
Article:
Lopez-Lopera, A.F., Durrande, N. and Alvarez, M.A. (2021) Physically-inspired Gaussian process models for post-transcriptional regulation in Drosophila. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18 (2). pp. 656-666. ISSN 1545-5963

https://doi.org/10.1109/tcbb.2019.2918774

PHYSICALLY-INSPIRED GAUSSIAN PROCESS MODELS FOR POST-TRANSCRIPTIONAL REGULATION IN DROSOPHILA

Andrés F. López-Lopera*
Mines Saint-Étienne
42000 Saint-Étienne, France
andres-felipe.lopez@emse.fr

Nicolas Durrande
PROWLER.io
Cambridge, CB2 1LA, UK
nicolas@prowler.io

Mauricio A. Álvarez
University of Sheffield
Sheffield, S1 4DP, UK
mauricio.alvarez@sheffield.ac.uk

ABSTRACT

The regulatory process of Drosophila is thoroughly studied for understanding a great variety of biological principles. While pattern-forming gene networks are analysed in the transcription step, post-transcriptional events (e.g. translation, protein processing) play an important role in establishing protein expression patterns and levels. Since the post-transcriptional regulation of Drosophila depends on spatiotemporal interactions between mRNAs and gap proteins, proper physically-inspired stochastic models are required to study the link between both quantities. Previous research attempts have shown that using Gaussian processes (GPs) and differential equations lead to promising predictions when analysing regulatory networks. Here we aim at further investigating two types of physically-inspired GP models based on a reaction-diffusion equation where the main difference lies in where the prior is placed. While one of them has been studied previously using protein data only, the other is novel and yields a simple approach requiring only the differentiation of kernel functions. In contrast to other stochastic frameworks, discretising the spatial space is not required here. Both GP models are tested under different conditions depending on the availability of gap gene mRNA expression data. Finally, their performances are assessed on a high-resolution dataset describing the blastoderm stage of the early embryo of Drosophila melanogaster

1 Introduction

Regulatory process modelling in molecular mechanisms has taken great attention due to its significant role in systems biology (Alon, 2006; Barrera et al., 2016). The gene regulation consists mainly of two steps. First, DNA sequences are encoded in mRNA molecules according to the need of proteins' production in cells (transcription step). Then, the mRNA is used as a template for protein synthesis (translation step) (Alon, 2006; Forgacs and Newman, 2005). Both transcription and translation steps are crucial to understand a great variety of biological phenomena such as the molecular mechanisms of cell survival and protein production (Alon, 2006; Barrera et al., 2016; Dilão and Muraro, 2010). While pattern-forming gene networks can be analysed in the former step, post-transcriptional events (e.g. RNA splicing, translation, protein processing) play an important role in establishing protein expression patterns and levels (Alon, 2006; Becker et al., 2013; Dilão and Muraro, 2010). This paper is focused on the post-transcriptional regulation at the mRNA level, more precisely, on modelling the gap gene network dynamics of Drosophila.

The genus Drosophila is one of the most studied examples of regulatory processes (Becker et al., 2013; Rogers et al., 2014). The regulation in Drosophila is commonly analysed to understand how protein concentrations are expressed and their influence on biological systems (Rogers et al., 2014; Hsiao et al., 2016; Dilão and Muraro, 2010). It is also used to investigate the interaction between the environment and biological processes (Rogers et al., 2014). For example, in recent work, it has been discovered that there are specific molecular mechanisms that control the circadian rhythm in living species (research awarded with the 2017 Nobel prize in Physiology and Medicine) (Yagita, 2018). Therefore, because of the versatility of Drosophila, its regulatory network is commonly used as a standard model.

Due to the importance of post-transcriptional regulation in Drosophila, accurate approaches for capturing the existing link between mRNAs and proteins are required (Becker et al., 2013; Dilão and Muraro, 2010; Wilson et al., 2010). In the last decades, several frameworks have been proposed encouraging the modelling of regulatory processes as stochastic processes (Barenco et al., 2006; Becker et al., 2013; Lipniacki et al., 2006; Dalessi et al., 2012; Erban et al., 2007). More precisely, they have suggested the use of Gaussian processes (GPs) and differential equations (Lawrence

*Part of this work was completed during an internship of A. F. López-Lopera at PROWLER.io.

et al., 2007; Gao et al., 2008; Álvarez et al., 2013; Vásquez Jaramillo et al., 2014). For the case of post-transcriptional regulation, GP-based approaches assume that both mRNA and protein concentrations are Gaussian-distributed (Liu and Niranjan, 2012; Álvarez et al., 2013). Then, due to the linearity of the differential equation used to link those quantities, mechanistic parameters can be encoded as parameters of GP covariance functions (Lawrence et al., 2007; Álvarez et al., 2013). According to experiments on both synthetic and real-world data in (Lawrence et al., 2007; Álvarez et al., 2013; Vásquez Jaramillo et al., 2014; Gao et al., 2008), physically-inspired GPs have yielded competitive and promising results for modelling regulatory processes.

Different classes of physically-inspired approaches have been proposed to model the regulation of the early embryo of Drosophila melanogaster (Dalessi et al., 2012; Liu and Niranjan, 2012; Álvarez et al., 2013; Vásquez Jaramillo et al., 2014). In (Dalessi et al., 2012), a Green's function method is introduced to model the Bicoid gradient establishment in Drosophila's embryos assuming the mRNA to be deterministic. Their approach presents two main drawbacks. First, it provides only deterministic solutions which do not allow expressing uncertainties on mRNA distributions. Second, their model is tested only on synthetic data due to issues on experimental settings. In contrast to (Dalessi et al., 2012), a stochastic framework based on GPs is proposed in (Liu and Niranjan, 2012), allowing the expression of uncertainty on inferred Bicoid concentrations. Although their results are promising, the discretisation of the spatial space is required; losing quality of resolution in predictions. More recently, in (Álvarez et al., 2013; Vásquez Jaramillo et al., 2014), an alternative physically-inspired GP framework using Green's functions is proposed to model continuous mRNA concentrations. There, it is assumed that the mRNA acting in the regulatory network was unknown and had to be estimated. Their assumption stands due to the common lack of (commonly expensive) mRNA data and the availability of protein concentration data. They place a GP prior over the mRNA and built up the resulting GP over the protein. Their framework requires explicitly solving the associated differential equation, followed by solving multiple integrals involving kernel functions which is not always feasible (Álvarez et al., 2013; Guarnizo and Álvarez, 2018). Assuming that closed-form solutions are available, then the resulting GP over the protein can be established, and the mRNA concentration can be inferred conditionally to the protein concentration data. The model proposed in (Álvarez et al., 2013; Vásquez Jaramillo et al., 2014) still presents some limitations. First, due to the lack of mRNA data, their approach could not be thoroughly tested for the inference of mRNA patterns. Second, in order to obtain closed-form expressions, their work is limited to a class of kernels.

An alternative approach is to assume the GP prior over the protein rather than over the mRNA, building up the resulting GP framework. This leads to a GP model where the solution of the differential equation is not required, but the differentiation of kernel functions is. For further discussions, we refer to GP-mRNA and GP-Protein to the physically-inspired GP with prior over the mRNA or protein concentrations, respectively. Therefore, our main contributions are threefold. First, we introduce the GP-Protein model as a novel alternative that does not require solving differential equations. Second, we further investigate both GP-mRNA and GP-Protein models, assessing their performances when data from both mRNA and protein concentrations are available. Three situations are analysed depending on the data availability: whether from the mRNA, from the protein or from both quantities. Third, we test both physically-inspired GP models on a high-resolution dataset describing the blastoderm stage of the early embryo of Drosophila (Becker et al., 2013).

This paper is organized as follows. In Section 2, we briefly describe the gap gene network associated with segmentation in early Drosophila organism development. In Section 3, we establish both physically-inspired GP models based on a diffusion equation. For the GP-mRNA model, we refer to the formulation from (Álvarez et al., 2013), but we write the main equations here for readability and further discussion. For the GP-Protein model, its formulation is completely detailed in Appendix A. We also assess both GP models on synthetic examples under different conditions depending on the data availability. In Section 4, we test the models using the Drosophila's database from (Becker et al., 2013). In Section 5, we summarise the conclusions, as well as potential future work.

2 Gap gene network of Drosophila

This work focuses on the role of post-transcriptional regulation within the context of gap gene networks associated with segmentation during the blastoderm stage of early Drosophila development. There, a set of molecules known as morphogens are responsible for embryo segmentation (Jaeger et al., 2012; Dalessi et al., 2012; Vásquez Jaramillo et al., 2014). Morphogens propagate spatially and establish maternal gradients along the anterior-posterior (A-P) axis of the embryo, describing a reaction-diffusion process (Jaeger et al., 2012; Meinhardt, 2015). Then, maternal gradients interact with specific trunk gap genes (e.g. *hunchback–hb*, *caudal–cd*, *Krüppel–kr*, *knirps–kni* and *giant–gt*), and this gap gene network of interactions constitutes the segmentation of the Drosophila (Becker et al., 2013; Surkova et al., 2013; Álvarez et al., 2013).

The reaction-diffusion process during early Drosophila embryo development is usually modelled through linear partial differential equations (PDEs) (Becker et al., 2013; Álvarez et al., 2013; Vásquez Jaramillo et al., 2014). For readability, and according to the structure of the dataset used in Section 4.2, we focus on the gap gene network dynamics,

$$\frac{\partial y(x,t)}{\partial t} = Su(x,t) - \lambda y(x,t) + D\frac{\partial^2 y(x,t)}{\partial x^2}, \tag{1}$$

where the relative gap protein concentration $y(x,t)$, at location x and instant t, is driven by the mRNA $u(x,t)$. Here, the translation rate constant S represents the rate of protein production from the mRNA, and parameters λ and D are the decay and diffusion rate constants.

3 Physically-inspired Gaussian processes for post-transcriptional regulation

Physically-inspired Gaussian process (GP) models are stochastic approaches where linear differential equations are encoded into kernel functions (Lawrence et al., 2007; Álvarez et al., 2013). From the data-driven point of view, they can be established without specifying all the physical interactions involved in mechanistic processes. From the physically-inspired models' point of view, they provide accurate predictions even in regions where data are not available. Since they can account for different types of differential equations, physically-inspired GPs have been applied successfully in several fields such as human motion capture and robotics (Álvarez et al., 2013; Agudelo-España et al., 2017; Guarnizo and Álvarez, 2018), neuroscience (Alvarado et al., 2014), and molecular biology and genetics (Lawrence et al., 2007; Álvarez et al., 2013; Vásquez Jaramillo et al., 2014; López-Lopera and Álvarez, 2019; Croix et al., 2018; Gao et al., 2008). In (Álvarez et al., 2013; Vásquez Jaramillo et al., 2014; Croix et al., 2018), they have been applied to model the early embryo development of Drosophila melanogaster.

