim at setting up a systematic strategy for correctly building hybrid models of biochemical systems. In such models, only the most relevant sources of noise will be represented via a fully detailed stochastic description.

The handy dimension of our model of gene regulation allows us to play with different models and techniques. On the one hand, we study a stochastic model of the negative feedback loop to construct an exact picture of its possible behavioural patterns and of the effects of noise. This precision comes with a high computational cost, especially for certain parameter sets. On the other hand, we systematically apply various forms of model abstraction, in order to mitigate the inefficiency of exact stochastic simulation methods. In particular, we abstract the discrete stochastic dynamics into continuous deterministic dynamics for some of the model components, thus obtaining a stochastic hybrid model. Afterwards, we simplify the model structure reducing the number of model components and reactions, by applying quasi-steady state approximations (QSSAs).

These abstractions are not blindly applied, but are rather considered only in the regions of the parameter space for which the hypotheses underlying these techniques are (reasonably) satisfied.

We believe that the methodology we adopt is a reliable approach for modelling in systems biology, where too often *in silico* techniques based on abstractions or approximations are used off-the-shelf, without any concern on their applicability and faithfulness.

2. The model

The model we consider, described among others in [27, 28], is a genetic network composed by the reactions reported in Table 1 and schematically rep-

resented in Figure 1(a).

The model is a minimal system that describes the self-regulated transcription/translation of a gene into a protein. The gene (G) is transcribed into an mRNA molecule (M), which in turn is translated into a protein P. P regulates its own expression by means of a negative feedback loop: P can bind to G, making it switch from its free active state to an inactive state (Gb). Finally, both M and P can be degraded. All reactions are assumed to follow mass action kinetics, i.e. the speed of the reaction is proportional to the product of the amounts of each reactant and a kinetic constant. In the following, the amounts of P, M, G, and Gb are denoted by X_P, X_M, X_G, and X_{Gb}, respectively.

Similarly to [27], we ignore copy-number variations (CNVs) and assume to have a single copy of the gene. The reason for this assumption is that CNVs may rise from different structural rearrangements (e.g. deletions, duplications, inversions) and can involve different genomic regions (e.g. enhancer, promoter, coding), hence having potentially different and sometimes unpredictable effects on the expression of the surrounding/overlapping genes. These effects may be difficult to model and may introduce a level of variability that goes beyond the aim of this work.

rate constant	**reaction**	**description**
$prod_M$	$G \to G + M$	mRNA production (transcription)
$prod_P$	$M \to M + P$	protein production (translation)
deg_M	$M \to \emptyset$	mRNA degradation
deg_P	$P \to \emptyset$	protein degradation
$bind_P$	$G + P \to Gb$	repressor binding
$unbind_P$	$Gb \to G + P$	repressor unbinding

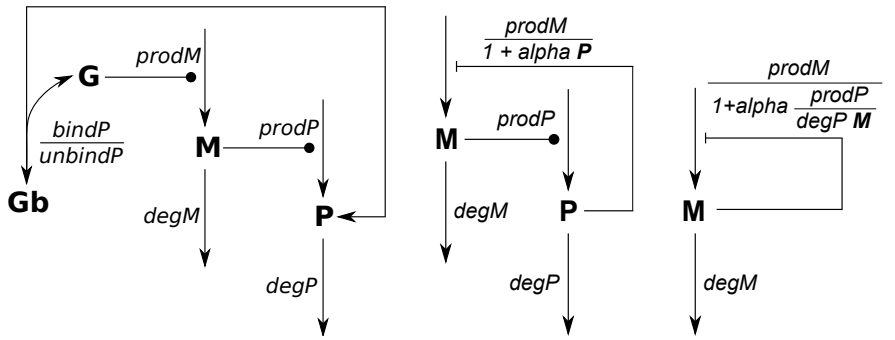
Table 1: Biochemical reactions of the self-repressing gene network considered. All reactions follow mass action kinetics and the units of their rate constants are s^{-1}.

In [27], a very detailed phenomenological description of the different behaviours of the model has been carried out, changing parameters in order to explore a large portion of the parameters' space. The authors claim that this simple feedback network has a *counter-intuitive* behaviour: while, in engineering, negative feedbacks are assumed to be a mechanism to decrease noise, the authors show that in this network, instead, noise increases with feedback strength.

Following on from those results, in this section we provide a simpler picture of the possible behavioural patterns of the model; we found out, in fact, that the seemingly counter-intuitive behaviour can be explained by analysing a combination of a minimal set of parameters.

First of all, with a preliminary high level analysis of the stochastic dynamics of M, P, gene repression, and of the interactions between dynamical regimes involved in the model (partly reported in the Supplementary Material, Appendix A), we identified the following two key parameters:

1) P binding/unbinding ratio $\alpha = bind_P/unbind_P$;



(a) Model with explicit protein binding/unbinding

(b) Model with QSSA on binding/unbinding

(c) Model with QSSA on binding/unbinding and protein

Figure 1: Diagrammatic representation of the gene regulatory model described in this section and of its variants with QSSAs described in Section 4. Arrows indicate the direction of reactions; dot-headed and T-headed lines represent reaction catalysts (i.e. species which are on both the left- and the right-hand side of the reaction) with, respectively, positive and negative roles.

2) P/M degradation ratio $\beta = deg_P/deg_M$.

The first parameter α was used in [27] as an indicator of the *strength of the feedback regulation*: the higher the value of α, the stronger the repression and the smaller the number of mRNA molecules on average. While in [27] the authors fix the binding rate and vary the unbinding rate (thus increasing feedback strength by increasing the binding strength), in our work we fix the unbinding rate and vary the binding rate (thus increasing feedback strength by increasing the binding affinity), hence using a different mechanism to represent repression intensity. This choice has an impact on the dynamics of the network, in particular on the behaviour of the bursty protein regime and on the feasibility of the QSSA (see also Section 3 and Appendix A in the Supplementary Material).

The second parameter β is the *ratio between the half-life of M and the half-life of P*. From a dynamical point of view, it describes the speed at which production and degradation of the protein reach equilibrium, relative to the mRNA. Essentially, it captures how the protein level reacts to fluctuations at the mRNA level: the higher the value of β, the faster P's response. This means that X_M and X_P will be highly correlated for high β, and so noise at the level of X_M level will propagate more effectively to X_P.

The values of the stochastic rate constants we use are based on those from [27] and their units are s^{-1}. After a prescreening in which we used several values according to [27] for stochastic simulations, we identified the values that proved to be representative in reproducing the complete gamut of all the interesting observable phenotypes. Therefore, in this work we choose to vary α in the set $\{0.0166, 16.6, 16600\}$, and β in the set $\{0.01, 1, 100\}$. In the following, we refer to these three values for α and β as *low* (**L**), *medium* (**M**), and *high* (**H**), and

we refer to the combinations of parameters with the names listed in Table 2.

The parameter α is varied by fixing the value for $unbind_P$ to 1 and modifying $bind_P$, defined as $\alpha \cdot unbind_P$. Similarly, we fixed the value for deg_M to 0.001 and vary $deg_P = \beta \cdot deg_M$ while varying β. Moreover, we define P production rate as the product of P degradation rate and the steady state value for P per M molecules, i.e. $prod_P = deg_P \cdot P_{steady}$ with P_{steady} equal to 3500. We have chosen this value because it allows us to obtain biologically meaningful numbers of proteins for all the parameter combinations. However, reasonably changing this parameter does not disrupt the observed behaviours. This enables us to explore parameter combinations with a low degradation rate, ensuring that translation rate does not become too large considering a maximum rate constant of 10^8 – 10^{10} $\mathrm{M}^{-1}\mathrm{s}^{-1}$ [29] (see Appendix B in the Supplementary Material for further details). We finally fix $prod_M$ to 35 considering an average RNA polymerase transcription rate of 24 – 79 nucleotides/s and 1100 base pairs as the average size of an mRNA molecule [30], i.e. an mRNA molecule is produced (on average) in 14 – 46 s.

Analysis of stochastic behaviour. We proceed now by building a fully stochastic model from our set of reactions, in order to undertake a thorough stochastic analysis of its dynamics. The simulation tool we use is COPASI [31], a framework that provides us with all the algorithms needed to run stochastic, hybrid, and deterministic simulations.

In order to outline the possible dynamics, we partition our parameter space into different regions, by creating nine different versions of our model, one for each parameter combination. For each model we then set the initial amounts of M and P to the corresponding steady state values and run 1000 stochastic simulations with limit time 10000 s.[1]

As listed in Table 2, for combinations having low α, the initial X_P and X_M are 86000 and 25, respectively. For combinations having medium α, the initial X_P and X_M are 2700 and 1, respectively. Finally, for combinations having high α, the initial X_P and X_M are 86 and 0, respectively.

Figure 2 shows, for each of the nine versions of the model, one representative simulated trajectory of X_P together with its coefficient of variation (CV) and the correlation between X_P and X_M at time $t = 10000$. A comprehensive summary of population statistics of X_P and X_M is reported in Table 3 (top panel), and further details are in the Supplementary Material.

[1] We empirically identified 10000 seconds as a limit time that is long enough to allow us to observe the specific steady state dynamics of P and M in all the parameters combinations we considered. Moreover, we are taking the population average at the chosen time 10000, and not the time average along a single trajectory, as commonly done. However, in our context, these two approaches are equivalent as all the models we consider are ergodic. This is obvious for the Continuous-Time Markov Chain (CTMC) models, but it can be shown also for the hybrid models, by applying a result in [32], characterising ergodicity in terms of a Discrete-Time Markov Chain (DTMC) obtained by sampling trajectories of the Piecewise Deterministic Markov Process (PDMP) at random times.

β \ α	**0.0166**	**16.6**	**16600**
100	**L-H**	**M-H**	**H-H**
1	**L-M**	**M-M**	**H-M**
0.01	**L-L**	**M-L**	**H-L**
X_{P0}	86000	2700	86
X_{M0}	25	1	0

Table 2: Summary of the combinations of parameters considered and their respective initial values for X_P and X_M.

