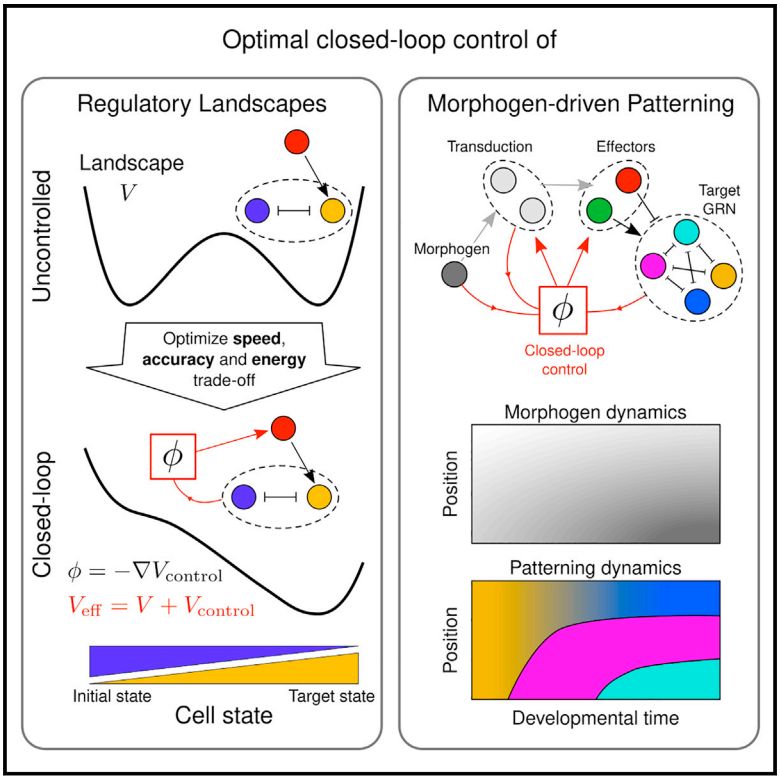
Cell Systems

Article

Optimal control of gene regulatory networks for morphogen-driven tissue patterning

Graphical abstract



Highlights

- Morphogen signaling controls the pattern of gene expression in developing tissues
- Optimal control theory identifies signaling mechanisms for morphogen patterning
- By incorporating feedback between signaling and gene regulation, it explains dynamics
- Provides an alternative framework to the "French Flag model" for morphogen patterning

Authors

Alberto Pezzotta, James Briscoe

Correspondence

a.pezzotta@ucl.ac.uk (A.P.),
james.briscoe@crick.ac.uk (J.B.)

In brief

An approach based on optimal control theory, using reinforcement learning to identify signaling dynamics that guide cells to specific fates in a timely, precise, and reproducible manner, offers insights into how morphogen-controlled gene regulatory networks pattern tissues and establishes an alternative framework to the French Flag model.

Pezzotta & Briscoe, 2023, Cell Systems *14*, 940–952
November 15, 2023
https://doi.org/10.1016/j.cels.2023.10.004

CellPress

CellPress
OPEN ACCESS

Cell Systems

Article

Optimal control of gene regulatory networks for morphogen-driven tissue patterning

Alberto Pezzotta[1,2,3,*] and James Briscoe[1,*]
[1]Developmental Dynamics Laboratory, The Francis Crick Institute, 1 Midland Road, NW1 1AT London, UK
[2]Gatsby Computational Neuroscience Unit, University College London, 25 Howland Street, W1T 4JG London, UK
[3]Lead contact
*Correspondence: a.pezzotta@ucl.ac.uk (A.P.), james.briscoe@crick.ac.uk (J.B.)
https://doi.org/10.1016/j.cels.2023.10.004

SUMMARY

The generation of distinct cell types in developing tissues depends on establishing spatial patterns of gene expression. Often, this is directed by spatially graded chemical signals—known as morphogens. In the "French Flag model," morphogen concentration instructs cells to acquire specific fates. How this mechanism produces timely and organized cell-fate decisions, despite the presence of changing morphogen levels, molecular noise, and individual variability, is unclear. Moreover, feedback is present at various levels in developing tissues, breaking the link between morphogen concentration, signaling activity, and position. Here, we develop an alternative framework using optimal control theory to tackle the problem of morphogen-driven patterning: intracellular signaling is derived as the control strategy that guides cells to the correct fate while minimizing a combination of signaling levels and time. This approach recovers experimentally observed properties of patterning strategies and offers insight into design principles that produce timely, precise, and reproducible morphogen patterning.

INTRODUCTION

Embryogenesis depends on positioning functionally distinct types of cells in the right place, at the right time in a developing tissue. In many cases, this is guided by chemical signals (usually termed morphogens). Emanating from a localized source, a morphogen spreads across a field of cells to form a gradient; hence, cells at different positions are exposed to different levels of the morphogen.[1] In the influential "French Flag model," cells are proposed to read the gradient, such that the local signal concentration instructs position-dependent cell fate.[2] It has become apparent, however, that morphogen concentration alone is insufficient to explain the interpretation of morphogen gradients. In many tissues, morphogen gradients are dynamic, and there is no simple relationship between morphogen concentration and position within the tissue.[3,4] It is also unclear how a simple gradient mechanism would allow timely and accurate cell-fate decisions, despite the presence of molecular noise and individual variability.