Using the reaction-diffusion model in (1) as mechanistic model, we can then assume a zero-mean GP prior with covariance function k either over u or y. Since (1) involves only linear operations, the Gaussianity holds for both sides no matter where the GP prior is placed. Let $\mathbf{u}$ and $\mathbf{y}$ be Gaussian vectors containing evaluations of the GPs u and y (respectively) at a given set of couples $(x_i,t_i)_{i=1}^N$. Then, the joint process, at $(x_i,t_i)_{i=1}^N$, follows a multivariate Gaussian distribution,

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{K}_{\mathbf{u},\mathbf{u}} & \mathbf{K}_{\mathbf{y},\mathbf{u}}^\top\\ \mathbf{K}_{\mathbf{y},\mathbf{u}} & \mathbf{K}_{\mathbf{y},\mathbf{y}}\end{bmatrix}\right), \tag{2}$$

with covariance matrices $(\mathbf{K}_{\mathbf{u},\mathbf{u}})_{i,j} = k_{u,u}(x_i,t_i,x_j,t_j)$, $(\mathbf{K}_{\mathbf{y},\mathbf{u}})_{i,j} = k_{y,u}(x_i,t_i,x_j,t_j)$, $(\mathbf{K}_{\mathbf{y},\mathbf{y}})_{i,j} = k_{y,y}(x_i,t_i,x_j,t_j)$ for $i,j = 1,\cdots,N$, with kernel structure $k_{z,z'}$ given by $k_{z,z'}(x_i,t_i,x_j,t_j) = \operatorname{cov}\{z(x_i,t_i), z'(x_j,t_j)\}$. One can note that the physically-inspired GP model is fully established if we encode the mechanistic model from (1) into the covariance matrix of the joint process in (2).

Next, we study two choices of physically-inspired GP models depending on whether the GP prior is placed over the mRNA (GP-mRNA) or the gap protein (GP-Protein).

3.1 GP-mRNA model

Since mRNA data are not always available, GP prior assumptions are commonly placed over mRNA profiles (Álvarez et al., 2013; Vásquez Jaramillo et al., 2014). This approach requires writing the output process y in terms of the driving-force u, with the explicit solution of the PDE in (1). Notice that the complexity of this solution depends on the initial and boundary conditions (Polyanin, 2001; Stakgold and Holst, 2011; Abramowitz and Stegun, 1965). Here we assume homogeneous conditions, i.e. $y(x,t=0)=0$ and $y(x=0,t)=y(x=l,t)=0$ for a diffusion evolution in $x\in[0,l]$ with $l\in\mathbb{R}^+$. These assumptions are made according to the structure of the dataset used in Section 4. Hence, the solution of (1) is given by (Polyanin, 2001; Stakgold and Holst, 2011)

$$y(x,t) = S\int_0^t\int_0^l u(\xi,\tau)G(x,\xi,t-\tau)d\xi d\tau, \tag{3}$$

where the Green's function $G(x,\xi,t)$ is defined as

$$G(x,\xi,t) = c(t)\sum_{n=1}^{\infty}\sin(\omega_n x)\sin(\omega_n\xi)\exp\{-D\omega_n^2\},$$

with $c(t) = \frac{2}{l}\exp\{-\lambda t\}$ and $\omega_n = \frac{n\pi}{l}$. In practice, the Green's function is commonly truncated, and the accuracy of the solution in (3) depends on the number of terms used in the approximation.

Then, a GP prior can be placed over the mRNA u. In order to obtain analytical expressions in further steps, we use a zero-mean GP prior with covariance function $k_{u,u}$ given by the product of two squared exponential (SE) kernel functions, i.e.

$$k_{u,u}(x,t,x',t') = \sigma^2 k(x,x')k(t,t') = \sigma^2 \exp\left\{-\frac{(x-x')^2}{\theta_x^2}\right\}\exp\left\{-\frac{(t-t')^2}{\theta_t^2}\right\}, \tag{4}$$

where θ_x and θ_t are the length-scale parameters.

Now, we aim at computing the covariance function for the output $k_{y,y}$, and the cross-covariance function between the output and the driving-force $k_{y,u}$.

3.1.1 Covariance function for the output

Since the PDE in (1) is linear, the output process y is also a GP with covariance function $k_{y,y}(x,t,x',t') = \text{cov}\{y(x,t), y(x',t')\}$ given by

$$k_{y,y}(x,t,x',t') = \sigma^2 S^2 \int_0^t \int_0^{t'} \int_0^l \int_0^l \widehat{G}(x,\xi,t,\tau,x',\xi',t',\tau') \times k(\xi,\xi')k(\tau,\tau')d\xi' d\xi d\tau' d\tau,$$

with $\widehat{G}(x,\xi,t,\tau,x',\xi',t',\tau') = G(x,\xi,t-\tau)G(x',\xi',t'-\tau')$.

After solving the multiple integrals, one can show that the covariance function $k_{y,y}$ is given by (Álvarez et al., 2013)

$$k_{y,y}(x,t,x',t') = \frac{4\sigma^2 S^2}{l^2} \sum_{\forall n} \sum_{\forall m} \sin(\omega_n x)\sin(\omega_m x')K(t,t',n,m)C(n,m), \tag{5}$$

where

$$K(t,t',n,m) = \frac{\theta_t\sqrt{\pi}}{2}[h(\beta_m,t',t) + h(\beta_n,t,t')],$$

and

$$h(\beta_m,t',t) = \frac{e^{\left(\frac{\beta_m\theta_t}{2}\right)^2}}{\beta_m+\beta_n}\left[e^{-\beta_m(t'-t)}\mathcal{H}(\beta_m,t,t') - e^{-(\beta_m t' + \beta_n t)}\mathcal{H}(\beta_m,0,t')\right],$$

with $\beta_n = \lambda + D\omega_n^2$, $\beta_m = \lambda + D\omega_m^2$, and $\mathcal{H}(\zeta,v,\varphi) = \text{erf}\left\{\frac{\varphi - v}{\theta} - \frac{\theta\zeta}{2}\right\} + \text{erf}\left\{\frac{v}{\theta} + \frac{\theta\zeta}{2}\right\}$. The operator erf denotes the error function (Polyanin, 2001).

The term $C(n,m)$ is defined according to (Álvarez et al., 2013). When $n \neq m$ such that m and n are both even or both odd numbers, $C(n,m)$ follows

$$C(n,m) = \frac{\theta_x l}{\sqrt{\pi}(m^2-n^2)}\left\{n\mathcal{I}\left[\mathcal{W}_{\theta_x}(m)\right] - m\mathcal{I}\left[\mathcal{W}_{\theta_x}(n)\right]\right\},$$

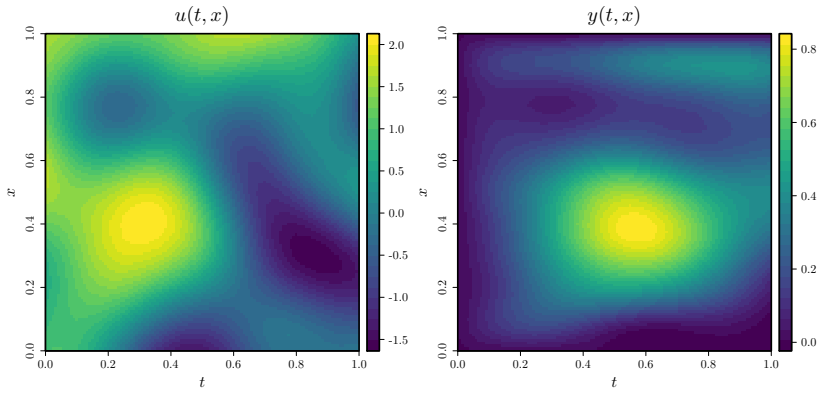
where $\mathcal{I}$ is the operator returning the imaginary part of the argument, and

$$\mathcal{W}_{\theta_x}(m) = w(jz_1^{\gamma_m}) - e^{-\left(\frac{l}{\theta_x}\right)^2}e^{-\gamma_m l}w(jz_2^{\gamma_m}),$$

with $\gamma_n = j\omega_n$, $\gamma_m = j\omega_m$, $z_1^{\gamma_n} = \frac{\theta_x\gamma_n}{2}$ and $z_2^{\gamma_n} = \frac{l}{\theta_x} + \frac{\theta_x\gamma_n}{2}$. We note that $w(z) = \exp\{-z^2\}\,\text{erfc}\,\{-jz\}$ is known as the Faddeeva function (Poppe and Wijers, 1990), with operator erfc denoting the complementary error function. Otherwise, if $n \neq m$ but m and n are not both even or both odd numbers, then $C(n,m) = 0$. If $n = m$, then

$$C(n,n) = \frac{\theta_x\sqrt{\pi}\,l}{2}\left\{\mathcal{R}\left[\mathcal{W}_{\theta_x}(n)\right] - \mathcal{I}\left[\mathcal{W}_{\theta_x}(n)\right]\left[\frac{\theta_x^2 n\pi}{2l^2} + \frac{1}{n\pi}\right]\right\} + \frac{\theta_x^2}{2}\left[e^{-(\frac{l}{\theta_x})^2}\cos(n\pi) - 1\right],$$

where $\mathcal{R}$ is the operator returning the real part of the argument.



(a) mRNA (b) Protein

Figure 1: Synthetic example generated by the GP-mRNA model.