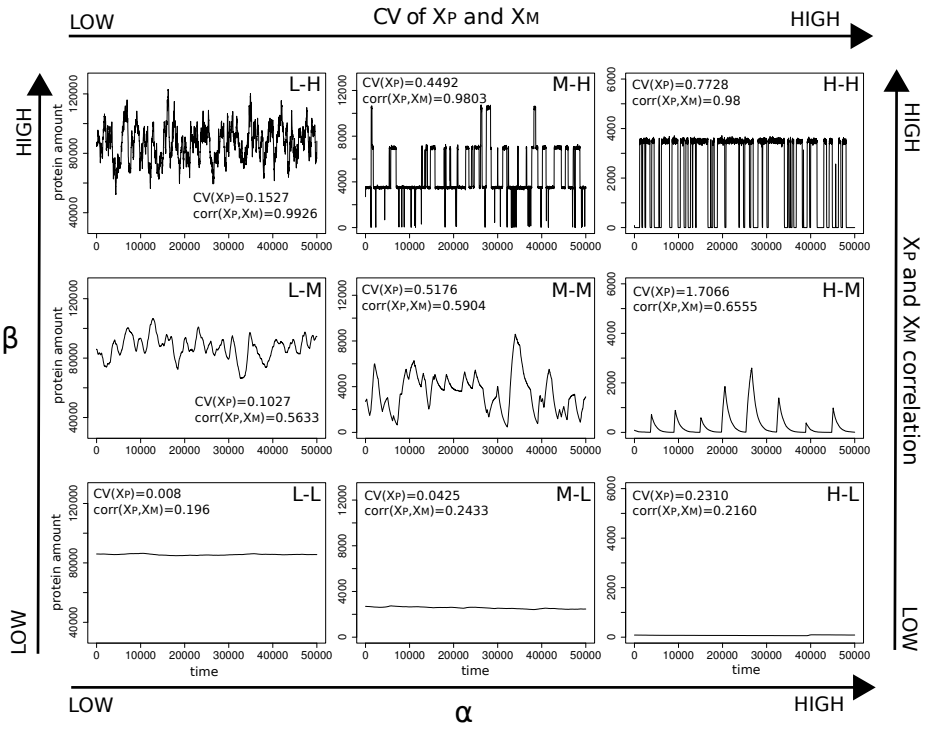


Figure 2: Sampled X_P trajectories from the explicit stochastic model computed using the Gibson-Bruck stochastic simulation algorithm [33] using different combinations of α and β values. The coefficient of variation and correlation trends indicated by the arrows and in the panels are reported in detail on the top panel of Table 3, described in Section 3, and are computed from 1000 simulated trajectories, sampled at time $t = 10000$ s.[1] Note that the trajectories are reported for a longer time ($t = 50000$ s) to better visualise the long term trends and to show that the choice of time $t = 10000$ for computing the statistics is valid.

With **L-L** combination we have that, since repression is weak, the average X_M value is relatively high and, consequently, X_P steady state is also high. Moreover, since P degradation is slower than M degradation, X_P dynamics are slower than X_M's, i.e. X_P fluctuates with a lower frequency than X_M does;

moreover, at steady state X_P behaves like a simple birth-death process with very low noise. We can say that X_P is very close to its deterministic average behaviour. It is not surprising that this is also the combination with the lowest $X_P - X_M$ correlation and coefficient of variation for X_P. In case of **M-L** combination P dynamics are still slower than M's but repression is stronger than before, causing X_M and X_P steady state values to be much lower. In particular, X_M starts to take very low values with fluctuations between 0 and 4 (see Figure S1 for instances of X_M simulated trajectories). This combination of discreteness and possible absence of M is immediately reflected in P dynamics. Indeed, such dynamics are still very close to a simple birth-death process but, as indicated also by the increment of the coefficients of variation for X_M and X_P, with higher noise. Combination **H-L** generates a strong gene repression that causes X_M to fluctuate between 0 and 1. Moreover, since P degradation is very low, the effect of repression is even amplified by the almost constant presence of P, hence X_M is 0 most of the time. As a consequence, X_P tends to slowly degrade but displays some small "steps" corresponding to those rare and short intervals in which X_M takes value 1.

By increasing β to a medium value, we have that X_M and X_P start to show a correlation higher than 0.5. In this parameter regions, each change in X_M is *followed* by a change in X_P, that is quantitatively equivalent (when scaled on X_M). With **L-M** combination X_M fluctuates at steady state around a value of 25, but in this case the coefficient of variation is higher, meaning that its dynamics are noisier. Since the number of P molecules varies in accordance with the number of M ones, we have that the noise in X_M is reflected and amplified at the level of X_P, generating a behaviour that is noisier than in **L-L**. With **M-M** combination we can observe that X_P starts to present an irregular fluctuating-like behaviour till reaching, with combination **H-M**, a behaviour that exhibits bursts. This is evidently caused by the high degree of discreteness in X_M (i.e. X_M takes very small values). In particular, bursts are emerging because P degradation is fast enough to increase the frequency of the intervals in which X_M is equal to 1 and to allow X_P to rapidly reach a low amount when X_M is back to 0. Note that the system alternates periods in which P is degraded to periods in which M is produced in few copies, and so X_P increases.

When β is high, P fluctuates with a higher frequency than M, hence it reaches stochastic equilibrium faster than M. This causes, for all three values of α, the correlation between X_P and X_M to be very high (greater than 0.98). With combination **L-H** X_P fluctuates around the mean value with high level of noise. This increase in noise in all combinations having low α can be visually observed also in the X_P distributions in Figure S2. With the intermediate combination **M-H**, the time course of P copy number starts to exhibit a multi-modal behaviour. This can be explained by observing that X_M fluctuates between 0 and 3 and considering that, for a high value of β, X_P is able to reach its steady state between successive changes in X_M level. Hence, P dynamics look like M's with discrete levels properly rescaled. From Figure S2 it is easy to see that we have three main discrete levels of X_P situated around 0, 3500, 7000 and 10500. By further increasing the repression with **H-H** combination, we end up with a

bimodal behaviour. This is clearly the consequence of the fact that, for high values of α, X_M fluctuates only between 0 and 1, a behaviour that is shown by X_P, which mainly alternates between two discrete levels. From Figure S2 we can observe that these two levels are 0 and 3500.

Our exploration of the parameter space is clearly model-specific and cannot be directly generalised to other models of gene regulatory networks; however, from a methodological point of view, our results clearly illustrate that this kind of thorough exploration is essential when studying genetic networks, in order to understand the role of feedback loops in generating different patterns of time courses of both mRNA and translated proteins. In particular, we showed that certain parameter combinations are responsible for bursty temporal patterns. While there is strong experimental evidence supporting transcriptional bursts both in prokaryotes [34] and (especially) in eukaryotes [35], the underlying biological mechanism still remains unclear. Even though many complex mechanisms, involving the biochemical machinery of DNA transcription, seem to contribute to pulsatile transcription (as suggested in [34]), our study supports the fact that gene regulation mechanisms themselves (and the relative speed of the different reactions involved) play an important role in determining this pattern of transcription.

Our case study also shows the importance of modelling without abstracting away any part of the feedback loop in order to reproduce the full spectrum of possible behaviours. Indeed, in a study similar to ours, Peccoud and Ycart [36] specified a Markovian model of gene regulation considering four parameters: λ and γ, the rates of gene switching from active to inactive state and *vice versa*, and μ and σ, the rates of mRNA transcription and degradation, respectively; while their model can describe both fluctuating and bursty temporal patterns, it does not account for bimodality in mRNA distribution. This suggests that abstracting away details even in a simple biological model can result in a loss of expressiveness. We will address this issue in Section 3.

Another result of our analysis which is worth considering in more detail is the trend of $X_P - X_M$ correlation. Evaluating this parameter in individual living cells is not an easy task, due to technical problems in detecting single molecules of mRNA and proteins in the same cell, at the same time [1]. Nevertheless this correlation analysis can give us useful insights into the process of gene expression, as it allows us to study the propagation of fluctuations in mRNA levels to the amount of produced proteins. Among the few experimental results currently available, it is worthwhile recalling those obtained by Raj et al. [35], which show that mRNA and protein levels are strongly correlated when protein lifetime is short, but that this correlation decreases when protein lifetime is long. These results are consistent with the predictions of our model, where the highest values for $X_P - X_M$ correlation are obtained when β is high.

3. Model abstractions

In the previous section, we considered the explicit stochastic model of the simple negative feedback loop. Modelling each part of the loop following Gille-

Stochastic model with explicit binding and unbinding					
α-β	mean(X_P)	CV(X_P)	mean(X_M)	CV(X_M)	cor(X_P,X_M)
L-L	85986.95	0.008	24.47	0.2037	0.196
L-M	86693.14	0.1027	24.908	0.1807	0.5633
L-H	85662.71	0.1527	24.49	0.1527	0.9926
M-L	2707.262	0.0425	0.788	0.8865	0.2433
M-M	3231.701	0.5176	0.936	0.8501	0.5904
M-H	4440.153	0.4492	1.27	0.4492	0.9803
H-L	84.058	0.2310	0.027	6.0061	0.2160
H-M	441.398	1.7066	0.165	2.2507	0.6555
H-H	2176.477	0.7728	0.622	0.7833	0.9824
Hybrid model with explicit binding and unbinding					
L-L	85857.524	0.0077	24.752	0.1970	0.2052
L-M	86285.457	0.1048	24.667	0.1839	0.5815
L-H	86956.046	0.1467	24.881	0.1479	0.9910
M-L	2724.2987	0.0426	0.797	1.1114	0.2315
M-M	3313.9473	0.5245	0.962	0.8841	0.5635
M-H	4454.3658	0.4429	1.274	0.4504	0.9803
H-L	84.8464	0.246	0.024	6.3802	0.2677
H-M	433.6356	1.7359	0.128	2.6114	0.649
H-H	2201.9696	0.7615	0.628	0.7733	0.9897
Stochastic model with QSSA on binding/unbinding					
L-L	85978.76	0.0079	25.539	0.2034	0.2178
L-M	85991.76	0.102	24.688	0.1759	0.5939
L-H	88768.61	0.14	25.37	0.1400	0.9847
M-L	2705.874	0.0429	0.751	1.1683	0.2280
M-M	3259.29	0.5105	0.921	0.9133	0.6054
M-H	4369.297	0.4506	1.249	0.4538	0.9828
H-L	85.423	0.2427	0.026	6.1236	0.1210
H-M	576.304	1.9408	0.21	2.753	0.6294
H-H	3386.57	0.3392	0.967	0.3553	0.9569
Hybrid model with QSSA on binding/unbinding					
L-L	85832.168	0.0077	24.457	0.2001	0.2124
L-M	86562.663	0.105	24.608	0.1784	0.5904
L-H	85935.261	0.1397	24.55	0.141	0.99
M-L	2727.1072	0.04431	0.799	1.1134	0.2287
M-M	3266.9487	0.5245	0.894	0.9301	0.5913
M-H	4503.4732	0.445	1.289	0.4482	0.9766
H-L	84.8650	0.2343	0.019	7.1891	0.1727
H-M	526.3546	1.8783	0.199	2.5999	0.6442
H-H	3407.6452	0.3234	0.972	0.3370	0.9689

Table 3: Statistics for different abstractions of the explicit stochastic model, computed from 1000 simulated trajectories, sampled at time $t = 10000$. [1]

spie's computational approach, we obtain an *exact* stochastic model (under the assumption that molecules are well stirred), whose analysis gives us a precise picture of the patterns of dynamical behaviour and of the effects of noise. However, even for this simple genetic network, the computational analysis can be difficult to carry out because exact simulation algorithms can be very inefficient under certain parameter configurations. This happens, for instance, for low α and high β, where P's dynamics are much faster than M's, and the number of P molecules ranges close to 100000; in this situation, the number of firing P production and degradation events is so large that explicit simulations are severely slowed down. When exact simulation is not feasible, one can either use approximate simulation algorithms, such as τ-leaping [37], or adopt some form of model abstraction (see, e.g., [38, 33, 39]).