The interpretation of morphogen signals involves gene regulatory networks (GRNs) in responding cells.[4] These comprise the intracellular signaling pathways of the morphogens and the downstream transcriptional responses and are central to transforming the continuous spatiotemporal input of morphogen signaling into discrete cell fates. Regulatory interactions between components of these networks appear to perform the equivalent of an analog-to-digital conversion.[4–7] GRNs have also been proposed to contribute to the accuracy and reproducibility of patterning in the presence of intracellular noise.[8–10] Moreover, non-linearities and feedback within the GRN can confer multi-stability, memory, and hysteresis to cellular decision-making. A consequence of this is that cell fate depends not only on the levels of signals and effectors but also on their temporal features. Taken together, the complexity of interactions within the GRN can produce rich dynamics in the signaling and gene expression in developing tissues. Understanding the origin and function of these dynamics offers insights into patterning. Moreover, the interplay between morphogen gradient and GRN allows cells to actively contribute to morphogen signaling, rather than being simply "instructed" by the gradient. This highlights the need for alternative paradigms to the French Flag model, in which the GRN plays a complementary and equally important role to the morphogen, to frame questions about morphogen activity.

The dorso-ventral patterning of the developing vertebrate neural tube is a well-established example of a morphogen-patterned tissue.[4,11] In the ventral neural tube, the secreted morphogen Sonic hedgehog (Shh)—produced from the notochord and floor plate, which are located at the ventral pole—forms a ventral to dorsal gradient.[12] Binding of Shh to its receptor Patched1 (Ptch1) releases the inhibition of downstream signaling and leads to the conversion of the transcriptional effectors—the Gli family of proteins—from their repressor to their activator forms. The Gli proteins regulate the expression of a set of transcription factors, which include members of the Nkx, Olig, Pax, and Irx families. This comprises the neural tube GRN. Interactions between

intracellular signaling and the transcriptional network generate a dynamic response of Gli activity to varying amounts of Shh and produce a sequence of genetic toggle switches that generate distinct gene expression states over time.[3,13] Feedback leads to the desensitization of cells to the morphogen signal,[12,14–16] resulting in adaption in Gli activity.[16] Similar effects of negative feedback have been observed for many signaling pathways, but its function and implications for morphogen-dependent pattern formation remain unclear.

Dynamical systems theory provides a framework to describe the activity of morphogens and GRNs. The behavior produced by such models can often be represented geometrically as a dynamical landscape. This provides an intuitive description of cell-fate decisions that correspond to the idea of an ''epigenetic landscape'' proposed by Waddington.[17] In this view, the developmental trajectory of a cell is analogous to a particle rolling on an undulating landscape, where valleys and watersheds represent fates and decision points, respectively. Morphogens can be thought of as tilting the landscape in such a way that the valleys can be made deeper, shallower, or disappear altogether. In this way, the morphogen controls the terrain and hence the valley a cell enters. Although originally introduced as a pictorial representation of development, this idea has been used to develop quantitative methods that reproduce key features of GRNs and make predictions about the effect of signals.[18–20] Nevertheless, it remains a challenge to construct landscape models that incorporate the knowledge of signals and GRNs. How is the landscape modified by an external signal and how are feedback mechanisms incorporated? How can experimentally inferred landscapes give insights into the signaling dynamics?

Here, we set out to develop a framework to understand the intracellular signaling strategies used by cells to interpret a morphogen signal. Are there design principles to the signaling pathways that contribute to timely, precise, and accurate morphogen-controlled tissue patterning? What role does feedback play and does this result in a trade-off between speed, accuracy, and robustness of the pattern formation? To this end, we adopt an optimization approach, with the aim of discovering strategies that underlie the solution of an information-processing task, namely morphogen-driven patterning, by maximizing the performance. Specifically, we cast the morphogen-driven patterning process as an optimal control problem, where a trade-off is sought that minimizes the distance from a specified target and the amount of signal given. One advantage of this approach is that it can easily interface with optimization techniques such as reinforcement learning (RL).[21] An additional feature that makes this strategy particularly suited for the study of morphogen-driven patterning is that it allows the activity of signaling effectors to be a generic function of both extracellular signal and target genes within the GRN. Thus, it provides a way in which the function of feedback within the signaling pathway and from the GRN to the signaling pathway can be investigated.

We first applied this approach to a Waddington-landscape model representing a genetic toggle switch—where analytical treatment is possible. We then extended the analysis to a dynamical-system model describing gene regulation in ventral neural tube progenitors. We show that desensitization of the signaling pathway to morphogen emerges as a means to minimize control inputs in the context of multi-stability. The approach discovers morphogen-patterning strategies that are widely used in biological systems and suggests an explanation for these strategies. Using this optimal control framework places morphogens and GRNs on the same footing, each playing complementary roles as parts of a whole decision-making unit. In this sense, the approach provides an alternative framework to the French Flag paradigm.