3.1.2 Covariance function between the output and the driving-force

The cross-covariance function $k_{y,u}(x,t,x',t') = \text{cov}\left\{y(x,t), u(x',t')\right\}$ between the output y and the driving-force u is given by

$$k_{y,u}(x,t,x',t') = \sigma^2 S \int_0^t \int_0^l G(x,\xi,t-\tau)k(\xi,x')k(\tau,t')d\xi d\tau.$$

After solving the double integral, one can show that (Álvarez et al., 2013)

$$k_{y,u}(x,t,x',t') = \frac{2\sigma^2 S}{l} \sum_{\forall n} \sin\left(\omega_n x\right) \widetilde{K}(t,t',n)\widetilde{C}(x',n), \tag{6}$$

where

$$\widetilde{K}(t,t',n) = \frac{\theta_t\sqrt{\pi}}{2} e^{\left(\frac{\beta_n\theta_t}{2}\right)^2} e^{-\beta_n(t-t')}\mathcal{H}(\beta_n,t',t),$$

$$\widetilde{C}(x',n) = \frac{\theta_x\sqrt{\pi}}{2}\mathcal{I}\left[\widetilde{\mathcal{W}}_{\theta_x}(x',n)\right],$$

$$\widetilde{\mathcal{W}}_{\theta_x}(x',n) = e^{-\left(\frac{x'-l}{\theta_x}\right)^2} e^{\gamma_n l} w(jz_2^{\gamma_n,x'}) - e^{-\left(\frac{x'}{\theta_x}\right)^2} w(jz_1^{\gamma_n,x'}),$$

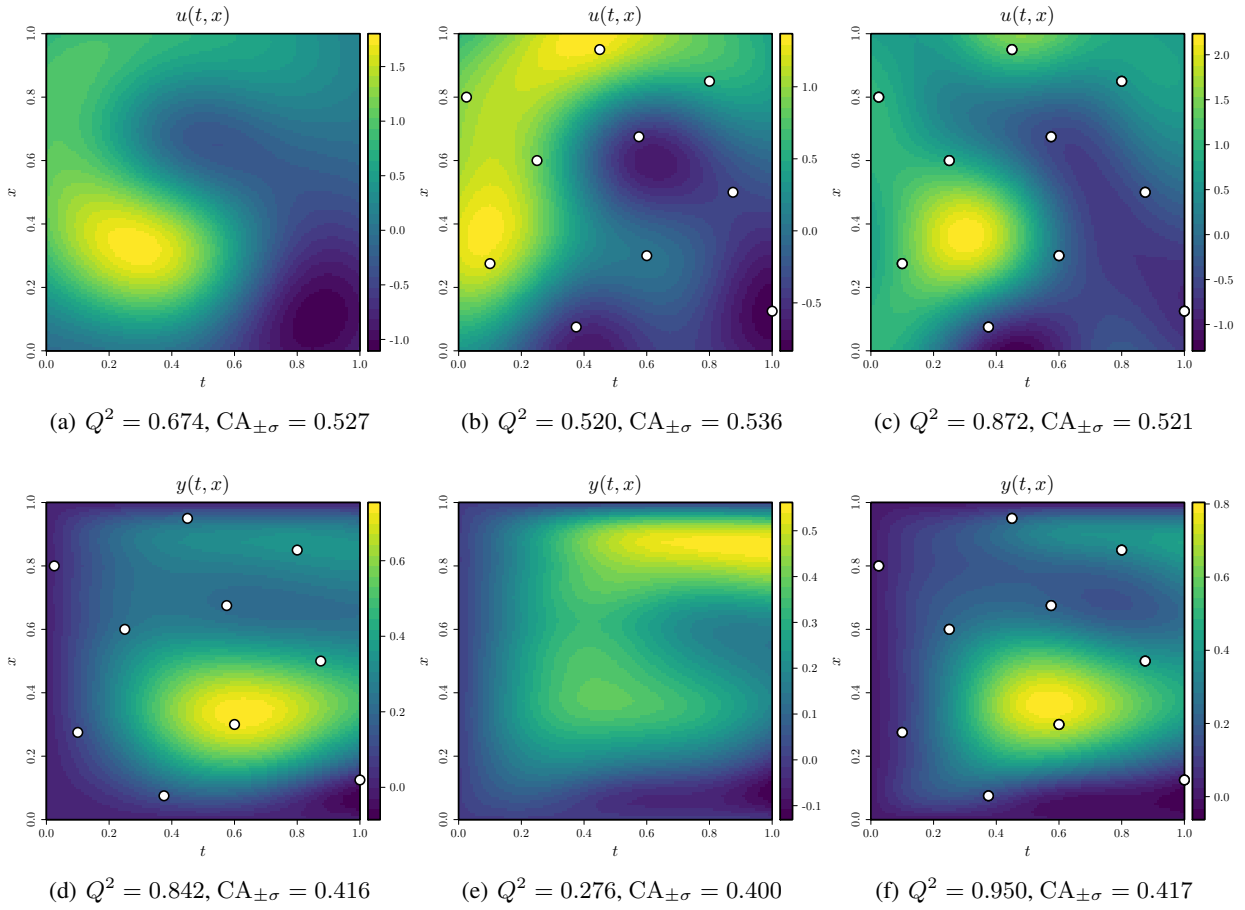
with $z_1^{\gamma_n,x'} = \frac{x'}{\theta_x} + \frac{\theta_x\gamma_n}{2}$ and $z_2^{\gamma_n,x'} = \frac{x'-l}{\theta_x} + \frac{\theta_x\gamma_n}{2}$. Finally, the process in (2) can be computed using (4), (5) and (6).

One must note that the stability of the GP-mRNA model, besides depending on the number of terms of the Green's function, it also depends on the computation of the erf and Faddeeva functions (see expressions (6) and (5)). Since those functions do not have closed-form expressions, they have to be computed numerically (see, e.g., Poppe and Wijers, 1990; Weideman, 1994).

3.1.3 Toy example: inference of simulated data

To illustrate the properties of the GP-mRNA model, we generate synthetic data by sampling from the joint GP using the kernels functions (4), (5) and (6). We consider the domain $(x,t) \in [0,1]^2$, and we use ten components of the Green's function. We fix as covariance parameters $\sigma^2 = 1$, $\theta_x = \theta_t = 0.3$, and as mechanistic parameters $S = 1$, $\lambda = 0.1$, and $D = 0.01$. Samples are generated using a 41×41 equispaced grid on $[0,1]^2$. Figure 1 shows the generated mRNA and protein profiles. One can observe that the homogeneous conditions are ensured in the protein profile. Also notice that the GP-mRNA model does not guarantee that the mRNA and gap proteins are strictly positive quantities. As for standard GP models, GP-mRNA cannot account for positivity but non-linear transformations of GPs can be applied for ensuring it everywhere (e.g. exponential of GPs Vanhatalo and Vehtari, 2007). However, those transformations do not yield analytical solutions of the resulting joint GP as we provided in Section 3.1. Therefore, the synthetic example proposed here is for illustrative purposes only.

We aim at testing the performance of GP-mRNA under three conditions. In the first two cases, we establish a joint GP using data only from the mRNA or the protein, and we then estimate both quantities in the whole domain $[0,1]^2$. We



(a) $Q^2 = 0.674$, $\mathrm{CA}_{\pm\sigma} = 0.527$ (b) $Q^2 = 0.520$, $\mathrm{CA}_{\pm\sigma} = 0.536$ (c) $Q^2 = 0.872$, $\mathrm{CA}_{\pm\sigma} = 0.521$

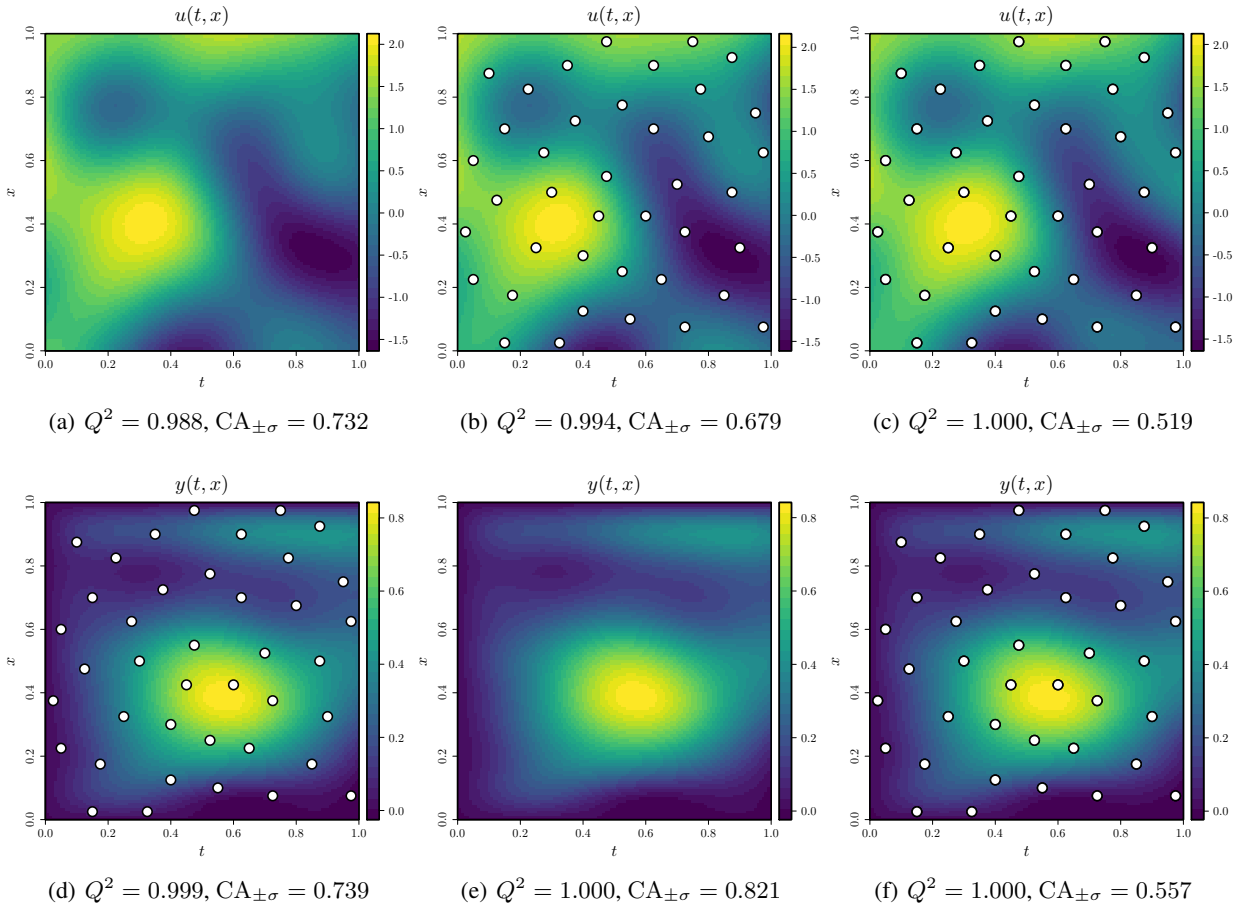
(d) $Q^2 = 0.842$, $\mathrm{CA}_{\pm\sigma} = 0.416$ (e) $Q^2 = 0.276$, $\mathrm{CA}_{\pm\sigma} = 0.400$ (f) $Q^2 = 0.950$, $\mathrm{CA}_{\pm\sigma} = 0.417$

Figure 2: GP-mRNA prediction results using conditioning data either only from the mRNA (left), or from the protein (centre), or from both of them (right). Conditioning points (white dots) were chosen using a maximin LHD with 10 points, and the quality of predictions is assessed using the Q^2 and $\mathrm{CA}_{\pm\sigma}$ criteria.

repeat the same procedure using conditioning data from both the mRNA and protein rather than only from one of them. We use the $Q^2 = 1 - \mathrm{SMSE}$ criterion, where SMSE is the standardised mean squared error (Rasmussen and Williams, 2005), to evaluate the quality of predictions over the points that were not used for training the GP model (test data). For noise-free observations, the Q^2 criterion is equal to one if the predictive mean of the resulting process is equal to the test data and lower than one otherwise. To evaluate the quality of the predictive variances, we use a criterion based on the coverage accuracy (CA) of the confidence intervals. For one standard deviation intervals, predictive variances should cover around 68% of the test points (Meyer, 1970). Departure from $\mathrm{CA}_{\pm\sigma} = 0.68$ may indicate that the predictive variances are either underestimated (i.e. $\mathrm{CA}_{\pm\sigma} < 0.68$) or overestimated (i.e. $\mathrm{CA}_{\pm\sigma} > 0.68$). For a further discussion on assessing the quality of predictions using GP-mRNA, we assume that both covariance and mechanistic parameters are known and are equal to the ones used to generate the synthetic data.