Here we discuss methods in the latter category, focusing both on techniques that abstract the dynamics, from discrete stochastic to continuous determinis-

tic, and on techniques, like quasi-steady state approximation [40], that simplify the model structure reducing the number of reactions and components. In particular, we consider three abstraction approaches: (i) an approximation of the stochastic discrete dynamics with continuous deterministic ones, applied to all model components, thus obtaining a model defined by a set of ordinary differential equations (ODEs), (ii) a replacement of the stochastic discrete dynamics with continuous deterministic ones, localised to some of the components of the model, thus obtaining a stochastic *hybrid* model, and (iii) an approximation obtained by the removal of specific model reactions performed by the computation of quasi-steady state values.

Establishing the quality of an abstraction *a priori* is a difficult issue, as there are no simply derivable analytic formulae to invoke. We will however discuss and use heuristic arguments, to justify the use of abstractions in certain regions of the parameter space, justifying them by a though *a posteriori* statistical evaluation.

Continuous deterministic abstraction. The most common approach for dynamics abstraction is to replace the stochastic model with a fully deterministic one based on ODEs. It is known that this type of approximation is exact in the thermodynamic limit [41] and it usually works well, provided that the number of molecules in the system is sufficiently large, while it fails if noise plays a relevant role in the system dynamics. Evaluating the system of ODEs associated with our model, we can easily verify that the molecule numbers for the different species converge to stable steady states for all parameter configurations considered. This behaviour is different from the one observed by performing stochastic simulations, in which not all the parameter configurations yield patterns that converge to a steady state (see Figure 2). A configuration of parameters where the deterministic approximation of the whole system turns out to be appropriate (with respect to the behaviour of P) corresponds to low α and β (low repression and slow P). In this case, in fact, the amount of M ranges around 25 and P's slow dynamics have the effect of averaging the noise at the level of M, so that noise at the level of P is extremely low. This also holds for G's dynamics. In general, for small β the approximation is reasonable, although less accurate as feedback strength increases, corresponding to a decrease in the number of M molecules (see Figure 2).

Hybrid abstraction. From the discussion above, it follows that the inherent discreteness of M and G evolution, especially in the high repression regime, plays a central role in determining the qualitative pattern of dynamical evolution. Then, noise at the level of M propagates more or less rapidly to P's dynamics. The intrinsic noise of P's dynamics, instead, should be less relevant in this respect if the steady state of X_P is sufficiently high. Therefore, a reasonable abstraction of this model would see P as a continuous quantity evolving according to an ODE, while maintaining discrete dynamics of M and gene repression.

Mathematically, this gives rise to a *stochastic hybrid* model, belonging to the class of Piecewise Deterministic Markov Processes (PDMPs) [42]. These

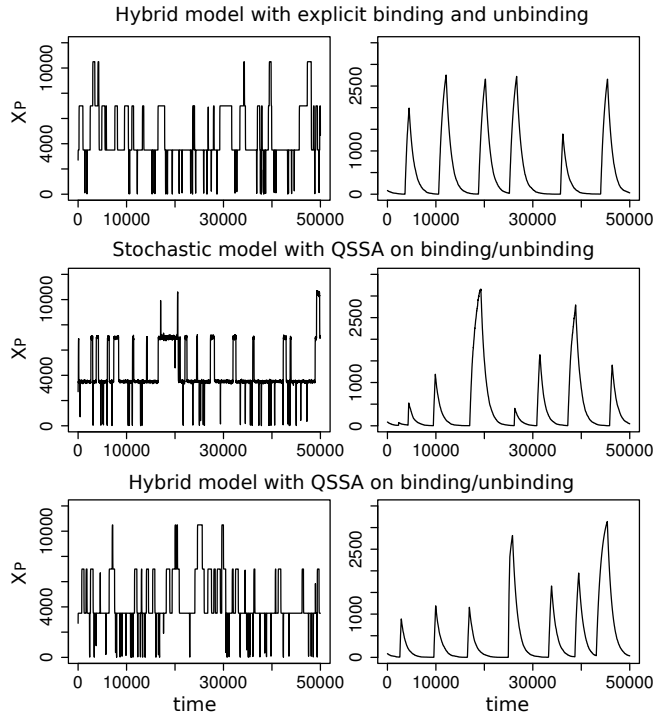


Figure 3: Comparison of multi-modal (medium α, high β) and peak-like (high α, medium β) trajectories of P for three different abstractions. Results for the stochastic model with explicit binding and unbinding can be found in the corresponding panels in Figure 2.

processes are described by a set of continuous variables and a set of discrete variables. Continuous variables are subject to continuous evolution, while discrete changes happen spontaneously at times determined by exponential distributions, as in CTMCs. In particular, as rates of discrete jumps can depend on the value of continuous variables, the discrete process is non-homogeneous in time. PDMPs can be analysed by simulation or even by numerical solution of their master equation, which is a partial differential equation, although the latter approach is computationally rather demanding [43]. Hybrid simulation algorithms for biochemical systems based on PDMPs [22, 44] usually implement some heuristic rule for partitioning species into discrete and continuous ones, in general according to their copy numbers in order to minimise errors caused by the continuous approximation. The partitioning rule can be either static or dynamic. Static partitioning is performed before simulation, after a screening of the current parameter set. Dynamic partitioning, instead, is performed during simulation: molecules that at a certain time fall below a specific threshold are treated discretely and stochastically.

We applied dynamic partitioning to our system[2], setting the continuous-to-discrete switching threshold to 10 (hence, when the number of molecules falls below 10 we switch to a fully stochastic system) and the discrete-to-continuous switching threshold to 20. The use of separate thresholds helps to avoid too many switches due to noise effects. These two thresholds have been heuristically selected to avoid unnecessary switching between discrete and stochastic models.

We expect this hybrid scheme to work well in most cases, with a possible loss of precision for high feedback repression (which reduces the overall number of P molecules).

In Table 3 and Figure 3 we compare the behaviour of the hybrid model with dynamic partitioning to the behaviour of the fully stochastic model. We can see that the hybrid simulation works very well, and essentially all behaviours of the fully stochastic model are qualitatively captured. In Figure 3 we report the most interesting combinations. At the quantitative level, the hybrid model seems to be also able to capture quite accurately the first two moments of the distribution as well as the correlation between X_M and X_P (Table 3), with some loss of precision for M and large values of α. This supports our initial conjecture that the inherent discrete and stochastic dynamics of G and M are what mainly determine noise modes of protein expression. Intrinsic fluctuations of the protein due to stochastic production and degradation are less relevant in this respect, hence can be safely abstracted away.

In terms of simulation time, the hybrid model outperforms the stochastic simulation in those parameter regions in which P's dynamics are the bottleneck for stochastic simulation (Table S6), as expected [44]. In fact, since we are approximating only P's dynamics as continuous, our hybrid abstraction will improve stochastic simulation only when the number of P production and degradation events dominates the simulation cost. In all other cases, the overhead introduced by hybrid simulation[3] can overcome the gain in computational efficiency. In our system, for instance, the hybrid approach is faster for high β. Indeed, for low α and high β, we obtain a 277-fold speed-up, while the performance is less appealing for larger feedback strength, as the P number is reduced, hence so is the translation frequency and P degradation reactions in the stochastic model. However, notice that the execution time of the hybrid model is essentially constant for all parameter sets, allowing us to study the behaviour of the system for parameter combinations in which stochastic simulation

[2]We do not consider further static partitioning, as it turned out to be not very accurate for high feedback values (data not shown). In this case, in fact, P trajectories often approach zero, and treating P as always continuous can introduce significant errors. In particular, in such an extreme, feedback strength is increased, as now P exerts a repression also when its value lies between 0 and 1.

[3]In hybrid simulation, the ODE integration engine needs to be coupled with an event detection mechanism [45] that requires to find the roots of non-linear equations to identify the firing time of stochastic events [22]. This is because the stochastic part of the process in the hybrid system is time inhomogeneous, with rates that depend on the continuous variables. Approximate hybrid strategies can avoid this overhead [22, 31], at the price of reducing the integration step of ODE solvers in order to avoid significant loss in accuracy.

is computationally unfeasible (e.g. for large values of β).

We stress that the choice of which hybrid or ODE solver to use is crucial. For large values of β, e.g. P's deterministic dynamics are stiff; in this case, a stiff solver [45] should be used. If a non-stiff integrator is employed (such as methods belonging to the explicit Runge-Kutta family [45]), then no speed-up may be observed at all (data not shown) compared to exact stochastic simulation.

Quasi-steady state approximation. A different abstraction technique consists in reducing the number of model variables and reactions, using the Quasi-Steady State Approximation (QSSA) [40, 46, 47]. The idea behind QSSA is that, if a set of reactions acting on one (or more) molecular species is very fast, then their dynamics will quickly reach an equilibrium. Therefore, one can remove these reactions from the model, assuming that the entities involved only in these reactions are at their steady state. In practice, model variables are partitioned into fast and slow, and the steady state of fast variables conditional on the slow ones being constant is computed. For a stochastic model, one obtains a steady state distribution for fast variables (assuming ergodicity). Then, a reduced system is constructed, containing only the slow species, averaging out the fast variables from the rate functions depending on them according to the previously computed steady state distribution. Since analytical expressions are rarely obtained in this way, such averaged rates can be approximated either by stochastic simulation [48, 49], or using the deterministic steady state distribution of fast variables, i.e. the one obtained from the ODE model.