RESULTS

A case for optimal control—computational vs. reductionist modeling

The interpretation of morphogen by cells in developing tissues can be regarded as an example of information processing in a biological system. Neuroscientists David Marr and Tomaso Poggio pointed out how ''understanding'' information-processing systems can be (and has to be) achieved at relatively independent levels,[22] introducing a framework that is often referred to as Marr's levels of analysis. This framework, made explicit for the visual system,[23] has since become widely accepted as a paradigm in understanding general cognitive processes—and information-processing in biology more generally.[24] The distinct levels pertain to the tasks performed, i.e., the problems solved by the system under consideration (computational level); the rules and representations used in solving these tasks, which can be described by input-output mathematical laws (algorithmic level); and, finally, the physical realization of those laws, i.e., the hardware and physical processes used to solve the computational problem (implementation level).

These levels are only relatively independent from one another. Questions can be formulated at each one of them, but understanding aspects at one level can shed light on aspects at another. An understanding of a system means connecting these different levels of analysis. Mathematical modeling constitutes reductionist attempts to step from the implementation level to that of the algorithms: what kind of logic can be obtained given a set of components, their behavior in isolation and their possible interactions? Due to the emergent complexity of the whole system, this often requires a level of abstraction such that specific details of the implementation are approximated, averaged in time and space, or even neglected, such that the description at the algorithmic level becomes relatively independent of them. The modeling framework in Cohen et al.[13] and Bintu et al.[25] are examples of such attempts (see STAR Methods section ventral neural-progenitor GRN model (PONI network)). However, capturing the qualitative features of the laws at the level of the algorithms from the ground up in this way is often so difficult that it becomes practically impossible.[26] Even in those instances where this can be successfully carried out, it could be argued that the laws at the algorithm level are qualitatively so different from those at the implementation level that they acquire relative independence from them and that a whole different description is more useful for their understanding. This is well understood in physics[27,28] and applies, to an even greater extent, to the study of biological systems[29] and beyond.[30] A fruitful approach to ''understand'' behavior is one that moves from the computational (pertaining to the goals of the system of interest) toward

the implementation level, rather than vice versa, from the top down, instead of from the bottom up.[26] This does not invalidate bottom-up, reductionist modeling, which remains a complementary approach and becomes more useful the more developed the understanding of a system's parts and their laws.

When a measure of performance associated with a computational problem can be formulated, the search for the strategies that maximize performance provides a means to discover the possible rules that can be used to solve the particular problem. This approach can draw from the well-grounded mathematical theory of optimization and a plethora of numerical techniques. Optimal control is a class of mathematical problems that can provide a framework for this approach, as it represents part of the broader theory of decision-making. Cell fate decision-making and the interpretation of morphogen signaling in developing tissues appear to be a suitable problem to which tools from optimal control can be applied. The goal is to identify and understand the rules (algorithms) cells use to solve the problem of (compute) pattern formation.

Dynamical systems and optimal control approach to cell-fate decisions

In this section, we describe how decision-making in the context of a cell or a tissue can be framed as an optimal control problem. The dynamics of gene regulation and cell-fate decisions can be described using a Langevin equation

$$\frac{dx}{dt} = f(x,u) + \sigma(x,u)\eta \qquad \text{(Equation 1)}$$

where x is the set of concentrations of the components of the network, and u is a set of inputs or control variables. The functions f and σ are the drift and the strength of the noise, respectively (η is a standard Gaussian white noise). In general, the noise term has a multiplicative form, i.e., it depends on the cell state x. This is the case when stochasticity arises not only from external disturbances but also from the finite copy number of each species in the network.[31–33] The drift and noise functions f and σ can incorporate mechanistic knowledge of the regulatory logic of the network and the effect of morphogen signaling, for instance, transcriptional control via binding/unbinding of transcription factors to their respective regulatory elements and cooperative and competitive effects.[13,25] Multiplicative noise can have very profound effects on the dynamics, altering the stability of the system substantially, and has to be dealt with carefully.[34] For instance, from the point of view of its analytical treatment, multiplicative noise poses the issue of the convention (notably, Itô or Stratonovich) used to solve the Langevin equation, as detailed in Coomer et al.[34] However, when Equation 1 is derived from more detailed microscopic models the choice of the convention is a natural one. When multiplicative noise is considered in this work, it is derived within the chemical Langevin equation framework,[31] which prescribes the Itô convention (see also STAR Methods section ventral neural-progenitor GRN model (PONI network)).

The dynamical systems that result from representing GRNs in this way are generally non-linear and may operate in multistable regimes. The input u can substantially change the dynamics of the network, altering the position of attractors (stable states) and saddle nodes (decision points). Moreover, the attractor reached by a system depends on the full history of the inputs. This can be seen, for example, in the neural progenitor GRN,[13] where the input u comprises the activating and repressing forms of the Shh morphogen-regulated Gli effectors (Figure 1).