Figure 2 shows the performance of the GP-mRNA model using a maximin Latin hypercube design (LHD) at ten locations.[2] For the case when only conditioning data from the protein concentration are used, we can observe that GP-mRNA presents accurate performances to reconstruct both the mRNA and protein profiles providing Q^2 results above 0.67. Since conditioning data belongs to the protein profile where the mechanistic parameters were encoded, this information is taken into account in the inference of the mRNA. On the other hand, when only conditioning mRNA data are used, predictions over both quantities are poor. There, the influence of the PDE over the conditional process seems to be weak due to conditioning data belonging to the GP prior. Finally, one can observe that predictions are

[2] A maximin LHD is a space-filling design consisting in the iterative maximisation of the distance between two closest design points from a random LHD. In this paper, we used the simulated annealing routine `maximinSA_LHS` from the R package `DiceDesign` (Dupuy et al., 2015).



(a) $Q^2 = 0.988$, $\mathrm{CA}_{\pm\sigma} = 0.732$ (b) $Q^2 = 0.994$, $\mathrm{CA}_{\pm\sigma} = 0.679$ (c) $Q^2 = 1.000$, $\mathrm{CA}_{\pm\sigma} = 0.519$

(d) $Q^2 = 0.999$, $\mathrm{CA}_{\pm\sigma} = 0.739$ (e) $Q^2 = 1.000$, $\mathrm{CA}_{\pm\sigma} = 0.821$ (f) $Q^2 = 1.000$, $\mathrm{CA}_{\pm\sigma} = 0.557$

Figure 3: GP-mRNA prediction results. Panel description is the same as in Figure 2. Conditioning data were chosen using a LHD with 40 points.

considerably improved when data from both the protein and mRNA are used. In that case, we obtained improvements of the Q^2 criterion obtaining values above 0.85. Although the predictive variances were commonly underestimated, resulting $\mathrm{CA}_{\pm\sigma}$ values are not far from the expected 68% (with departures of 15-28$\%$).

In Figure 3, we show that if the number of conditioning points increases, the performance of the GP-mRNA model improves, obtaining Q^2 results above 0.98 (and equal to one when data from both sides are used), and $\mathrm{CA}_{\pm\sigma}$ values closer to 68% with maximum departures of 20%.

3.2 GP-Protein model

As shown in Section 3.1, the establishment of the GP-mRNA models requires the explicit solution of the reaction-diffusion PDE in (1) and the multiple-integration of kernel functions. Both calculations commonly require the evaluation of cumbersome terms which are not always feasible to compute (see discussion in Section 3.1). In this paper, we suggest placing the GP prior over the protein y rather than over the mRNA u. This leads to a novel alternative where building up the resulting GP model is simpler since the solution of the PDE is not required.

The reaction-diffusion dynamics from (1) can be written in terms of the protein y, obtaining

$$u(x,t) = \frac{1}{S}\left[\frac{\partial y(x,t)}{\partial t} + \lambda y(x,t) - D\frac{\partial^2 y(x,t)}{\partial x^2}\right]. \tag{7}$$

As for the GP-mRNA model, we use the same SE kernel structure of (4) for the covariance function of the GP prior over y, i.e.

$$k_{y,y}(x,t,x',t') = \sigma^2 k(x,x')k(t,t'), \tag{8}$$

where $k(z, z') = \exp\{-(z - z')^2/\theta_z^2\}$ with length-scale θ_z.

One can note that the establishment of the GP-Protein model is not restricted to the use of SE kernel functions, and that other classes of differentiable kernels can be used instead (e.g. Matérn family of covariance functions) (Rasmussen and Williams, 2005).

Now, we need to compute the covariance function for the driving-force $k_{u,u}$, and the cross-covariance function between the output and the driving-force $k_{y,u}$. For ease of readability, next we summarise the expressions required for the computation of $k_{u,u}$ and $k_{y,u}$. We refer to Appendix 1 for further details.

3.2.1 Covariance function for the driving-force

Since (7) involves only the differentiation of the output process y, and due to the symmetry of SE kernel functions, the covariance function for the mRNA is given by

$$k_{u,u}(x,t,x',t') = \frac{\sigma^2 D}{S^2}\left[Dk^{iv}(x,x') - 2\lambda k^{ii}(x,x')\right]k(t,t') - \frac{\sigma^2}{S^2}\left[k^{ii}(t,t') - \lambda^2 k(t,t')\right]k(x,x'), \tag{9}$$

where $k^j(z, z')$ is the j-th derivative of the SE kernel $k(z, z')$ w.r.t. the input z. Then, the complexity of the problem is in the computation of the corresponding derivatives of the SE kernel function, and they follow

$$\begin{aligned} k^i(z,z') &= \left[-\frac{2(z-z')}{\theta_z^2}\right]k(z,z'), \\ k^{ii}(z,z') &= \left[-\frac{2}{\theta_z^2} + \frac{4(z-z')^2}{\theta_z^4}\right]k(z,z'), \\ k^{iv}(z,z') &= \left[\frac{12}{\theta_z^4} - \frac{48(z-z')^2}{\theta_z^6} + \frac{16(z-z')^4}{\theta_z^8}\right]k(z,z'). \end{aligned} \tag{10}$$

Notice that, since we have to differentiate partially $k_{y,y}$ four times, the GP-Protein model is limited to the use of differentiable kernels (e.g. Matérn family of covariance functions with regularity parameter $\nu > \frac{5}{2}$) (Rasmussen and Williams, 2005; Stein, 1999).

3.2.2 Covariance function between the driving-force and the output

The cross-covariance function between the output y and the force u, $k_{y,u}(x,t,x',t') = \text{cov}\{y(x,t), u(x',t')\}$ is given by

$$k_{y,u}(x,t,x',t') = \frac{\sigma^2}{S}[\lambda k(x,x')k(t,t') - k(x,x')k^i(t,t') - Dk^{ii}(x,x')k(t,t')], \tag{11}$$

with derivatives of the SE kernel given in (10).

3.2.3 Toy example: inference of simulated data

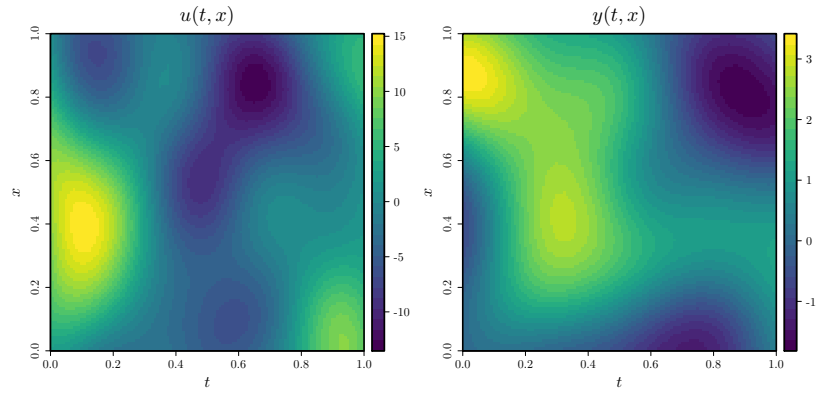
As in Section 3.1.3, we generate a synthetic example by sampling from the GP in (2) using the kernel functions (8), (9) and (11). We assume the same parametrisation used for GP-mRNA. Figure 4 shows the generated mRNA and protein. One can observe that since we did not enforce the initial and boundary conditions, homogeneous conditions are not necessarily ensured by the GP-Protein model.

Now, we test the performance of the GP-Protein model under the three conditions studied in Section 3.1.3. Figure 5 shows the performance of GP-Protein using a fixed maximin LHD at ten locations. As observed in Section 3.1.3, if only conditioning points are used from the mRNA or protein, predictions over the unobserved quantity are less reliable; and they improve when data are available from both sides. In the latter case, we obtain Q^2 values above 0.87 and $\text{CA}_{\pm\sigma}$ values around 57-68%. In figure 6, one can observe that if the number of conditioning points increases, the performance of the GP-Protein also improves with Q^2 values close to one in almost all the cases, and $\text{CA}_{\pm\sigma}$ values almost equals to 68% when using data from both quantities.