In our model the binding and unbinding of the gene repressor turns out to be a natural candidate for QSSA due to the dynamics of the reaction. Indeed, QSSA of gene dynamics is assumed in the majority of models of genetic networks. For the simple binding/unbinding mechanism considered in our model, applying QSSA we obtain a nice closed form for the production rate of mRNA. In fact, keeping X_M and X_P fixed, we get that $G_{steady} = 1$ with probability $1/(1 + \alpha \cdot X_P)$.[4] Then, the two species describing the gene state (G and Gb) and the binding and unbinding reactions can be removed from the model, and the transcription rate, with the gene variable G averaged out, becomes $prod_M/(1 + \alpha \cdot X_P)$. In this case, this expression coincides with the one that can be obtained from the ODE model. The resulting model contains only four reactions and two species, and is schematically illustrated in Figure 1(b). In Figure 3 and Table 3, we show the results of stochastic simulation of this reduced model, assuming QSSA on binding and unbinding (also in this case Figure 3 reports the most interesting combinations).

A typical "rule of thumb" to apply QSSA [44, 46, 49] consists in comparing the time scales of the reactions that are to be removed by QSSA with the time scales of the remaining reactions. If the former are much smaller that the latter,

[4]Conditional on X_M and X_P, the remaining Markov chain has two states, one for $G = 1$ and one for $G = 0$. The rate of going from $G = 1$ to $G = 0$ is then $\alpha \cdot unbind_P \cdot X_P$, while the rate of going in the other direction is $unbind_P$. The steady state probability π for $G = 1$ can be obtained by solving the balance equation $\alpha \cdot unbind_P \cdot X_P \cdot \pi = unbind_P(1 - \pi)$.

i.e. the corresponding rates are faster, then QSSA is usually safe. Practically, we have to check if the binding and unbinding rates are a few orders of magnitude larger than the remaining rates. This is not always the case in our system: for large β, the unbinding rate is comparable to P's speed (more precisely, it is only one order of magnitude larger). Therefore, QSSA is predicted to be valid only for medium to low β. However, simulating the model we observed that the dynamics are qualitatively similar to those of the explicit model for all parameters and, in most cases, they are in good agreement also from a quantitative point of view. This can be explained by observing that gene repression influences directly only M and that the dynamics of binding/unbinding reactions are much faster than those of production/degradation of M and, thus, the dynamics of gene repression reach equilibrium (in the stochastic sense) much sooner than M's.

If the unbinding rate is reduced then the QSSA assumption on equilibrium of gene repression will cease to be valid. In this case, it is known that the QSSA can introduce large errors in the stochastic dynamics [46], and this is indeed the case in our model, as observed also in [27]. This is confirmed in Figure 4, where we compare the behaviour of the full and the QSSA models for different parameter combinations. In particular, we fixed β as medium and varied only α but, differently from Section 2, we fixed the binding rate to $bind_P = 0.0166$ and varied the unbinding rate accordingly to $unbind_P = bind_P/\alpha$. Note how the model with QSSA severely underestimates noise and exhibits more and much lower peaks w.r.t. the full stochastic model for large α (i.e. very low unbinding rate).

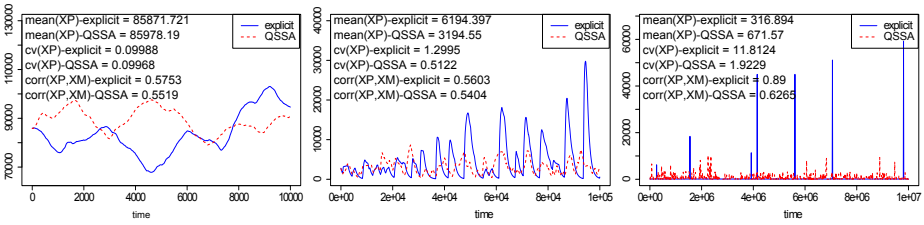


Figure 4: Comparison of P trajectories for the explicit stochastic model (continuous blue line) and the QSSA stochastic model (dotted red line) for medium β, different values of α, and variable unbinding rates (decreasing from left to right).

The application of QSSA is by no means restricted only to gene dynamics: it can be applied to any variables and sets of reactions that are sufficiently fast (with respect to the species they interact with). For instance, in our model we may apply it to P itself, when β is large.

For example, consider a variation of the model in which we apply QSSA to both G and P, for high β and varying α (with unbinding rate fixed, as in Section 2). In this case, the model has only one molecular species left, namely M, and two reactions, transcription and M degradation. In particular, we approximate the stochastic QSSA by computing reduced rates from the deterministic system, giving a rate of transcription of the form $prod_M/(1 + \alpha \cdot prod_P/deg_P \cdot X_M)$,

where X_P is replaced by its deterministic steady state value $prod_P/deg_P \cdot X_M$.[5] This version of the model is illustrated in Figure 1(c), and simulation results are shown in Figure 5. As we can see, from a qualitative point of view, the dynamics are preserved for low and medium α. From a quantitative point of view, as well, if we compare the X_M distributions for the full stochastic model and those for this minimal model, we obtain that the two models are very similar for low and medium α (Table S5). We stress that QSSA on P can be applied only for large values of β, i.e. when the correlation between P and M is close to 1.

It is clear that the hybrid abstraction and the QSSA can be combined together [44]. In Figure 3 and Table 3, we show the results of hybrid simulation of the reduced model where QSSA is used to abstract binding and unbinding. Also in this case the dynamics are essentially similar to those of the explicit model. This can be explained following and combining the considerations we already made for the two abstractions separately.

Some statistical insights. Looking at the histograms of the distribution of the amounts of P and M at time $t = 10000$ s (see Figures S2, S3, S4, S5 for the distributions of P) for all the models considered, we can observe that the distributions look visually quite similar. This is confirmed by statistically testing the difference between the empirical distributions of the abstract models and the empirical distribution of the stochastic model.

In particular, we computed the *histogram distance* [50] and we performed both a *Mann-Withney test* [51] and a *Kolmogorov-Smirnov test* [52] (Tables S1, S2, and S3). The histogram distance is used in Monte Carlo simulation to calculate the distance between histogram functions that approximate probability density functions of different group of samples. The Mann-Whitney test is a non-parametric test used to check for a significant statistical dominance between two samples. Its two-sided version can be used to check for a significant difference between two populations. It is also used as a test for detecting a difference between locations (means or medians), assuming that the two samples come from distributions with the same shape. The (two-samples) Kolmogorov-Smirnov test, instead, is a classical test for goodness of fit between two samples, testing the null hypothesis that the two samples come from the same distribution (for the two-sided case). These three tests can detect different aspects of differences between distributions, hence we present the results of all of them in the Supplementary Material.

All these tests reported no significant difference between the empirical distributions for most parameter combinations. However, some statistically significant differences are detected, mainly for large values of α and in relation to the

[5] If we apply the recipe of stochastic QSSA [40, 49], to get the correct QSSA mRNA production rate we need to compute the steady state distribution of the gene and the protein simultaneously, and then compute the marginal probability that $G_{steady} = 1$. In this case, we obtain that this probability equals $\sum_{k=0}^{\infty} \frac{1}{1+\alpha k} \frac{1}{k!} e^{-\lambda} \lambda^k$, with $\lambda = P_{steady} \cdot X_M$. However, we are not aware of a closed form solution of the previous summation, hence we preferred to stick to the simpler deterministic approximation.

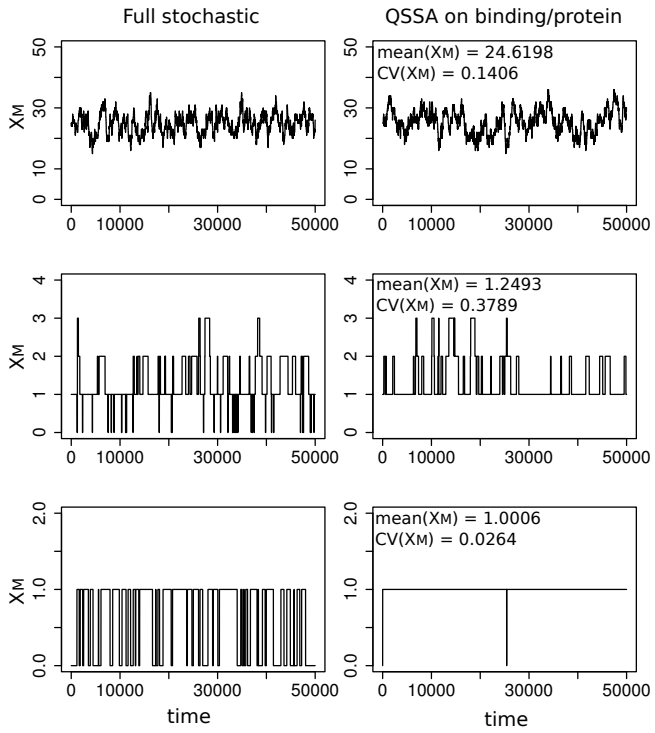


Figure 5: Comparison of M trajectories for the explicit stochastic model and the stochastic model with QSSA both on gene repression and protein dynamics, for high β and different values of α (increasing from top to bottom).

hybrid approach. Moreover, the tests give different results for some combinations of parameters (with Kolmogorov-Smirnov and Mann-Withney detecting more significant differences). There could be several reasons for these differences, related to the number of samples considered, or to the particular shape of the density functions.

4. Discussion

Aiming at defining a rigourous strategy for building correct hybrid models of biochemical systems, we studied the dynamics of a simple and paradigmatic gene regulatory network. In this context, we set up a systematic investigation for evaluating the relative importance of different sources of noise (i.e. the reactions composing the model) in determining the overall behaviour.

The results of this analysis allowed us to identify which of the considered reactions can be specified through a deterministic formalisms, while preserving the accuracy of the corresponding fully stochastic model in reproducing the described dynamics.

First, we identified two main parameters that govern the system's behaviour, grounding on the fundamental information about the basic "building blocks" of the system (i.e. protein, mRNA, and gene dynamics) and checking the validity of different plausible hypotheses. We also identified biologically reasonable ranges for the values of these parameters. Subsequently, in order to carry out numerical analysis, we considered several possible abstractions of the exact discrete stochastic model, relating their validity to the possible values of parameters identified. The use of different abstract models also allowed us to identify the key species and interactions that determine the overall behaviour of the system. Finally, we analysed the model quantitatively, also comparing the viability of different abstractions.