The behavior of such systems can be visualized as a dynamical landscape with valleys representing the stable states of the network and signals tilting the landscape to determine which valleys are accessible or inaccessible. The dynamical-system function f is thus given by the gradient of the landscape, V, parametrically dependent on the effector u. This approach has been used to reproduce the qualitative features of GRNs as well as to predict patterning processes in embryos[18,19] and proportions of cell types in differentiation protocols.[20]

Given this dynamical system's view of patterning, how does the signaling input to a GRN generate a sufficiently precise pattern in a developmentally relevant time period? To address this, we recast patterning as an optimization problem and ask what sort of signal input is necessary to produce precise, reliable, and timely cell-fate decisions. The framework that naturally deals with these types of problems is optimal control theory. We are faced with the task of choosing a dynamic signaling regime u (here referred to as control) that minimizes the average of a cost accumulated along the trajectory (referred to as running cost), plus a cost determined by the distance from the target at the termination of the decision task (hence called terminal cost) occurring upon a differentiation event. The running cost quantifies how much, during the task, gene expression deviates from the target and how much control is exerted (respectively, through a function q, minimum at the target state, and through a term quadratic in u, weighted by a trade-off parameter ϵ). The terminal cost also measures the distance from the target but is evaluated only at the end of the time window through a function Q. If the random termination of the task occurs with a homogeneous probability rate τ^{-1}, the cost function so defined can be written as

$$C = \int_0^\infty dt\, e^{-t/\tau}\left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t)) + \tau^{-1}Q(x(t))\right). \qquad \text{(Equation 2)}$$

From the point of view of decision-making, the constant rate of differentiation assigns more weight to more imminent events while discounting those further in the future (see STAR Methods section exponential discounting). Notice that, in Equation 2, the terminal target-state cost can be formally interpreted as a running cost, as a result of exponential discounting (see STAR Methods section terminal cost and discounting for details), so that q and Q can be used interchangeably up to a τ scale factor. Additionally, τ features as a trade-off parameter setting a relative weight between terminal and running costs. In this work, we chose q and Q to be the square distance between the state and the target (see STAR Methods section optimal control in a potential). In general, any convex function that has a minimum at the target state will lead to qualitatively similar results, upon appropriate choice of the trade-off parameters. It is also important to note that there is no mechanistic interpretation of the target cost function: as discussed in the

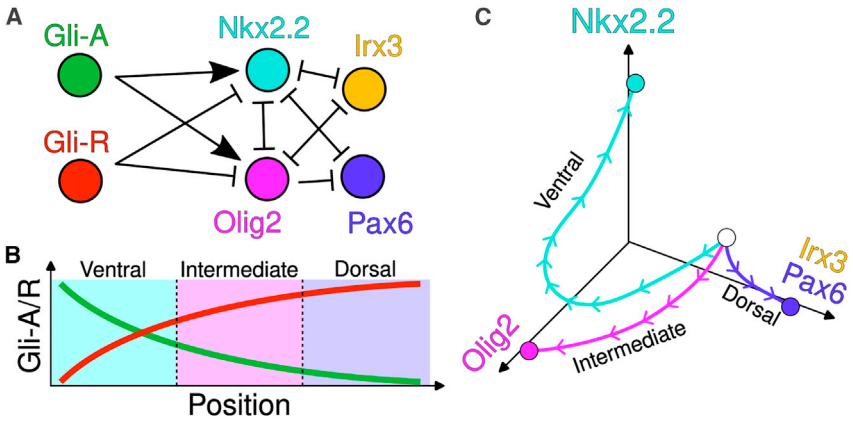


Figure 1. External input changes the stability properties of the dynamical system

(A) We consider a model of gene regulation that describes the patterning dynamics in the ventral neural tube with the addition of intrinsic noise,[8,13] whereby the Gli effectors (activator green and repressor red) control differential expression of downstream genes interacting in a network.

(B and C) Cartoons showing the qualitative spatio-temporal behavior of such networks. (B) When exposed to a static gradient of activator/repressor, starting from identical initial conditions, cells at different positions in the tissue attain different steady-state expression, and a pattern is established over time, with three qualitatively distinct regions, labeled "ventral" (high Nkx2.2, cyan), "intermediate" (high Olig2, magenta), and "dorsal" (high Pax6 and Irx3, co-expressed, blue). (C) In the framework of dynamical systems, we understand this as the effect of a control parameter changing the stability properties of the system, whereby the qualitatively different steady states correspond to distinct point attractors. The parameters used in this model throughout the manuscripts are reported in STAR Methods, see the table in section ventral neural-progenitor GRN model (PONI network).

previous section, the cost function is a quantification of performance that lives at the computational level and has, in principle, no connection with the molecular mechanisms involved—other than the fact that it is expressed in terms of the genes of interest. By contrast, the cost for control could be given a mechanistic interpretation as a proxy for an energetic cost. For Markov processes, stochastic thermodynamics functions such as work and entropy production can be expressed as an additive functional similar to the running control cost introduced here.[35] However, this form is not valid for non-Markov processes, except in particular limiting cases.[36,37] Here, the particular form chosen does not quantify the energetics of the production/regulation of the signaling effectors mechanistically, which would require a detailed knowledge of the physical processes involved in it. Instead, despite the simplistic Markovian assumption that is convenient for the mathematical and computational analysis, it retains the functional significance as the (evolutionary) pressure to reduce effort/energy expenditure during the developmental process.