4 Results and Discussions

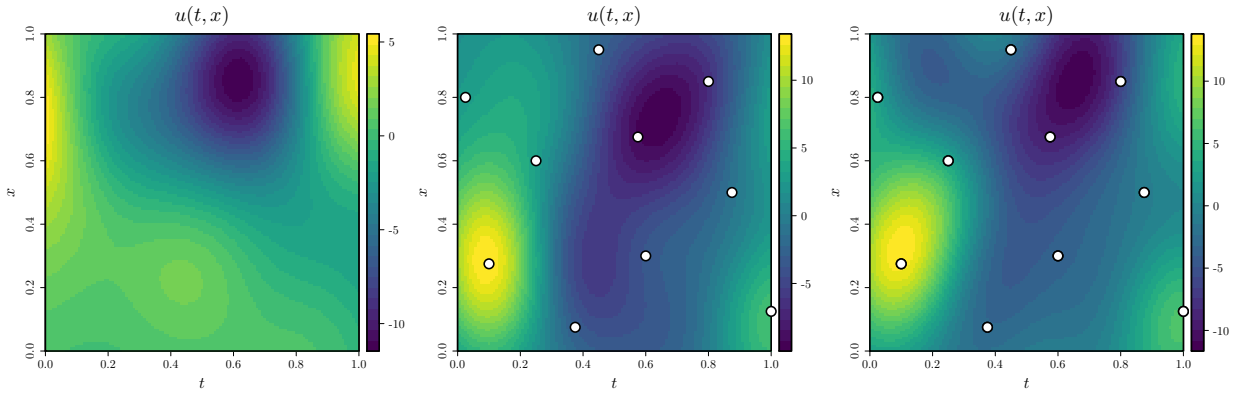
4.1 Numerical setup

Both physically-inspired GP approaches for modelling the post-transcriptional regulation of the early embryo of Drosophila melanogaster were implemented in R, and the codes are available on Github: https://github.com/

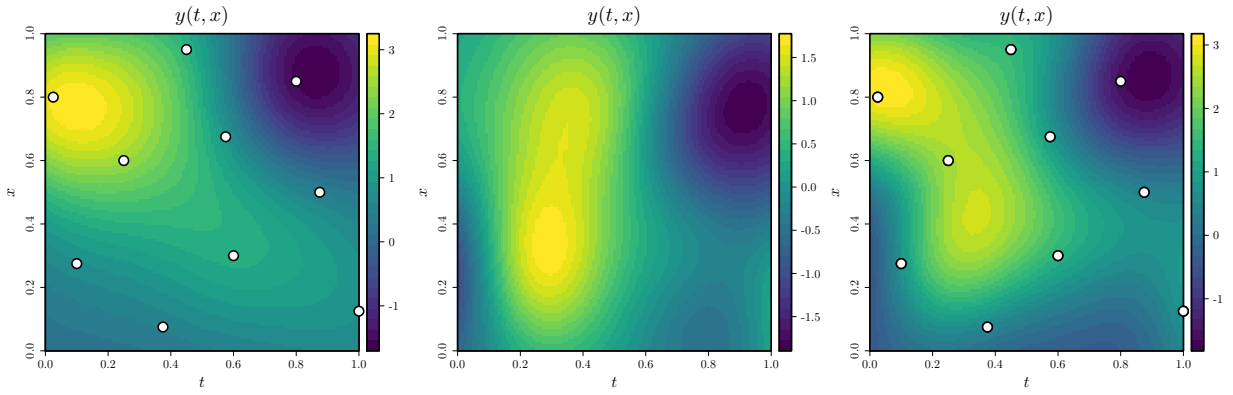


(a) mRNA (b) Protein

Figure 4: Synthetic example generated by the GP-Protein model.



(a) $Q^2 = 0.373$, $\mathrm{CA}_{\pm\sigma} = 0.504$ (b) $Q^2 = 0.756$, $\mathrm{CA}_{\pm\sigma} = 0.685$ (c) $Q^2 = 0.893$, $\mathrm{CA}_{\pm\sigma} = 0.571$

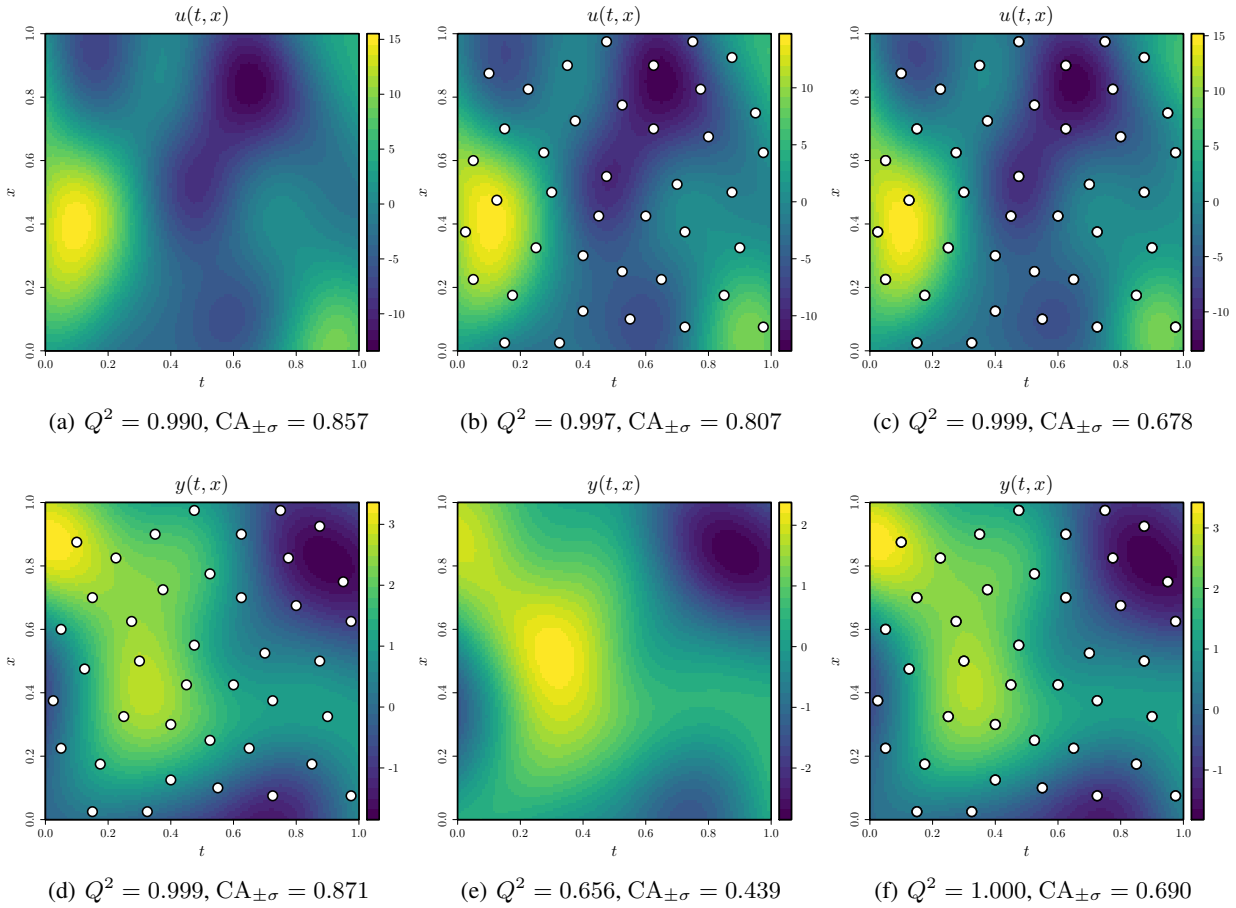


(d) $Q^2 = 0.714$, $\mathrm{CA}_{\pm\sigma} = 0.558$ (e) $Q^2 = 0.467$, $\mathrm{CA}_{\pm\sigma} = 0.555$ (f) $Q^2 = 0.948$, $\mathrm{CA}_{\pm\sigma} = 0.683$

Figure 5: GP-Protein prediction results using conditioning data either only from the mRNA (left), or from the protein (centre), or from both of them (right). Conditioning points (white dots) were chosen using a maximin LHD with 10 points, and the quality of predictions is assessed using the Q^2 and $\mathrm{CA}_{\pm\sigma}$ criteria.

`anfelopera/PhysicallyGPDrosophila`. They are based on the R package `kergp`.[3] For the computation of the erf

[3]The `kergp` project is an open source R package available in CRAN for Gaussian Process models with customised covariance kernels (Deville et al., 2015).



(a) $Q^2 = 0.990$, $\text{CA}_{\pm\sigma} = 0.857$ (b) $Q^2 = 0.997$, $\text{CA}_{\pm\sigma} = 0.807$ (c) $Q^2 = 0.999$, $\text{CA}_{\pm\sigma} = 0.678$

(d) $Q^2 = 0.999$, $\text{CA}_{\pm\sigma} = 0.871$ (e) $Q^2 = 0.656$, $\text{CA}_{\pm\sigma} = 0.439$ (f) $Q^2 = 1.000$, $\text{CA}_{\pm\sigma} = 0.690$

Figure 6: GP-Protein prediction results. Panel description is the same as in Figure 5. Conditioning data were chosen using a LHD with 40 points.

Algorithm 1 Physically-inspired GPs for transcriptional regulation of the early embryo of Drosophila melanogaster.

1: **procedure** PREDICTION OF MRNA u AND GAP PROTEIN y
2: Input: training set $\mathcal{D} = (\mathbf{x}, \mathbf{t}, \mathbf{u}, \mathbf{y})$,[a] initial set of hyperparameters $\boldsymbol{\theta} = \{S, \lambda, D, \sigma^2, \theta_x, \theta_y\}$
3: Compute the covariance matrices $\mathbf{K}_{\mathbf{y},\mathbf{y}}$, $\mathbf{K}_{\mathbf{u},\mathbf{u}}$, and $\mathbf{K}_{\mathbf{y},\mathbf{u}}$ according to Section 3.
4: Estimate the hyperparameters $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log\{p_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{y})\}$.
5: According to (Rasmussen and Williams, 2005), compute the conditional distribution for the test set $\mathcal{D}^* = (\mathbf{x}^*, \mathbf{t}^*)$, i.e. $p(\mathbf{u}^*, \mathbf{y}^* | \mathcal{D})$.

[a] One may note that is not necessary to have access to conditioning data from both y and u simultaneously.

and Faddeeva functions, after testing the numerical stability of various R packages, we chose the *pracma* (Borchers, 2012) and `NORMT3` (Nason, 2012) packages. The hyperparameters $\boldsymbol{\theta} = (S, \lambda, D, \sigma^2, \theta_x, \theta_t)$ are estimated by maximising the joint marginal log-likelihood $p_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{y})$ using gradient-descent methods (Rasmussen and Williams, 2005), i.e. $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log\{p_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{y})\}$. Algorithm 1 shows the pseudo-code of both GP-mRNA and GP-Protein for the post-transcriptional regulation of Drosophila.