Different dynamic behaviours: role of parameters. A collection of distinct dynamic behaviours emerged, and we were able to link (relative) values of model parameters with observed behaviours, covering (to the best of our knowledge) the full landscape of behaviours reported in the literature. This analysis justified our assumption that feedback repression strength and relative degradation speed are good descriptors of the network behaviour. We provided a quantitative analysis of the relative contribution of various parameters to the global dynamics, correlating precise parameter combinations with a corresponding "behavioural phenotype". For instance, low values of both α and β yield the "continuous phenotype", while high values of α and β correspond to the "multi-modal phenotype". It is interesting to notice that gene expression dynamics switching from continuous to oscillating behaviours are common in biological systems and can be used as mechanisms to trigger alternative responses [53].

Different dynamic behaviours: emergence and relation to model abstraction. Analysing different abstract models allowed us to better clarify which interactions in the model are crucial for distinct behavioural phenotypes to emerge. All the used abstractions are compared in terms of computational cost and of the potential of the underlying model to faithfully reproduce exact behaviours and on the deriving cost/benefit ratio. We showed that while the fully continuous approximation failed to capture relevant behaviours (as expected), a hybrid approximation scheme, where only the protein is treated continuously, worked very well also for moderately low protein numbers. However, a discrete treatment of the protein was shown to be necessary for very small numbers in order to avoid significant loss in accuracy. We finally investigated the use of QSSAs on gene dynamics and its interaction with hybrid abstraction, recovering known patterns in the accuracy of this approximation also for the hybrid case [46].

Noise sources. The hybrid and QSSA abstractions allowed us to selectively turn on and off the different (internal) noise sources of the system, i.e. the different reactions, by treating them as continuous or removing them via QSSA. In this way, we were able to identify the reactions which are essential for producing the observed noise patterns. Our experimentation showed that the key reactions are those involved in modifying the value of molecular species present in low

numbers and that are slow (hence not amenable to QSSA). To analyse to what extent noise patterns are correctly reproduced, a model describing a negative feedback loop can be safely simplified abstracting away the non-key reactions. Our methodology of screening and targeting key reactions can be generalised to more complex models. Hence, it is the intrinsic discreteness of the system (in terms of low copy numbers of some molecular species) that drives the noise dynamics.

In summary, we systematically analysed a fully detailed discrete stochastic model of a negative feedback loop, collecting interesting insights on its dynamics; in this respect, our results can be considered an extension of the ones presented in [27]. Furthermore, we precisely analysed the impact of abstractions on the viability of model analysis. By systematically comparing the exact model with its abstractions, we highlighted the key interactions that determine the different behavioural phenotypes exhibited by the model.

We believe that our methodology, i.e. a preliminary computational screening for identifying the key structural parameters and the regions in which different abstractions are applicable, followed by an extensive computational analysis, allows us to gain a deep understanding of the model behaviour and, at the same time, it reduces the overall computational effort compared to a more blind and extensive exploration of the parameter space. In this sense, our approach can be extended to a thorough analysis of more complex genetic networks. On a more general level, the methodology proposed here for the study of feedback loops is based on the following assumptions and considerations:

- fully discrete stochastic models are the most reliable and can be used as "benchmarks" for the evaluation of alternative models;
- a complete and detailed landscape of parameter values must be determined and tested for each proposed model;
- it is important to identify the "hot points" in control structures, i.e. the key points for the regulation dynamics. Furthermore, for each control point, it is crucial to investigate to which extent the different approximations capture the emerging behaviours, possibly using deductive arguments;
- the best trade-off between computational costs and biological faithfulness is to be found in hybrid models preserving such "hot points";
- large genetic networks may be studied by exploiting an extensive analysis of the simpler modules composing them. In particular, the noise properties of simple feedback mechanisms of genetic networks and a precise characterisation of the validity of hybrid and QSSAs can be used to construct accurate abstract models in a modular way.

An interesting extension of this work would be the definition of a (semi)-automatic procedure which, following the analysis steps presented here, will aid in building hybrid models of biological networks. Such methodology should proceed with an attempt to (semi-)automate parameters and (especially) "hot

points" search exploiting modularity and it would have as ultimate goal the identification of the "right" level of abstraction of a network, with respect to the fully discrete stochastic model. This approach should produce a classification of relevance/weight of each point in the overall evaluation of the chosen level of abstraction: a classification certainly useful from a biological point of view, but very difficult to realise without a systematic *modus operandi*. In this setting, it would be also interesting to investigate to what extent the proposed methodology can be extended to preserve specific classes of behaviours. This would in principle allow to select for abstractions exhibiting a dynamics that closely relates to the expected dynamics of the real system under consideration.

Acknowledgements

Davide Chiarugi acknowledges the support of the Flagship "InterOmics" project (PB.P05), supported by the Italian MIUR and CNR organizations. Maria Luisa Guerriero was partially supported by Science Foundation Ireland research grant SFI 13/IF/B2792; her current affiliation is AstraZeneca, Cambridge, UK. Alessandro Romanel's contribution was partially supported by the ABSTRACTCELL ANR-Chair of Excellence.

References

[1] A. Raj, A. van Oudenaarden, Single-molecule approaches to stochastic gene expression, Annu. Rev. Biophys. 38 (2009) 255–270.

[2] A. Novic, M. Weiner, Enzyme induction as an all-or-none phenomenon, Proc. Natl. Acad. Sci. USA 43 (7) (1957) 553–566.

[3] I. L. Ross, C. M. Browne, D. A. Hume, Transcription of individual genes in eukaryotic cells occurs randomly and infrequently, Immunol. Cell Biol. 72 (2) (1994) 177–185.

[4] M. B. Elowitz., A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell, Science 297 (5584) (2002) 1183–1186.

[5] K. F. Murphy, R. M. Adams, X. Wang, G. Balázsi, J. J. Collins, Tuning and controlling gene expression noise in synthetic gene networks, Nucleic Acids Research 38 (8) (2010) 2712–2726.

[6] N. V. Mantzaris, From single-cell genetic architecture to cell population dynamics: Quantitatively decomposing the effects of different population heterogeneity sources for a genetic network with positive feedback architecture, Biophysical Journal 92 (12) (2007) 4271–4288.

[7] M. Stamatakis, K. Zygourakis, A mathematical and computational approach for integrating the major sources of cell population heterogeneity, Journal of Theoretical Biology 266 (1) (2010) 41–61.

[8] M. Stamatakis, K. Zygourakis, Deterministic and stochastic population level simulations of an artificial lac operon genetic network, BMC Bioinformatics 12 (2011) 301.

[9] G. Balázsi, A. van Oudenaarden, J. J. Collins, Cellular decision making and biological noise: From microbes to mammals, Cell 144 (6) (2011) 910–925.

[10] A. Raj, S. A. Rifkin, E. Andersen, A. van Oudenaarden, Variability in gene expression underlies incomplete penetrance, Nature 463 (7283) (2010) 913–918.

[11] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, Garland Science, 2002.

[12] G. Lillacci, M. Khammash, Parameter estimation and model selection in computational biology, PLoS Computational Biology 6 (3) (2010) e1000696.

[13] A. Raj, A. van Oudenaarden, Nature, nurture, or chance: Stochastic gene expression and its consequences, Cell 135 (2) (2008) 216–226.

[14] D. A. Hume, Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression, Blood 96 (7) (2000) 2323–2328.

[15] J. R. Chabot, J. M. Pedraza, P. Luitel, A. van Oudenaarden, Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock, Nature 450 (7173) (2007) 1249–1252.

[16] A. Becksei, B. Séraphin, L. Serrano, Positive feedback in eukaryotic gene networks: cell differentiation by graded binary response conversion, EMBO J. 20 (10) (2001) 2528–2535.

[17] H. Resat, L. Petzold, M. F. Pettigrew, Kinetic modeling of biological systems, Methods in Molecular Biology 541 (10) (2009) 311–335.

[18] D. McQuarrie, Stochastic approach to chemical kinetics, J. Appl. Prob. 4 (3) (1967) 413–478.

[19] A. F. Bartholomay, Stochastic models for chemical reactions, Bull. Math. Biophys. 20 (3) (1958) 175–190.

[20] T. Dauxois, F. Di Patti, D. Fanelli, A. J. McKane, Enhanced stochastic oscillations in autocatalytic reactions, Phys Rev E Stat Nonlin Soft Matter Phys 79 (2009) 036112.

[21] D. Gillespie, Exact stochastic simulation of coupled chemical reactions, Journal of Physical Chemistry 81 (25) (1977) 2340–2361.

[22] J. Pahle, Biochemical simulations: stochastic, approximate stochastic and hybrid approaches, Brief Bioinform. 10 (1) (2009) 53–64.

[23] L. Bortolussi, A. Policriti, Hybrid semantics of stochastic programs with dynamic reconfiguration, in: Proceedings of CompMod'09, 2009, pp. 63–76.

[24] L. Bortolussi, A. Policriti, (Hybrid) automata and (stochastic) programs: The hybrid automata lattice of a stochastic program, Journal of Logic and Computation 23 (4) (2013) 761–798.

[25] H. Salis, Y. Kaznessis, Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions, Journal of Chemical Physics 122 (5) (2005) 54103.

[26] H. Salis, V. Sotiropoulos, Y. N. Kaznessis, Multiscale Hy3S: Hybrid stochastic simulation for supercomputers, BMC Bioinformatics 7 (2006) 93.

[27] T. T. Marquez Lago, J. Stelling, Counter-intuitive stochastic behavior of simple gene circuits with negative feedback, Biophysical Journal 98 (9) (2010) 1742–1750.

[28] D. J. Stekel, D. J. Jenkins, Strong negative self regulation of prokaryotic transcription factors increases the intrinsic noise of protein expression, BMC Systems Biology 2 (2008) 6.

[29] G.-Q. Zhou, W.-Z. Zhong, Diffusion-controlled reactions of enzymes, Eur. J. Biochem. 128 (2–3) (1982) 383–387.

[30] U. Vogel, K. F. Jensen, The RNA chain elongation rate in Escherichia coli depends on the growth rate, J Bacteriol. 176 (10) (1994) 2807–2813.

[31] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, U. Kummer, COPASI: a COmplex PAthway SImulator, Bioinformatics 22 (24) (2006) 3067–3074.

[32] O. L. V. Costa, F. Dufour, Stability and Ergodicity of Piecewise Deterministic Markov Processes, SIAM J. Control Optim. 47 (2008) 1053–1077.

[33] M. Gibson, J. Bruck, Efficient exact simulation of chemical systems with many species and many channels, J. of Phys. Chem. 104 (9) (2000) 1876–1889.

[34] I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, Real-time kinetics of gene activity in individual bacteria, Cell 123 (6) (2005) 1025–1036.

[35] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA synthesis in mammalian cells, PLoS Biol. 4 (10) (2006) e309.