When the system to be controlled is completely deterministic, i.e., $\sigma = 0$ in Equation 1, given the initial conditions we can predict the optimal trajectory $x^*(t)$, and an expression of optimal control u^* can be given in the open-loop form, i.e., as a function explicitly dependent on time only. An example of open-loop control of a toggle-switch GRN is depicted in Figure 2A. However, open-loop control strategies are generally suboptimal when the system is stochastic. In such a case, u^* cannot be planned in advance, since fluctuations in the state of the system need to be constantly monitored by the controller in order to behave optimally. The minimization of a cost function in the form of Equation 2 naturally yields the optimal control in the closed-loop (feedback) form, $u^*(t) = \varphi^*(x^*(t))$, i.e., the dependence on time is through the state of the system (see STAR Methods section optimal control in a potential). For the toggle-switch GRN, this case is depicted in Figure 2C.

Although the open-loop picture can be used for modeling scenarios such as optogenetic driving or differentiation protocols in a dish, involving pre-defined signaling protocols, the closed-loop picture is particularly relevant in the context of the control of gene expression in a cell, where aspects of the signal transduction pathway and the signal effector can be under the control of the transcription factors in the GRN. Solving for the optimal control u^* yields optimal feedback designs and can shed light on the functional role of observed feedback mechanisms.

In general, the optimality equations are difficult to solve exactly, but approximate solutions can be found via techniques such as RL.[21] Although deterministic control problems—which admit open-loop optimal solutions—are special cases of more general stochastic control problems, in practice, their solution is generally approached by very different techniques.[38,39] Only in some limiting cases can the optimality equations be solved analytically or numerically. In the next section, we analyze in detail a simple model of a toggle-switch that lends itself to analytical treatment and that can be accurately solved with numerical methods. Despite its simplicity, this example is useful in illustrating the framework and provides a guide for understanding the parameters of the cost function.

Controlling the epigenetic landscape of a genetic switch

Here, we consider a simple model for a binary cell-fate decision, which can be implemented via a toggle-switch network with two mutually repressing genes integrating an external activator (Figure 2A), as in Perez-Carrasco et al.[40] A landscape model for this system is represented by one-dimensional double-well potential $V(x)$ with minima at, e.g., ± 1, which corresponds to two possible cells fates (see STAR Methods section optimal control in a potential). In this example, the noise is modeled as an additive and independent of control, so that σ in Equation 1 is given by $\sqrt{2D}$, with constant D. We model morphogen signaling as an additive drift contribution u, which "tilts" the landscape, $V(x, u) = V(x) - u \cdot x$, from which it follows that in Equation 1, $f = -\nabla V + u$ (Figure 2B). Note that here the noise is assumed to depend neither on the state x nor on the control u: this is somewhat limiting with respect to the general case in Equation 1. Also, the additivity of the control is somewhat artificial and would generally not apply if we were to "derive" a landscape model from a microscopic model of gene regulation. For

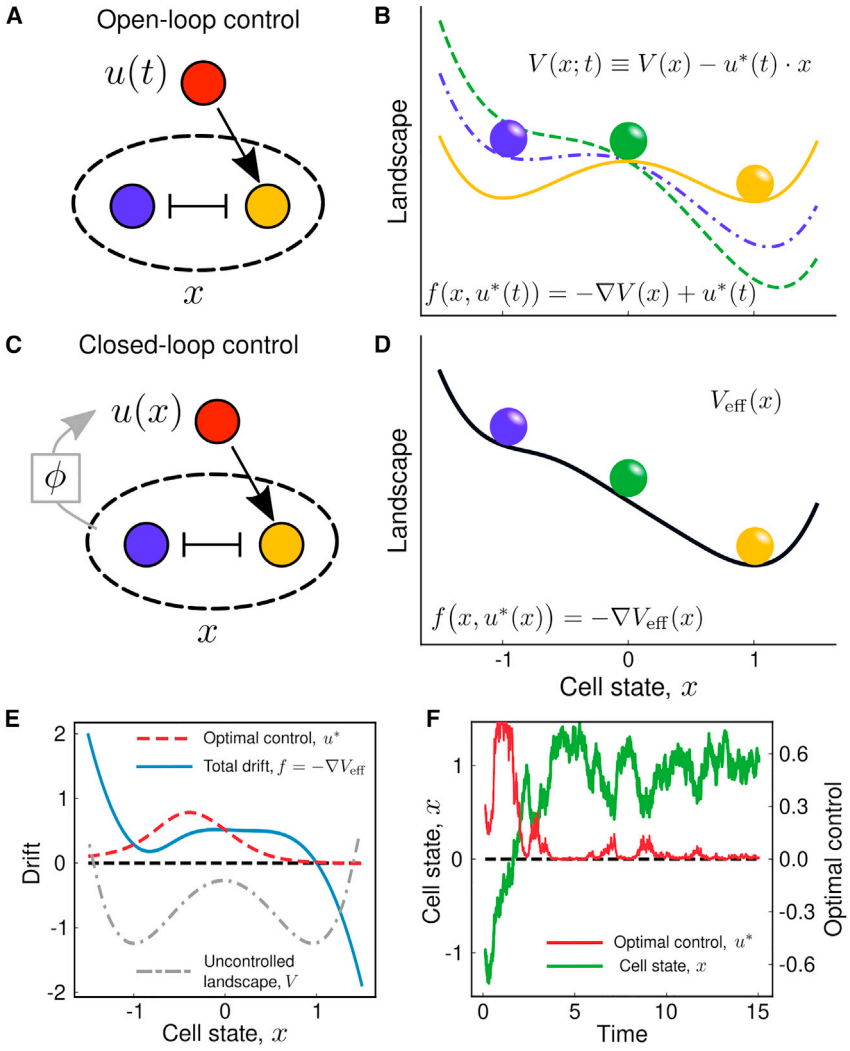


Figure 2. Optimal control representation of a Waddington landscape

(A) A GRN for a simple toggle-switch network with two genes can be dynamically controlled to reach a target state by explicitly defining a signaling protocol $u(t)$ (open-loop control).