4.2 Quantitative gap gene mRNA expression data

Here we aim at testing the performance of both physically-inspired GP models from Section 3 on the high spatial and temporal resolution dataset used in (Becker et al., 2013), describing the entire duration of the blastoderm stage for the early embryo of Drosophila melanogaster. This dataset exhibits homogeneous conditions, and it contains quantified independent time-series of gap gene mRNA expressions for the trunk gap genes *Krüppel* (kr), *knirps* (kni) and *giant*

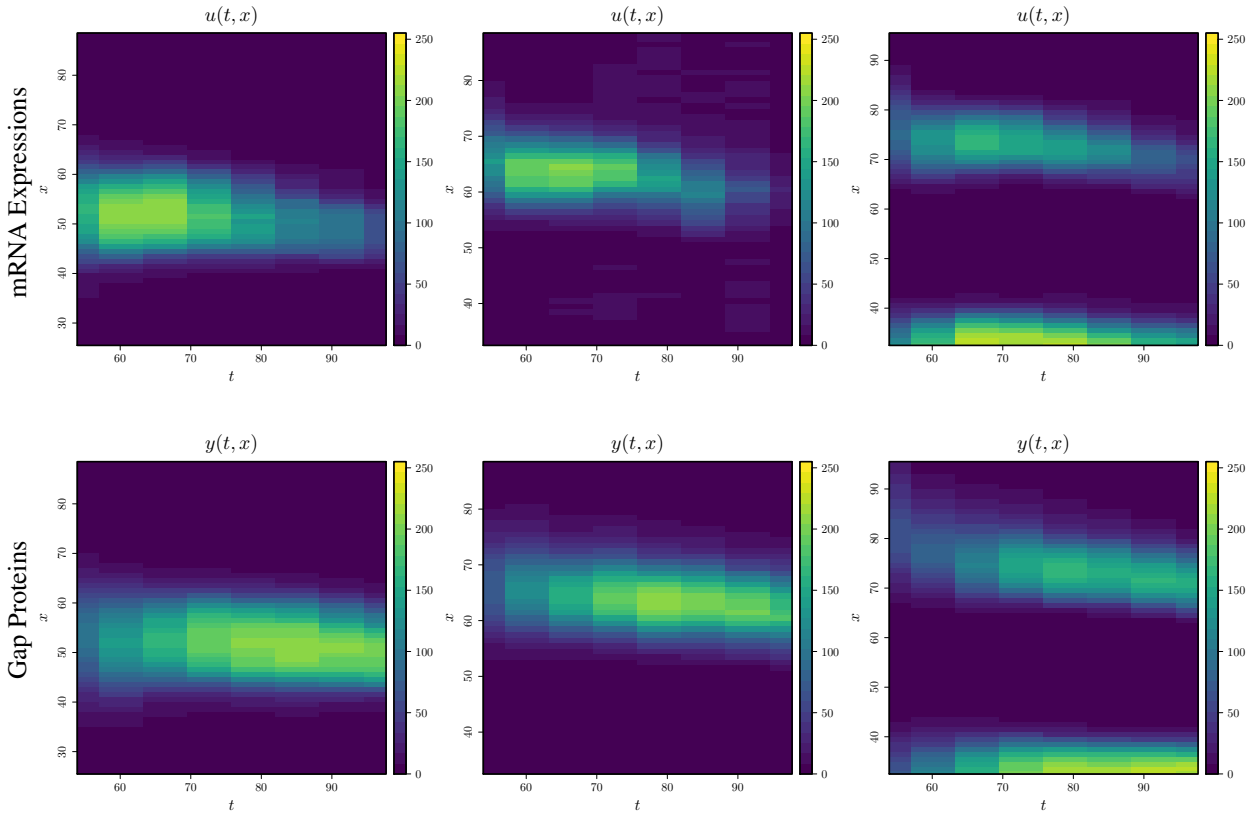


Figure 7: Gap gene mRNA expression data from (Becker et al., 2013) for *Krüppel* (left), *knirps* (centre), and *giant* (right) trunk gap genes.

Table 1: Estimated parameters by (Becker et al., 2013) via LSA-based global optimisation.

Parameter	Trunk Gap Gene		
	Krüppel (kr)	*knirps* (kni)	*giant* (gt)
Translation rate (S)	0.0970	0.0783	0.1107
Decay rate (λ)	0.0764	0.0770	0.1110
Diffusion rate (D)	0.0015	0.0125	0.0159

(gt). We ignored the observations in the range $t \in [0, 50]$ due to the poor quality of the data, and we focused on the late part where the data is equispaced: after the first 53 min and around A-P positions of $x = 25.5\%$ for kr, and $x = 32.5\%$ for kni and gt. This leads to a total of 512, 456 and 512 measurements for the trunk gap genes kr, kni and gt, respectively.

Figure 7 shows both mRNA and gap protein concentrations for each trunk gap gene of the dataset. We observe that both profiles exhibit similar patterns for each trunk gap gene with an active synthesis of proteins in the A-P intervals between 50-80%, except for the trunk gap gene gt where there is also synthesis between 30-40%. We can also note that the synthesis of the three gap proteins remains almost equal along the time axis.

4.3 Inference results

For each numerical experiment, we randomly selected the 30% of the available data from both biological profiles to train each of the GP model, and the remaining 70% of the data is used to test the quality of the models. For the hyperparameter parameter estimation, the mechanistic parameters (S, λ, D) were fixed to be equal to the ones estimated by (Becker et al., 2013) via Lam simulated annealing (LSA)-based global optimisation (see Table 1), and the covariance parameters $(\sigma^2, \theta_x, \theta_t)$ are estimated via maximum likelihood. We repeat this procedure for ten different random

Table 2: Prediction assessment of the physically-inspired GP models using the dataset from (Becker et al., 2013). Predictive accuracies of ten repetitions with different training sets are evaluated using the Q^2 (left) and $\mathrm{CA}_{\pm\sigma}$ (right) criteria. The mean μ and the standard deviations σ of the results are shown using GP-mRNA and GP-Protein models for: *Krüppel* (kr), *knirps* (kni) and *giant* (gt). Best Q^2 results for each trunk gap gene (rows) are shown in bold.

Q^2[%] Results

Trunk Gap Gene	GP-mRNA mRNA $\mu \pm \sigma$	GP-mRNA Gap Protein $\mu \pm \sigma$	GP-Protein mRNA $\mu \pm \sigma$	GP-Protein Gap Protein $\mu \pm \sigma$
Training data only from the gap protein concentration				
kr	$\mathbf{90.5 \pm 2.0}$	$\mathbf{90.8 \pm 0.6}$	$\mathbf{92.0 \pm 0.6}$	$\mathbf{90.6 \pm 0.5}$
kni	$\mathbf{81.1 \pm 2.5}$	$\mathbf{88.7 \pm 0.8}$	77.6 ± 4.7	88.6 ± 0.7
gt	91.2 ± 1.9	92.3 ± 0.6	$\mathbf{93.2 \pm 1.3}$	$\mathbf{92.8 \pm 0.5}$
Training data only from the mRNA concentration				
kr	$\mathbf{86.7 \pm 1.4}$	$\mathbf{97.5 \pm 0.7}$	84.0 ± 2.1	60.6 ± 1.2
kni	$\mathbf{82.9 \pm 2.1}$	$\mathbf{86.7 \pm 1.3}$	80.7 ± 3.2	55.2 ± 12.7
gt	$\mathbf{91.2 \pm 0.7}$	$\mathbf{93.9 \pm 0.3}$	88.2 ± 3.1	84.3 ± 1.7
Training data from both biological quantities				
kr	96.8 ± 0.5	97.9 ± 0.3	$\mathbf{98.6 \pm 0.6}$	$\mathbf{99.6 \pm 0.2}$
kni	91.2 ± 2.9	95.0 ± 0.7	$\mathbf{94.5 \pm 3.5}$	$\mathbf{99.4 \pm 0.3}$
gt	95.2 ± 1.4	96.2 ± 0.6	$\mathbf{97.7 \pm 1.7}$	$\mathbf{99.3 \pm 0.2}$

$\mathrm{CA}_{\pm\sigma}$[%] Results

Trunk Gap Gene	GP-mRNA mRNA $\mu \pm \sigma$	GP-mRNA Gap Protein $\mu \pm \sigma$	GP-Protein mRNA $\mu \pm \sigma$	GP-Protein Gap Protein $\mu \pm \sigma$
Training data only from the gap protein concentration				
kr	71.7 ± 6.7	46.2 ± 4.0	69.2 ± 1.3	57.2 ± 1.3
kni	83.0 ± 5.3	37.1 ± 4.4	49.7 ± 6.3	32.7 ± 5.3
gt	85.6 ± 6.9	31.8 ± 2.7	68.2 ± 3.3	42.2 ± 1.8
Training data only from the mRNA concentration				
kr	57.3 ± 1.5	80.7 ± 2.8	58.2 ± 2.3	78.0 ± 1.3
kni	48.4 ± 5.4	64.7 ± 4.2	54.9 ± 7.8	74.3 ± 3.5
gt	40.0 ± 3.1	63.0 ± 2.4	53.1 ± 4.2	66.8 ± 1.4
Training data from both biological quantities				
kr	34.5 ± 2.8	21.9 ± 4.6	81.9 ± 2.8	87.7 ± 2.0
kni	20.5 ± 2.4	13.7 ± 2.2	74.4 ± 7.1	86.6 ± 1.9
gt	19.6 ± 4.5	14.6 ± 3.1	79.8 ± 5.6	83.5 ± 2.6

training sets. For the GP-mRNA model, we choose the number of terms of the Green's function according to two criteria: the quality of the predictions and the computational cost. We gradually increased the number of terms, starting with the first five terms, and we checked the quality of the resulting model in terms of the Q^2 and $\mathrm{CA}_{\pm\sigma}$ criteria. We observed that the results became stable and accurate after the first twenty terms. Finally, we tested the same three conditions of data availability discussed in Sections 3.1.3 and 3.2.3.

Table 2 shows the performance of both physically-inspired GP models for ten repetitions using different training sets. The mean μ and the standard deviations σ of the Q^2 and $\mathrm{CA}_{\pm\sigma}$ results are shown. One can note that when only mRNA or gap protein data were used to train the models, better Q^2 values were commonly obtained when the GP prior was placed over the process where data were available. Although both models yielded similar departures of $\mathrm{CA}_{\pm\sigma}$ percentages. This result agrees with the log-likelihood performances of both GP models (see Table 3). This comparison is valid since the biological parameters (S, λ, D) were assumed to be known. However, these parameters are commonly unknown, and they have to be estimated in real applications. Since (S, λ, D) are not encoded in the covariance function of the GP prior, they cannot be learned if training data are available only from the prior. In this sense, in real applications, GP priors should be placed over the unobserved processes as suggested in (Lawrence et al., 2007; Álvarez et al., 2013).

Another interesting result can be pointed out by the performance of GP-mRNA model. This model was previously studied in (Álvarez et al., 2013; Vásquez Jaramillo et al., 2014) for the inference of mRNA using gap protein data only. However, it could not be further tested due to the lack of mRNA data, and inference results were justified according to qualitative criteria. Here, one can observe from Table 2 that GP-mRNA yielded accurate quantitative results, with Q^2 values over the 80%, on both biological quantities independently on the training data availability. However, we must note that the GP-mRNA model led to costly procedures due to the evaluation of more expensive kernel structures (e.g. depending on the number of terms from the Green's function, and the computation of the erf and Faddeeva functions). More precisely, while the parameter estimation using the GP-Protein model takes a couple of minutes, the running time for GP-mRNA is in the order of hours.