[36] J. Peccoud, B. Ycart, Markovian modelling of gene-product synthesis, Theor Popul Biol. 48 (2) (1995) 222–234.

[37] D. T. Gillespie, L. Petzold, Numerical Simulation for Biochemical Kinetics, MIT Press, 2006, Ch. System Modelling in Cellular Biology, pp. 1–29.

[38] M. Rathinam, L. R. Petzold, Y. Cao, D. T. Gillespie, Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method, Journal of Chemical Physics 119 (24) (2003) 12784–12794.

[39] R. Ramaswamy, I. Sbalzarini, A partial-propensity variant of the composition-rejection stochastic simulation algorithm for chemical reaction networks, Journal of Chemical Physics 132 (4) (2010) 044102.

[40] C. V. Rao, A. P. Arkin, Stochastic Chemical Kinetics and the Quasi-Steady State Assumption: Application to the Gillespie Algorithm, Journal of Chemical Physics 118 (11) (2003) 4999–5010.

[41] D. T. Gillespie, The chemical Langevin equation, Journal of Chemical Physics 113 (1) (2000) 297–306.

[42] M. H. A. Davis, Markov Models and Optimization, Chapman & Hall, 1993.

[43] K. S. Trivedi, V. G. Kulkarni, FSPNs: Fluid stochastic Petri nets, in: Application and Theory of Petri Nets, 1993, pp. 24–31.

[44] A. Crudu, A. Debussche, O. Radulescu, Hybrid stochastic simplifications for multiscale gene networks, BMC Systems Biology 3 (2009) 89.

[45] R. L. Burden, J. D. Faires, Numerical analysis, Thomson Brooks/Cole, 2005.

[46] Y. Cao, D. T. Gillespie, L. R. Petzold, The slow-scale stochastic simulation algorithm, Journal of Chemical Physics 122 (1) (2005) 014116.

[47] E. A. Mastny, E. L. Haseltine, J. B. Rawlings, Two classes of quasi-steady-state model reductions for stochastic kinetics, Journal of Chemical Physics 127 (9) (2007) 094106.

[48] W. E, D. Liu, E. Vanden-Eijnden, Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates, Journal of Chemical Physics 123 (2005) 194107.

[49] W. E, D. Liu, E. Vanden-Eijnden, Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales, Journal of Computational Physics 221 (1) (2007) 158–180.

[50] Y. Cao, L. Petzold, Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems, J. Comput. Phys. 212 (1) (2006) 6–24.

[51] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The Annals of Mathematical Statistics 18 (1) (1947) 50–60.

[52] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 3rd Edition, Chapman & Hall, 2004.

[53] J. J. Walker, J. R. Terry, K. Tsaneva-Atanasova, S. P. Armstrong, C. A. McArdle, S. L. Lightman, Encoding and decoding mechanisms of pulsatile hormone secretion, Journal of Neuroendocrinology 22 (12) (2010) 1226–1238.

[54] L. J. S. Allen, An Introduction to Stochastic Processes With Applications in Biology, Prentice Hall, 2003.

[55] H. Bremer, P. P. Dennis, Modulation of chemical composition and other parameters of the cell by growth rate, Neidhardt et al., 1996, Ch. Escherichia coli and Salmonella, pp. 1553–1569.

Histogram distance for protein					
α-β	Self	QSSA	Hybrid	QSSA+Hybrid	Self bound
L-L	0.268	0.262	0.296	0.354	1.0488
L-M	0.348	0.316	0.324	0.34	1.1248
L-H	0.185	0.402	0.234	0.160	1.0747
M-L	0.26	0.316	0.364	0.358	1.1728
M-M	0.34	0.282	0.266	0.284	1.1008
M-H	0.124	0.132	0.646	0.996	1.0088
H-L	0.176	0.32	1.194	**1.45**	0.9688
H-M	0.240	0.136	0.234	0.13	1.0568
H-M	0.108	0.652	0.746	**1.108**	0.9808
Histogram distance for mRNA					
L-L	0.252	0.166	0.16	0.156	0.9090
L-M	0.16	0.19	0.206	0.228	0.8170
L-H	0.185	0.347	0.211	0.223	1.1118
M-L	0.152	0.038	0.038	0.038	0.4322
M-M	0.052	0.006	0.05	0.046	0.3321
M-H	0.028	0.04	0.006	0.026	0.2838
H-L	0	0	0	0	0.1981
H-M	0	0.038	0	0.024	0.1981
H-H	0.004	0.078	0	0.074	0.2328

Table S1: Histogram distance for protein P and mRNA M between the explicit stochastic model and different abstract models. The histograms are based on 1000 trajectories sampled at time $t = 10000$. In the column `Self`, we report the self distance [50] between two histograms of the explicit stochastic model. In the column `Self bound`, we add to such self distance three standard deviations, estimated from an upper bound as in [50], as a (very conservative) significance threshold for comparing histogram distances of abstract model. Distances above this threshold are very unlikely to be due to random effects related to the limited amount of samples compared.

Kolmogorov-Smirnov test for protein				
α-β	Self	QSSA	Hybrid	QSSA+Hybrid
L-L	0.9022	0.9689	**9.0800e-05**	**9.0800e-05**
L-M	0.2574	0.0257	0.2193	0.5003
L-H	0.8971	0.0558	0.1847	0.6555
M-L	0.8186	0.5004	0.5066	**3.8958e-03**
M-M	0.6654	0.5004	0.5726	0.6475
M-H	0.6654	0.2193	**<1.0e-16**	**<1.0e-16**
H-L	0.8632	**1.1148e-06**	**<1.0e-16**	**<1.0e-16**
H-M	0.0815	0.0692	0.2877	0.2193
H-M	0.0693	**<1.0e-16**	**<1.0e-16**	**<1.0e-16**
Kolmogorov-Smirnov test for mRNA				
L-L	0.9959	0.9999	0.1082	0.8592
L-M	0.9987	0.7944	0.8279	0.3135
L-H	0.9748	0.1790	0.8452	0.95334
M-L	0.1725	0.9987	0.9688	1
M-M	0.9895	1	0.9969	0.9540
M-H	1	0.9987	1	1
H-L	1	1	1	1
H-M	0.9780	0.9936	0.4324	0.9999
H-H	0.9959	**<1.0e-16**	1	**<1.0e-16**

Table S2: Results of the Kolmogorov-Smirnov test to compare the empirical distributions of mRNA and protein of the explicit stochastic model at time $t = 10000$ with different abstract models. We report p-values of the test at 95% confidence level. Values in bold represent statistically significant differences between the two histograms. In the column `Self`, we report the result of the Kolmogorov-Smirnov test of two samples of the explicit stochastic model.

Mann-Withney test for protein				
α-β	Self	QSSA	Hybrid	QSSA+Hybrid
L-L	0.7381	0.7217	**2.1225e-05**	**9.9179e-07**
L-M	0.4150	0.0617	0.1612	0.6930
L-H	0.7424	0.1455	0.4591	0.9729
M-L	0.7850523	0.6461	**1.6075e-03**	**5.1110e-04**
M-M	0.4955	70.6832	0.4141	0.7696
M-H	0.7717	0.0985	0.6343	0.9958
H-L	0.8278	**2.4536e-07**	**8.7372e-04**	**3.5874e-03**
H-M	0.1127	0.0423	0.2431	0.9903
H-M	0.1499	**3.6361e-38**	0.7018	**1.8533e-42**
Mann-Withney test for mRNA				
L-L	0.7559	0.7695	0.1305	0.9166
L-M	0.8414	0.5699	0.5496	0.2749
L-H	1	0.0948	0.3456	0.8820
M-L	0.0464	0.3879	0.631	0.6891
M-M	0.4192	0.6860	0.4669	0.2444
M-H	0.8835319	0.3066	0.9944	0.5885
H-L	0.5588	0.8894	0.6706	0.2329
H-M	0.2037	0.8964	0.0139	0.6767
H-H	0.4212	**3.1888e-61**	0.7126	**1.9916e-68**

Table S3: Results of the Mann-Withney test to compare the empirical distributions of mRNA and protein of the explicit stochastic model at time $t = 10000$ with different abstract models. We report p-values of the test at 95% confidence level. Values in bold represent statistically significant differences between the two histograms. In the column `Self`, we report the result of the Mann-Withney test of two samples of the explicit stochastic model.

Appendix A. Dynamics of the feedback circuit

In this section, we carry out a high level discussion on the stochastic dynamics of M, P, gene repression, and on the interactions between the dynamical regimes involved in the model. Our goal is to illustrate the reasoning which led us to identify the key parameters ruling the behaviour of this simple feedback loop.

We start by discussing the simpler model with no repression, i.e. a model containing just transcription, translation, and M and P degradation, where the gene is always on [13].

Histogram distance, MW and KS test for protein with variable unbinding rate					
α-β	self distance	stoch-QSSA distance	self distance bound	MW p-value	KS p-value
L-M	0.288	0.306	1.1048	0.8838	0.9355
M-M	0.212	0.690	1.0289	**1.9224e-10**	**<1.0e-16**
H-M	0.036	0.626	0.8528	**1.35e-110**	**<1.0e-16**
Histogram distance, MW and KS test for mRNA with variable unbinding rate					
L-M	0.252	0.136	0.8888	0.84341	0.9356
M-M	0.072	0.304	0.8888	0.1988	**1.85e-10**
H-M	0.004	0.026	0.4469	**1.76e-40**	**1.55e-12**

Table S4: Mann-Withney (MW) test, Kolmogorov-Smirnov (KS) test, and histogram distances for comparing the empirical distributions of mRNA and protein of the explicit stochastic model at time $t = 10000$ (with variable unbinding rate) with the stochastic model with QSSA on binding, obtained from 1000 sampled trajectories. We report p-values of the test at 95% confidence level. Values in bold represent statistically significant differences between the two histograms. Histogram self distance is also reported, together with a (conservative) significance threshold computed as above.

Histogram distance, MW and KS test for mRNA for QSSA on binding and protein					
α-β	self distance	stoch-QSSA distance	self distance bound	MW p-value	KS p-value
L-H	0.259	0.265	1.112	0.547	0.5468
M-H	0.028	0.036	0.283	0.679	0.6797
H-H	0.004	0.002	0.233	**1.476e-103**	**<1.0e-16**

Table S5: Mann-Withney (MW) test, Kolmogorov-Smirnov (KS) test, and histogram distances for comparing the empirical distributions of mRNA of the explicit stochastic model at time $t = 10000$ with the stochastic model with QSSA on binding and protein, obtained from 1000 sampled trajectories. We report p-values of the test at 95% confidence level. Values in bold represent statistically significant differences between the two histograms. Histogram self distance is also reported, together with a (conservative) significance threshold computed as above.