(B) In the Waddington-landscape picture, we can think of the external control as "tilting" the landscape over time; the colored lines represent the instantaneous landscape felt by the "particle" of the same color.

(C) Alternatively, the signal can be placed under control of the target genes through a feedback function φ. This results in closed-loop, or feedback, control.

(D) The optimal closed-loop control is incorporated into a "static" effective landscape, describing the dynamical properties of the signaling and GRN system as a whole.

(E) The solution for the optimal control (dashed red line) exhibits adaptation near the target, when this corresponds to a stable fixed point of the uncontrolled landscape (dashed-dotted gray line, not in scale).

(F) This can also be seen in a sample trajectory of the dynamics of a cell (green line), where the control (red line) is switched off after an initial transient and is activated only to prevent large fluctuations away from the target. For (E) and (F), the parameters used are $D = 0.10$, $\tau = 10$, and $\epsilon = 10$.

instance, these modeling assumptions do not hold for the model depicted in Figure 1 and are considered in the next section. However, they allow the problem to be analytically tractable, while still being able to shed light on general principles of control, in addition to being easily interpretable geometrically.[18,19] We note that for the Langevin equation in this example, the quadratic cost for control has an information-theoretical interpretation as a Kullback-Leibler divergence, which makes it a convenient regularization term in machine-learning applications (also STAR Methods section optimal control in a potential).

We then seek a control protocol u (the dynamics of signal) that drives a cell from state $x = -1$ to the state $x = 1$ in the optimal way, i.e., minimizing the combination of how far the cell is from its target and the amount of control exerted to accomplish this (Equations 12 and 26 in STAR Methods). As highlighted in the previous section, the optimal control strategy in the presence of noise is naturally expressed in the closed-loop form (Figure 2C). Due to the specific form of the dynamics and the control cost (dynamical system linear in u and quadratic cost for control), the deterministic part of the optimally controlled dynamics can be expressed analytically as the negative gradient of a landscape function V_{eff}, i.e., $f = -\nabla V_{\text{eff}}$. This equals the original landscape function V plus the optimal cost expected to be paid from a given state x, i.e., the optimal cost-to-go function:

$$V_{\text{eff}}(x) = V(x) + \min_u \mathbb{E}[C \mid x_0 = x] \qquad \text{(Equation 3)}$$

where C is given by Equation 2 and should be regarded as a function of the trajectory $x(t)$ in state space and the control function u evaluated along the trajectory (see STAR Methods section optimal control in a potential for details). The additional term is also referred to as (minus) the value function—that is its more customary name in the context of decision-making. The optimally controlled dynamics take the form

$$\frac{dx}{dt} = -\nabla V_{\text{eff}}(x) + \sqrt{2D}\,\eta \qquad \text{(Equation 4)}$$

Thus, rather than thinking of the control as tilting the landscape over time, it can be incorporated into a new landscape that describes the system as a whole (Figure 2D). Such a landscape can be inferred by measuring the steady-state distribution of cell state which, according to the Langevin dynamics in Equation 4, is related to V_{eff} via

$$\rho_{\text{ss}}(x) = Z^{-1} \exp\left(-V_{\text{eff}}(x)/D\right) \qquad \text{(Equation 5)}$$

where $Z = \int dx \exp\left(-V_{\text{eff}}(x)/D\right)$ is a normalization constant. This observation suggests that the inverse problem might provide insights into the function of