Finally, when both mRNA and gap protein concentration data are used to train the models, we can note that the GP-Protein model outperformed the results provided by GP-mRNA with Q^2 improvements around 3-5% in all the cases. Here, the hyperparameters $\boldsymbol{\theta} = (S, \lambda, D, \sigma^2, \theta_x, \theta_t)$ were estimated via maximum likelihood with an initial set of biological parameters (S, λ, D) given by Table 1. The choice of using the estimated values of Table 1 as starting point is due to, according to numerical experiments, it seems that (S, λ, D) cannot be estimated consistently. As some covariance parameters from certain GP models cannot be estimated consistently due to their non-microergodicity (Stein, 1999; Zhang, 2004), we believe that both GP-mRNA and GP-Protein models suffer from the same downside. Finally, after convergence of the maximum likelihood estimation, we observed that the estimated values of (S, λ, D) remained around the ones from Table 1.

According to the $\mathrm{CA}_{\pm\sigma}$ results, one can observe that the GP-Protein model provide a more reasonable predictive variances than GP-mRNA. The harsh underestimation of the predictive intervals by GP-mRNA is produced by numerical

Table 3: Log-likelihood performance of GP-mRNA and GP-Protein for one repetition. First and second best results are shown in bold and grey.

Model	Training Data Usage	*Krüppel* (kr)	*knirps* (kni)	*giant* (gt)
GP mRNA	Protein	-53518.6	-50542.3	-42692.8
	mRNA	**-53502.0**	**-50537.0**	**-42654.9**
	Protein & mRNA	**-19603.9**	**-36868.5**	**-31246.4**
GP Protein	Protein	**-53502.0**	**-50537.0**	**-42654.9**
	mRNA	-53696.7	-50742.9	-42905.7
	Protein & mRNA	**-1238.7**	**-1181.0**	**-1269.7**

instabilities in the computation of their covariance matrices and the gradients of the joint process. In practice, one possible solution to avoid this overfitting is the early stopping of the maximum likelihood optimisation. In terms of the likelihood performance, we also noticed that results using GP-Protein are of a lesser order of magnitude to those obtained by GP-mRNA (see Table 3). This suggests that the GP-Protein model better describes the behaviour of the three trunk gap genes. Since homogeneous conditions are enforced in GP-mRNA, we believe those constraints may affect the likelihood performance of the model.

Figure 8 shows the obtained predictive means for one of the repetitions of the three trunk gap genes for both the GP-mRNA and GP-Protein models. One can observe that both models are able to capture the correlation between the mRNA and gap protein concentrations. In particular, we note that both models are capable of precisely recovering the time lag between the peaks in the mRNAs and the ones in the proteins for the three trunk gap genes: about 15-20 min. We also observe that the GP-Protein model provides smoother profiles than the ones obtained by GP-mRNA.

As pointed out in Section 3, one must note that both GP-mRNA and GP-Protein models do not necessarily guarantee that the mRNA and gap proteins are strictly positive quantities. In practice, positiveness assumptions can be fulfilled by positive non-linear transformations (e.g. exponential of GPs Vanhatalo and Vehtari, 2007). However, those transformations do not yield analytical solutions of the resulting joint GP as we provided in Section 3. Another possible approach to guarantee positiveness conditions is based on finite-dimensional approximations of GPs (Maatouk and Bay, 2017; López-Lopera et al., 2018). This approach could be potentially investigated in future implementations.

5 Conclusion

We have studied two types of physically-inspired Gaussian process (GP) regression approaches to model the post-transcriptional regulation of the early embryo of Drosophila. Both approaches are based on a continuous version of the linear reaction-diffusion differential equation. The main difference between both GP models lies on whether the GP prior is placed: either over the mRNA (GP-mRNA) or gap protein (GP-Protein). First, for the GP-mRNA model (framework known as *latent force model*), previous studies have been restricted to the use of gap protein data due to the lack of mRNA data (Álvarez et al., 2013; Vásquez Jaramillo et al., 2014). In this paper, we tested it when the information from both mRNA and protein concentrations are available, and we analysed their performance under different situations depending on the availability of data. Second, we introduced the GP-Protein model as a novel alternative where the complexity of computations is reduced to the differentiation of kernels.

We studied three conditions depending on the availability of data: whether from the mRNA, the gap protein or both quantities. We concluded that both models provide promising predictions when the number of training data is large enough. We also tested them in a real-world biological problem to model the early embryo of Drosophila. One interesting result we could pointed out from the GP-mRNA model is referred by its reliable inference results. We observed it yielded accurate results, with Q^2 values over the 80%, independently of the training data availability. Finally, we also observed that GP-Protein model slightly outperformed the prediction results provided by GP-mRNA when training data from both quantities were used.

According to numerical experiments proposed in this paper, we have different recommendations depending on the data availability. First, we recommend the use of the GP-Protein model when data are available from both mRNA and gap protein concentrations. This proposition stands due to its numerical stability, computational cost and accurate performance. Second, if data are available from only one of the biological quantities, we suggest placing the GP prior assumption over the unobserved profile in order to obtain more accurate inference results and to be able to learn both the mechanistic and covariance parameters via maximum likelihood estimation. Finally, we may prefer one of the GP

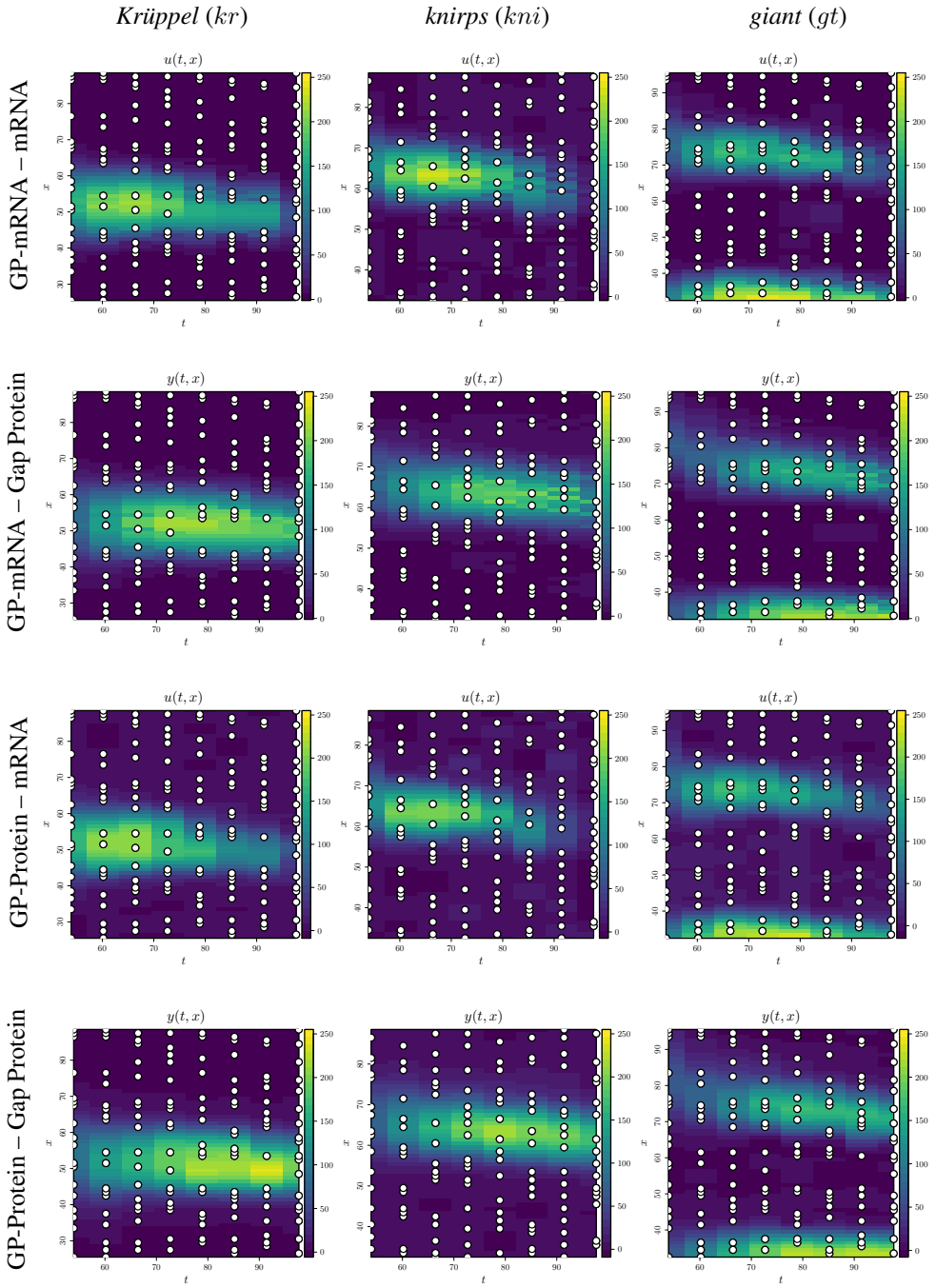


Figure 8: Prediction results on the three trunk gap proteins kr, kni and gt (columns) using GP-mRNA (first and third row) and GP-Protein (second and fourth row). Training points (white dots) correspond to the 30% of the dataset.

models over the other depending on the nature of the biological data (e.g. initial and boundary conditions, smoothness of observations). As an example, when data exhibit homogeneous conditions (both initial and boundary conditions equal to zero), we may recommend the use of GP-mRNA rather than GP-Protein as these conditions are explicitly encoded in the structure of the GP-mRNA model.

Both frameworks discussed in this paper could be improved in different ways. First, it is known that some regulatory processes exhibit delays during the translation step. We consider that taking into account those delays would be interesting as a future work (see, e.g. Honkela et al., 2015). Second, the GP-Protein model has been introduced for

single-input single-output schemes. Other possible potential future work is to derive the GP-Protein framework with multiple mRNA profiles (driving-forces) and multiple gap proteins (outputs). Finally, one may consider accounting for positiveness constraints into the GP framework using finite-dimensional Gaussian approximations (see, e.g., Maatouk and Bay, 2017; López-Lopera et al., 2018).