Simulation CPU time					
α-β	full stoch	stoch QSSA binding	hybrid	hybrid+QSSA binding	stoch QSSA binding+protein
L-L	0.046	0.037	0.211	0.076	n.a.
M-L	0.037	0.028	0.207	0.072	n.a.
H-L	0.036	0.028	0.200	0.073	n.a.
L-M	0.864	0.865	0.215	0.075	n.a.
M-M	0.07	0.059	0.213	0.073	n.a.
H-M	0.04	0.039	0.157	0.071	n.a.
L-H	77.798	80.34	0.280	0.083	0.067
M-H	4.130	4.174	0.216	0.077	0.136
H-H	2.129	3.108	0.171	0.075	0.204

Table S6: Comparison of simulation times (in seconds) for single runs, for $t = 10000$ seconds, for different abstraction methods.

In this case, X_M is independent of X_P,[6] and it is a birth-death process, with constant birth rate $prod_M$ and linear death rate $deg_M \cdot X_M$. Processes of this kind have a dynamics that is well understood from a theoretical point of view [54]: X_M converges to a Poisson stationary distribution with mean (and variance) equal to $prod_M/deg_M = M_{steady}$. Due to linearity of production and degradation rates, we can explicitly compute the transient dynamics of the mean $\mathbf{E}_t[X_M]$ and the variance $\mathbf{VAR}_t[X_M]$ of the process (assuming $X_M(0) = 0$ for simplicity):

$$\mathbf{E}_t[X_M] = \mathbf{VAR}_t[X_M] = \frac{prod_M}{deg_M}(1 - e^{-deg_M \cdot t}) \ .$$

From these equations, we can observe how the speed of convergence to the stationary distribution depends only on the degradation rate deg_M, while $prod_M$ has the effect of changing the average equilibrium value of M; in other words, $prod_M$ changes the scale of the process.[7]

The dynamics of X_P, conditional on $X_M = m$, is also a birth-death process

[6] This means that $\mathbb{P}(X_M = m|X_P = p) = \mathbb{P}(X_M = m)$, hence the joint probability distribution $\mathbb{P}(X_M = m; X_P = p)$ can be factorised as $\mathbb{P}(X_P = p|X_M = m)\mathbb{P}(X_M = m)$: we can study the processes X_M and X_P conditional on X_M separately, and then discuss how the dynamics of X_M propagates to X_P.

[7] The coefficient of variation of X_M, a standard measure of the noise level of a stochastic process (defined as its standard deviation divided by its mean), equals $1/\sqrt{prod_M/deg_M}$, hence noise is inversely proportional to the steady state level of X_M.

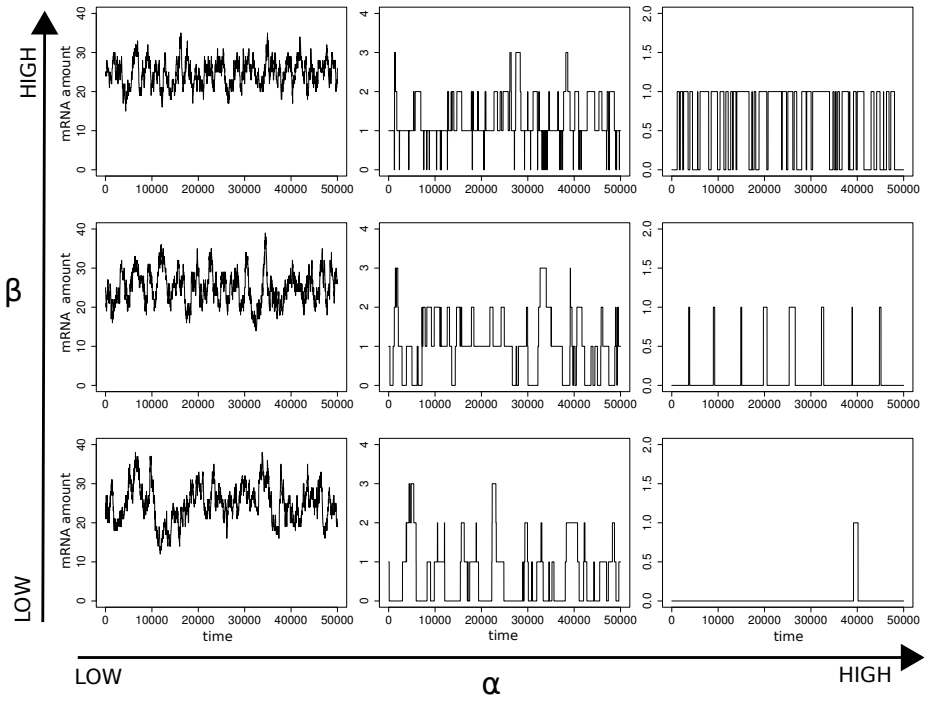


Figure S1: Sampled trajectories of M from the explicit stochastic model, using the Gibson-Bruck stochastic simulation algorithm, for different combinations of α and β.

with birth rate $prod_P \cdot m$ and death rate $deg_P \cdot X_P$, so that the average steady state of X_P is $prod_P \cdot m/deg_P$, while the speed of convergence to it depends only on deg_P.

The combined dynamics of X_M and X_P can be understood by looking at the ratio between their degradation rates (i.e. by the relative speed of the processes). If the dynamics of P is slow with respect to M, $deg_P \ll deg_M$, then X_M will rapidly reach equilibrium and fluctuate on a faster time scale than X_P, so that X_P tends to be insensitive to the fluctuations of X_M (provided that the ratio $prod_P/deg_P$ is not too large). On the other hand, if $deg_P \gg deg_M$ then X_P will react quickly to each change in the value of X_M, reaching a new steady state value every time X_M changes. If the ratio $prod_P/deg_P$ is sufficiently large (so that the fluctuations of X_P due to its birth-death dynamics are small with respect to the variation of the steady state value), then the trajectories of X_P will "look like" those of X_M, rescaled by a factor of $prod_P/deg_P$, and X_M and X_P will have a high correlation [13, 27]. This has the effect of increasing the noise on P. Such noise can be large even for slow P dynamics if the production rate is very large: in this case, a change in the value of M induces a very large change in P steady state value, increasing the magnitude of P's fluctuations [13]. However, very large translation rates are not to be expected in a biological setting, therefore we fixed the ratio $prod_P/deg_P$ and focussed on the ratio deg_P/deg_M

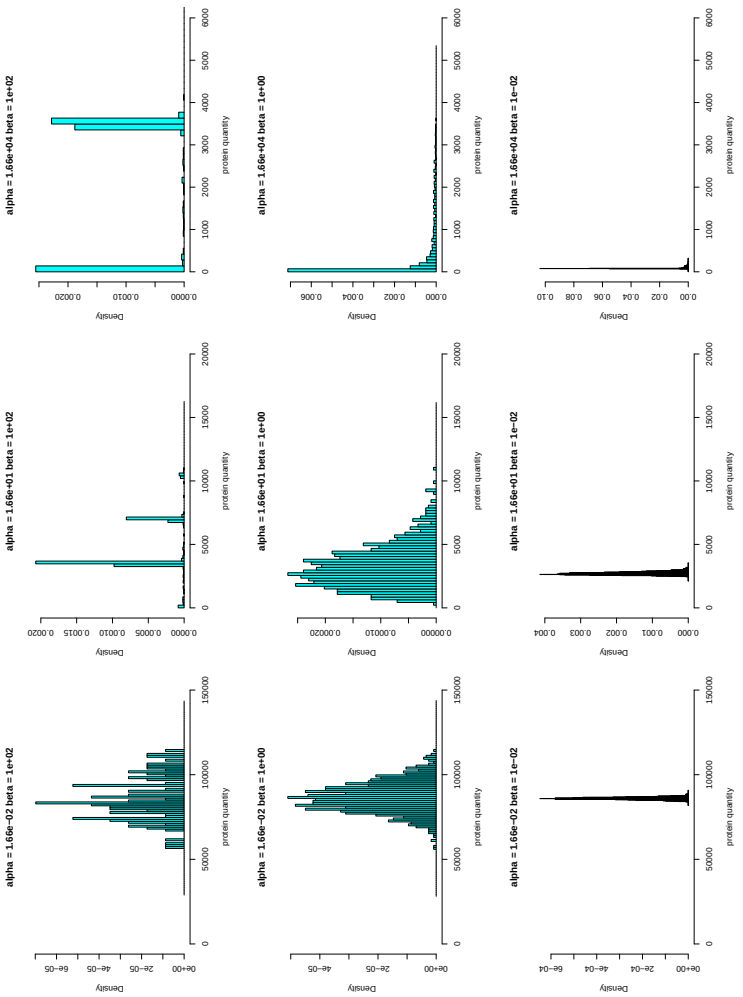


Figure S2: Empirical histogram distributions at $t = 10000$ of the protein P for the explicit stochastic model, obtained from 1000 sampled trajectories, using the Gibson-Bruck stochastic simulation algorithm, for different combinations of α and β.

as a descriptor of the combined effects of transcription and translation on the system dynamics.