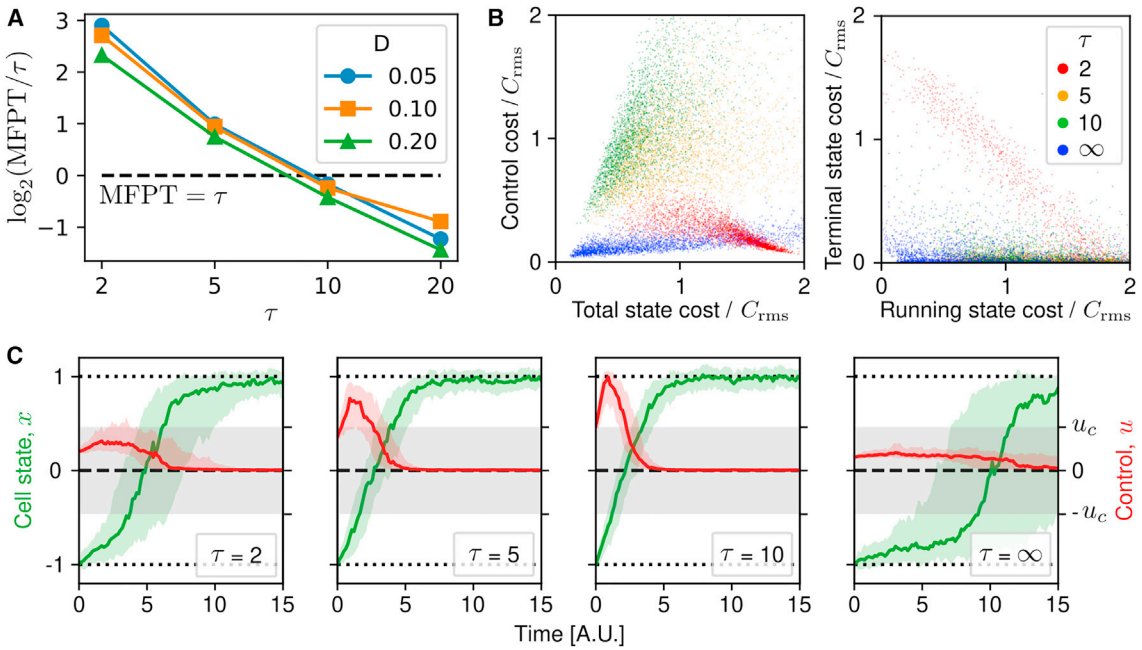


Figure 3. Effect of the discounting (differentiation) time τ

(A) The mean first passage time (MFPT) at the target $x = 1$ from $x_0 = -1$ as a function of τ, from the numerical integral of the analytical formula, under the optimal control. This is shown relative to the value of τ on a logarithmic scale. For high (low) values of τ, the MFPT for the optimally controlled dynamics is far lower (higher) than τ itself.

(B) State and control costs from 5,000 simulations for various values of τ (color coded). The optimal control for "small" or "large" values of τ effectively minimizes cost for control, whereas for intermediate values of τ, a non-trivial trade-off is observed (left). Only for low values of $\tau \simeq 1$ does the terminal cost for the distance from the target have a large contribution to the overall cost (right).

(C) Statistics of 100 samples of the dynamics for the state (green) and the control (red). Solid lines are the median values, shaded areas the 25th to 75th percentile. The gray-shaded area highlights the values of the control variable u for which the controlled landscape is still bistable, i.e., between the bifurcation values $\pm u_c$. In all panels, $\epsilon = 10$; in (B) and (C) $D = 0.05$. For intermediate values, when the MFPT is comparable with τ, the switch is driven by a non-trivial transient dynamics for the control, resulting from competition between control and target running costs.

feedback mechanisms in cell-fate decisions: given experimental observations and a landscape associated with the underlying GRN, it might be possible to distinguish the contributions of the controlled system (the GRN) from the feedback mechanisms (Figure 2E).

We notice that the resulting optimal control protocol leads to adaptive dynamics: high levels of control are necessary to leave the initial attractor; then, as the system approaches the target attractor, the amount of control is minimal and only required to prevent noise from reversing the transition (Figures 2F and S1). From this example, we see that the optimal solution minimizes control by taking advantage of the multi-stability built into the system.

This example also provides intuition into the effect of the differentiation rate—equivalently, of discounting cost over time. What is the optimal behavior of the system before a cell differentiates?

At one limit, when the differentiation rate is comparable with the overall timescale of the system, $\tau \lesssim 2$, and the noise, D, is low, only imminent running costs and the terminal costs have a meaningful effect on the total cost, and the optimally controlled dynamical system is bistable. This is because when the system is far from its target, a substantial reduction in the distance of the system from its target within a short time τ would have a very high cost for control. Therefore, the only part of the cost that the controller can minimize is the cost of the control itself. This leads to low values of the control at every state, and the system remains within the bistable regime (Figures 3A, 3C, and S1, bottom left). Such small values of τ would mean that a cell only rarely reaches its target before differentiation.

Very similar dynamics are observed in the opposite limit, when $\tau \gg 1$ (Figures 3A, 3C, and S1, top left). Here, no terminal cost is paid, and the problem consists of optimizing the average cost per unit of time at steady state. For low D, when multiple stable fixed points are present (as in the case of small u—bistable regime), the system spends long periods of time near each of them, with rare stochastic transitions between. In STAR Methods section optimal control in a potential, we demonstrate how the steady-state average of the cost q is exponentially small in u/D, when D is small: this allows very low values of u to yield large discrepancies between the probabilities of being in either attractor at steady state. This explains why, in such a limit, it is optimal to choose u well within the bi-stability regime.

For intermediate values, $5 \lesssim \tau \lesssim 20$, the optimally controlled dynamics are such that the time needed to perform the switch is comparable with τ itself. When this is the case, characteristic transient dynamics are observed: in the first phase, high levels of control are applied to the system in order to drive the transition; in the second phase, the control can be reduced to very

low levels, within the bistable regime. This suggests that, in these scenarios, the optimal strategy is for the controller to apply high levels of control for a short time, resulting in a lower cost from being off target for a shorter period of time (Figures 3A and 3D).

These observations are partially reflected in the distribution of the different terms in the total cost C. For low and high values of τ, the control term of the cost is relatively less important than target-state term of the cost, compared with the intermediate values of τ (Figure 3B, left). Also, the smaller τ is, the shorter the average time given to complete the task and therefore the higher the contribution from the terminal cost (Figure 3B, right). As already seen analytically, τ plays the role of a trade-off parameter between the running and terminal target-state costs, q and Q.