A Single-input single-output GP-Gene model

Let the reaction-diffusion model in (1). Next, we compute the covariance function for the driving-force $k_{u,u}$, and the cross-covariance function between the output and the driving-force $k_{y,u}$

Covariance function for the driving-force

For the computation of the covariance function of the driving-force, we assume that y is a zero-mean GP with covariance function given by Equation (4). Then, the mRNA expression u is also a zero-mean GP with covariance function $k_{u,u}(x,t,x',t') = \text{cov}\{u(x,t), u(x',t')\}$ given by

$$\begin{aligned}
k_{u,u}(x,t,x',t') =& \mathbb{E}\left\{\frac{1}{S}\left[\frac{\partial y(x,t)}{\partial t} + \lambda y(x,t) - D\frac{\partial^2 y(x,t)}{\partial x^2}\right] \times \frac{1}{S}\left[\frac{\partial y(x',t')}{\partial t'} + \lambda y(x',t') - D\frac{\partial^2 y(x',t')}{\partial x'^2}\right]\right\} \\
=& \frac{1}{S^2}[D^2 k^{xxx'x'}(x,t,x',t') - Dk^{xxt'}(x,t,x',t') - Dk^{x'x't}(x,x',t,x',t') - D\lambda k^{xx}(x,t,x',t') \\
& - \lambda D k^{x'x'}(x,t,x',t') + k^{tt'}(x,t,x',t') + \lambda k^{t}(x,t,x',t') + \lambda k^{t'}(x,t,x',t') + \lambda^2 k(x,t,x',t')],
\end{aligned}$$

where k^x is the derivative of k w.r.t. the space variable x, and k^{xt} is the derivative of k^x w.r.t. the time variable t. The other derivatives follow the same structure. Due to the symmetry of the derivatives of the SE kernel, then we obtain

$$k_{u,u}(x,t,x',t') = \frac{1}{S^2}\left[D^2 k^{xxxx}(x,t,x',t') - 2D\lambda k^{xx}(x,t,x',t') - k^{tt}(x,t,x',t') + \lambda^2 k(x,t,x',t')\right],$$

with the derivatives of $k(x,t,x',t')$ w.r.t. x and t given by

$$\begin{aligned}
k^t(x,x',t,t') &= \sigma^2 k(x,x')k^i(t,t'), \\
k^{tt}(x,x',t,t') &= \sigma^2 k(x,x')k^{ii}(t,t'), \\
k^{xx}(x,x',t,t') &= \sigma^2 k^{ii}(x,x')k(t,t'), \\
k^{xxt}(x,x',t,t') &= \sigma^2 k^{ii}(x,x')k^i(t,t'), \\
k^{xxxx}(x,x',t,t') &= \sigma^2 k^{iv}(x,x')k(t,t').
\end{aligned}$$

The derivatives of the SE kernel function are given in (10).

Covariance function between the driving-force and the output

The covariance function between the output y and the force u, $k_{y,u}(x,t,x',t') = \text{cov}\{y(x,t), u(x',t')\}$, is given by

$$\begin{aligned}
k_{y,u}(x,t,x',t') &= \frac{1}{S}\mathbb{E}\left\{y(x,t)\left[\frac{\partial y(x',t')}{\partial t'} + \lambda y(x',t') - D\frac{\partial^2 y(x',t')}{\partial x'^2}\right]\right\} \\
&= \frac{1}{S}\left[\lambda k(x,t,x',t') - k^t(x,t,x',t') - Dk^{xx}(x,t,x',t')\right].
\end{aligned}$$

Acknowledgements

This work was funded by the project "Probabilistic spatio-temporal models based on partial differential equations for the description of the regulatory dynamics for the Bicoid protein in the Drosophila Melanogaster body segmentation" (by Colciencias, Colombia and ECOS-NORD, France) with grant number C15M02. MAA has been partially financed by the EPSRC Research Projects EP/N014162/1 and EP/R034303/1.

References

Abramowitz, M. and Stegun, I. A., editors (1965). *Handbook of Mathematical Functions: With Formulas, Graphs and Mathematical Tables*. Dover Publications Inc., New York, NY, (USA), new edition edition.

Agudelo-España, D., Álvarez, M. A., and Orozco, Á. A. (2017). Definition and composition of motor primitives using latent force models and hidden Markov models. In Beltrán-Castañón, C., Nyström, I., and Famili, F., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 249–256, Cham. Springer International Publishing.

Alon, U. (2006). *An Introduction to Systems Biology*. Chapman and Hall, London.

Alvarado, P. A., Álvarez, M. A., Daza-Santacoloma, G., Orozco, A. A., and Castellanos-Domínguez, G. (2014). A latent force model for describing electric propagation in deep brain stimulation: a simulation study. In *Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE*, pages 2617–2620.

Álvarez, M. A., Luengo, D., and Lawrence, N. D. (2013). Linear latent force models using Gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2693–2705.

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25+.

Barrera, L. A., Vedenko, A., Kurland, J. V., Rogers, J. M., Gisselbrecht, S. S., Rossin, E. J., Woodard, J., Mariani, L., Kock, K. H., Inukai, S., Siggers, T., Shokri, L., Gordân, R., Sahni, N., Cotsapas, C., Hao, T., Yi, S., Kellis, M., Daly, M. J., Vidal, M., Hill, D. E., and Bulyk, M. L. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, 351(6280):1450–1454.

Becker, K., Balsa-Canto, E., Cicin-Sain, D., Hoermann, A., Janssens, H., Banga, J. R., and Jaeger, J. (2013). Reverse-engineering post-transcriptional regulation of gap genes in Drosophila melanogaster. *PLoS Comput Biol*, 9(10):1–16.

Borchers, H. W. (2012). Pracma: practical numerical math functions. https://cran.r-project.org/web/packages/pracma/index.html. [Online; 30-Jan-2018].

Croix, J.-C., Durrande, N., and Álvarez, M. (2018). Bayesian inversion of a diffusion evolution equation with application to biology. *ArXiv e-prints*.

Dalessi, S., Neves, A., and Bergmann, S. (2012). Modeling morphogen gradient formation from arbitrary realistically shaped sources. *Journal of Theoretical Biology*, 294:130 – 138.

Deville, Y., Ginsbourger, D., and Roustant, O. (2015). kergp: Gaussian process models with customised covariance kernels. https://cran.r-project.org/web/packages/kergp/index.html. [Online; 23-Dec-2015].

Dilão, R. and Muraro, D. (2010). mRNA diffusion explains protein gradients in Drosophila early development. *Journal of Theoretical Biology*, 264(3):847 – 853.

Dupuy, D., Helbert, C., and Franco, J. (2015). DiceDesign and DiceEval: Two R packages for design and analysis of computer experiments. *Journal of Statistical Software*, 65(i11).

Erban, R., Chapman, J., and Maini, P. (2007). A practical guide to stochastic simulations of reaction-diffusion processes. *ArXiv e-prints*.

Forgacs, G. and Newman, S. A. (2005). *Biological Physics of the Developing Embryo*. Cambridge University Press.

Gao, P., Honkela, A., Rattray, M., and Lawrence, N. D. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(10):170–175.

Guarnizo, C. and Álvarez, M. A. (2018). Fast Kernel Approximations for Latent Force Models and Convolved Multiple-Output Gaussian processes. *ArXiv e-prints*.

Honkela, A., Peltonen, J., Topa, H., Charapitsa, I., Matarese, F., Grote, K., Stunnenberg, H. G., Reid, G., Lawrence, N. D., and Rattray, M. (2015). Genome-wide modeling of transcription kinetics reveals patterns of rna production delays. *Proceedings of the National Academy of Sciences*, 112(42):13115–13120.

Hsiao, Y. T., Lee, W. P., Yang, W., Müller, S., Flamm, C., Hofacker, I., and Kügler, P. (2016). Practical guidelines for incorporating knowledge-based and data-driven strategies into the inference of gene regulatory networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(1):64–75.

Jaeger, J., Manu, and Reinitz, J. (2012). Drosophila blastoderm patterning. *Current Opinion in Genetics & Development*, 22(6):533 – 541.

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 785–792. MIT Press.

Lipniacki, T., Paszek, P., Marciniak-Czochra, A., Brasier, A. R., and Kimmel, M. (2006). Transcriptional stochasticity in gene expression. *Journal of Theoretical Biology*, 238(2):348 – 367.

Liu, W. and Niranjan, M. (2012). Gaussian process modelling for bicoid mRNA regulation in spatio-temporal Bicoid profile. *Bioinformatics*, 28(3):366–372.

López-Lopera, A. F. and Álvarez, M. A. (2019). Switched latent force models for reverse-engineering transcriptional regulation in gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1):322–335.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255.

Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582.

Meinhardt, H. (2015). Models for patterning primary embryonic body axes: The role of space and time. *Seminars in Cell & Developmental Biology*, 42:103 – 117.

Meyer, P. (1970). *Introductory Probability and Statistical Applications*. Addison-Wesley, 2 edition.

Nason, G. (2012). NORMT3: evaluates complex erf, erfc, Faddeeva, and density of sum of Gaussian and Student's t. https://cran.r-project.org/web/packages/NORMT3/index.html. [Online; 31-Oct-2012].

Polyanin, A. (2001). *Handbook of Linear Partial Differential Equations for Engineers and Scientists*. CRC Press.

Poppe, G. P. M. and Wijers, C. M. J. (1990). More efficient computation of the complex error function. *ACM Trans. Math. Softw.*, 16(1):38–46.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rogers, W. A., Grover, S., Stringer, S. J., Parks, J., Rebeiz, M., and Williams, T. M. (2014). A survey of the trans-regulatory landscape for Drosophila melanogaster abdominal pigmentation. *Developmental Biology*, 385(2):417 – 432.

Stakgold, I. and Holst, M. J. (2011). *Green's functions and boundary value problems*. Pure and applied mathematics. Wiley.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York.

Surkova, S., Golubkova, E., Manu, Panok, L., Mamon, L., Reinitz, J., and Samsonova, M. (2013). Quantitative dynamics and increased variability of segmentation gene expression in the Drosophila Krüppel and knirps mutants. *Developmental Biology*, 376(1):99 – 112.

Vanhatalo, J. and Vehtari, A. (2007). Sparse log Gaussian processes via MCMC for spatial epidemiology. In Lawrence, N. D., Schwaighofer, A., and Candela, J. Q., editors, *Gaussian Processes in Practice*, volume 1 of *Proceedings of Machine Learning Research*, pages 73–89, Bletchley Park, UK. PMLR.

Vásquez Jaramillo, J. D., Álvarez, M. A., and Orozco, A. A. (2014). Latent force models for describing transcriptional regulation processes in the embryo development problem for the Drosophila melanogaster. In *Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE*, pages 338–341.

Weideman, J. (1994). Computation of the complex error function. *SIAM Journal on Numerical Analysis*, 31(5):1497–1518.

Wilson, M. J., Havler, M., and Dearden, P. K. (2010). Giant, Krüppel, and caudal act as gap genes with extensive roles in patterning the honeybee embryo. *Developmental Biology*, 339(1):200 – 211.

Yagita, K. (2018). Dual-scope of circadian rhythm biology. *Sleep and Biological Rhythms*, 16(1):1–2.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.