The crucial point of the dynamics of X_M and X_P, in our network, is the influence exerted by the negative feedback loop. Repression is obtained by the binding of P to a promoter region of the gene. This event blocks the transcriptional activity of the gene, usually by masking the binding site of the RNA polymerase. Clearly, the intensity of repression depends both on binding and unbinding rates: strong repression can be obtained either by a high binding rate or by a low unbinding rate. In the former case, the probability that a

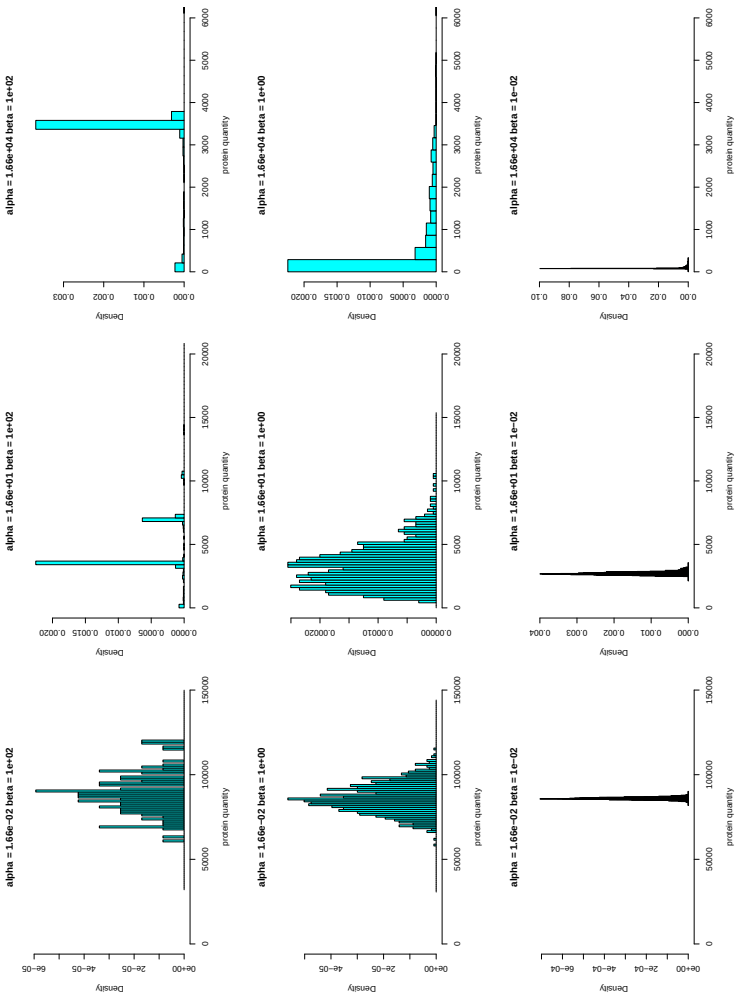


Figure S3: Empirical histogram distributions at $t = 10000$ of the protein P for the stochastic model with QSSA on binding, obtained from 1000 sampled trajectories, using the Gibson-Bruck stochastic simulation algorithm, for different combinations of α and β.

molecule of repressor binds to the gene is (much) larger than the probability of starting a transcription, hence this latter event will be less frequent. In the latter case, the time for which P remains bound to the gene increases, *de facto* reducing the time span in which the gene is available for transcription. In both cases, the net effect is a reduction in transcription rate and, hence, in the amount of M present in the system.

From a dynamical point of view, the repression mechanism, conditional to $X_P = p$, is a simple telegraph process [54] with constant rates $bind_P \cdot p$ and $unbind_P$. The probability of the system being in the repressed state at time t

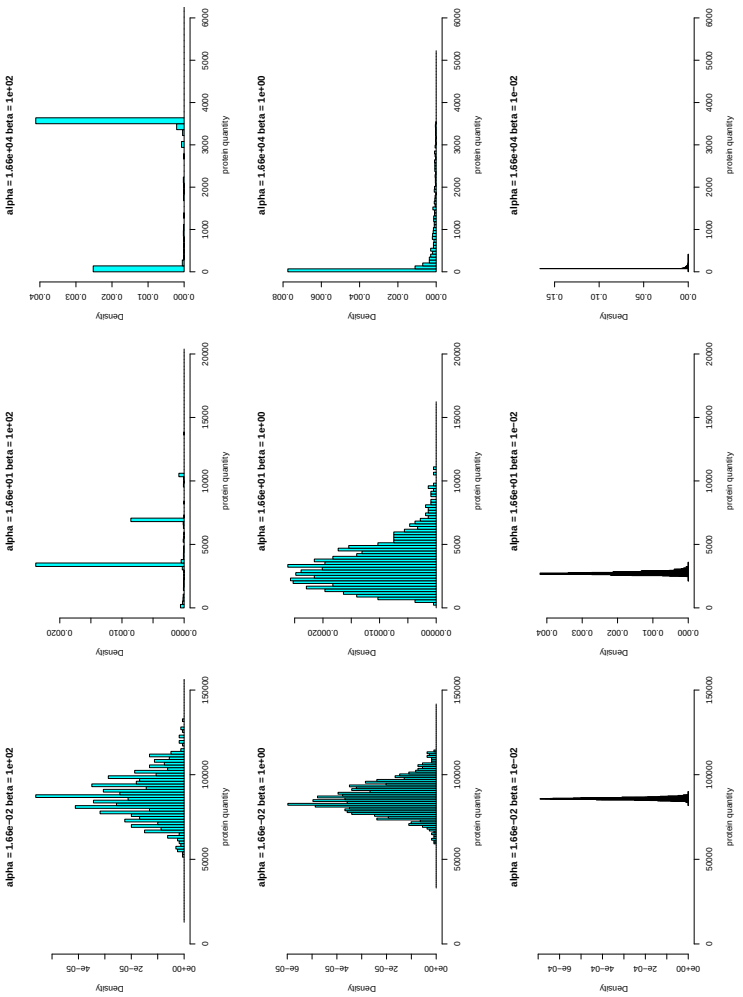


Figure S4: Empirical histogram distributions at $t = 10000$ of the protein P for the explicit hybrid model, obtained from 1000 sampled trajectories, using the COPASI LSODA hybrid algorithm, for different combinations of α and β.

is $bind_P \cdot p/(bind_P \cdot p + unbind_P)(1 - e^{-(bind_P \cdot p + unbind_P)t})$. In particular, the speed of convergence to the equilibrium distribution is $bind_P \cdot p + unbind_P$.

A standard way to evaluate the repression intensity is to consider the ratio $bind_P/unbind_P$, as done in [27]. In order to study the effect of the feedback loop, we vary such ratio fixing either the binding or the unbinding rate, and changing the other one. The choice of which one to fix is not irrelevant, as these two rates act on different aspects of the repression mechanism. If, as in [27], we fix the binding rate, we can achieve strong repression by increasing the binding strength between the repressor and the gene, so that P will remain bound for a long time.

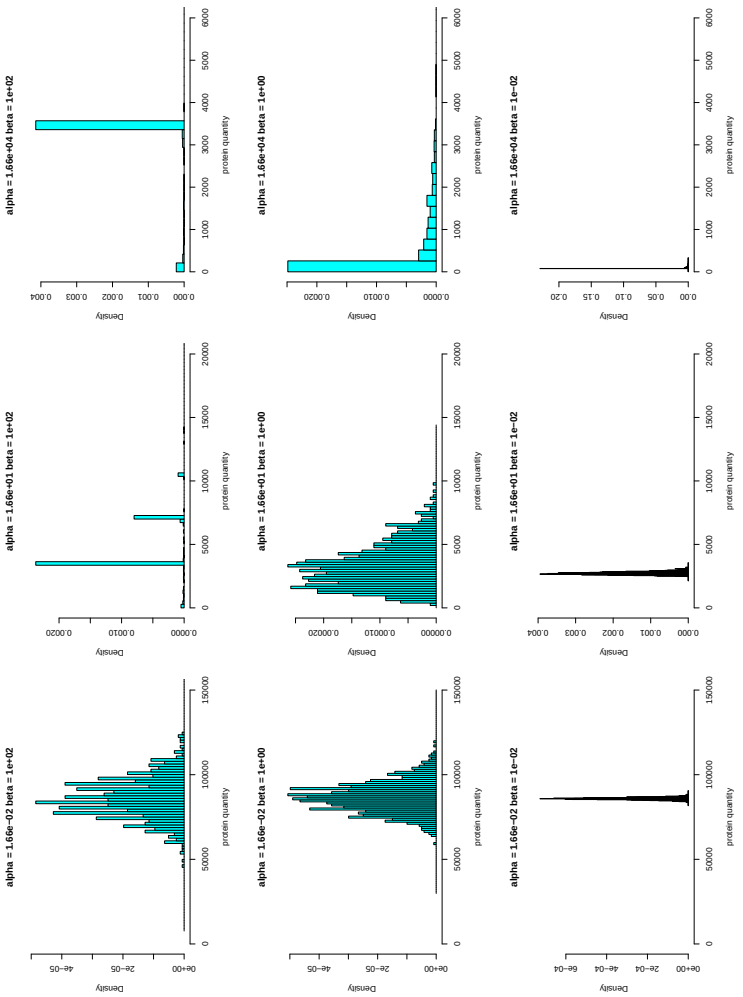


Figure S5: Empirical histogram distributions at $t = 10000$ of the protein P for the hybrid model with QSSA on binding, obtained from 1000 sampled trajectories, using the COPASI LSODA hybrid algorithm, for different combinations of α and β.

This choice reduces the dynamics repression speed (in $bind_P \cdot p + unbind_P$ both p and $unbind_P$ are reduced; for $p = 0$, the speed equals $unbind_P$, which can be very small). In particular, if the dynamics of binding and unbinding are slower than that of P and of M (i.e. $unbind_P < deg_P$ and $unbind_P < deg_M$), then P will be partly or fully degraded when the gene is shut off. In this case, we may observe a peak-like or burst-like behaviour. The magnitude of bursts will depend on the speed of P dynamics (and on the translation rate): the quicker P reacts to variations in M (and the larger the translation rate), the larger the bursts. On the other hand, if we fix the unbinding rate and we

increase the feedback strength by increasing the binding rate, then the speed at which repression reaches equilibrium is bounded below by $unbind_P$, which is a constant. Therefore, if $unbind_P$ is greater than deg_M, as commonly happens in real biological situations (including our parameter set), then the dynamics of repression will quickly reach equilibrium (i.e. the probability of the gene being repressed will be equal to the steady state probability [46]) between two consecutive events changing the value of X_M. Also in this case the transcription rate is reduced, but we should not see the long repression windows caused by a low unbinding rate. Therefore, the set of behaviours that we can observe as a function of feedback strength can be different if we fix the binding or the unbinding rate, especially at the strong repression regime. Here, we mostly stick to a fixed unbinding rate, investigating the alternative option in Section 3.3 of the main text, where we discuss a quasi-steady state approximation for binding and unbinding.

Appendix B. Estimating the upper bound of enzymatic reactions

The most efficient enzymes (diffusion-controlled enzymes) catalyse reactions with a rate in the range 10^8 – 10^{10} $M^{-1}s^{-1}$. These values are at the upper bound of the observed rates and have been predicted by theoretical studies on bimolecular reaction in solution [29]. Such enzymes are considered to be perfect, since their rate-limiting step is not due to any chemical event but to the diffusional association rate between the enzyme and the substrate [29]. These arguments can be useful to estimate a theoretical upper bound for the translation rate which is an enzymatically catalysed process. In our discrete framework the rates indicated above should be converted from $M^{-1}s^{-1}$ to number of molecules per second. Considering a cytoplasmic volume of $\approx 10^{-15}$l and the Avogadro's number $\approx 10^{24}$mol$^-$1, the maximum rate will be ≈ 10 molecules/s. Considering that, in E. coli, each mRNA is translated by ≈ 90 ribosomes (maximum value) [55], the highest possible rate value will be $\approx 10^3$.