We also quantified the dependency of the distribution of transition times from $x = -1$ to 1 under the optimally controlled dynamics on the noise strength D (Figures 3A and S2). As remarked above, for extreme values of τ, the control strategy is very similar, and the time to target is primarily controlled by noise, with a smaller and smaller average transition time as D increases—as given by theory of stochastic transitions (Figure S2; STAR Methods section optimal control in a potential). Instead, for intermediate values of τ, this distribution has a non-trivial dependency on D, with the mean transition time showing a non-monotonic behavior (see, e.g., $\tau = 20$, in Figures 3A and S2). These effects are subtle, and their rigorous mathematical analysis is outside the scope of this work. However, taken together, these results show that the requirement of optimality in the presence of multi-stability and noise can yield control strategies with counter-intuitive resulting dynamics.

By making use of a simple Waddington landscape model, this example shows how optimal control theory can make sense of adaptation as the least-effort strategy to drive a cell to a desired target while exploiting the multi-stability of a downstream network and its stochastic dynamics. The analytical results suggest an explanation for optimal signaling in the face of varying degrees of noise and multi-stability and for different values of differentiation rates, which set the exponentially distributed time horizon within which cell-fate decision needs to take place.

Control of cell-fate in ventral neural progenitors

We applied this optimal control approach to a GRN model that captures the patterning dynamics in the ventral region of the developing neural tube.[13] In this model, noise from fluctuations in the copy number of components of the system have been introduced using the chemical Langevin equation approximation[8,31] (Figure 1; reported in STAR Methods section ventral neural-progenitor GRN model (PONI network)). The control here is a two-component vector representing the activator and repressor form of the morphogen-controlled Gli effectors. These directly regulate the two most ventral markers, Nkx2.2 and Olig2 (Figure 4A). In this case, we find an approximate solution of the optimal control equations via RL.[21] RL provides the means to identify optimal control strategies, without knowledge of the dynamical-system function f, by sampling states, actions (controls), and running costs (or reward signals). Here, and in the following section, we use the twin delayed deep deterministic (TD3) algorithm,[41] which is a state-of-the-art RL algorithm for continuous control problems (see algorithm 1 in STAR Methods for details). Using this approach, we identify optimal control strategies for the system to adopt an Olig2 state or a Nkx2.2 state.

In all cases, we optimize the discounted cost function, Equation 26, with $\tau \simeq 5$ (arbitrary units, A.U. —see table in ventral neural-progenitor GRN model (PONI network)): this can be compared with the half-life of Nkx2.2 and Olig2, $t_{1/2} \simeq 0.35$. Thus, if $t_{1/2} \simeq 4h$, then $\tau \simeq 2.5$ days, consistent with the developmental timescales in the embryonic mouse neural tube. For both targets, the control input shows a very clear transient, whereby the activator Gli is initially high and then drastically reduces at the steady state.

Acquiring and maintaining the Olig2 state requires a very high sensitivity of control with respect to Olig2 levels, which is reflected in the high variability of the repressive form of Gli effector at the population level (Figures 4D and 4E) and in the learning curves (Figure S4). The learned control is such that below a threshold value of Olig2, Gli repressor is high, and above the threshold, Gli repressor is low (Figure 4E, right). One explanation for this could be that higher levels of repressor are necessary to restrain the system from bifurcating to Nkx2.2 when levels of Olig2 are too low. This is consistent with the experimental evidence that Olig2 may provide negative feedback onto the expression of Gli3, which is the dominant repressor for Shh signaling.[16,42,43]

Different runs of the RL algorithm yield quantitatively different transient dynamics of the signaling effectors; however, the control strategies at steady state are consistent across runs, and the same qualitative features commented on above are maintained even in the transient states (Figure S3). This is because convergence to the optimal solution in the neighborhood of a given configuration depends on how many times states in that neighborhood are visited during learning: transient configurations are visited less frequently than those at the steady state, leading to higher run-to-run variability in those regions.

The results for the Nkx2.2 target can be compared with those for the Olig2 target (Figures 4B and 4C). Similar values for the activator form of Gli are found at steady state, but much lower values for the Gli repressor are observed. The overall low levels of the effectors are also consistent with the repressive role of Nkx2.2 on Gli gene expression, as supported by experimental data.[15,16,43] It is notable that under the optimally controlled dynamics, a cell reaching the Nkx2.2 target must transition through the Olig2 state before acquiring Nkx2.2 expression.

Morphogen-driven patterning

In the previous section, we identified optimal control strategies independently for two target states. Here, we extend the approach to identify an integrated optimal control strategy that would generate a morphogen-patterned tissue comprising multiple states in response to a spatially graded and dynamic morphogen signal. We then define the state of the controlled system to comprise the GRN state and the signal as subsystems (Figure 5A).

Patterning, as an optimal control problem, can be conceived as a cooperative multi-agent task, whereby multiple cells have to reach their respective targets simultaneously, but where the shared morphogen input provides the positional information. Collectively, cells minimize a global shared cost, with the constraint that the controller function—representing the

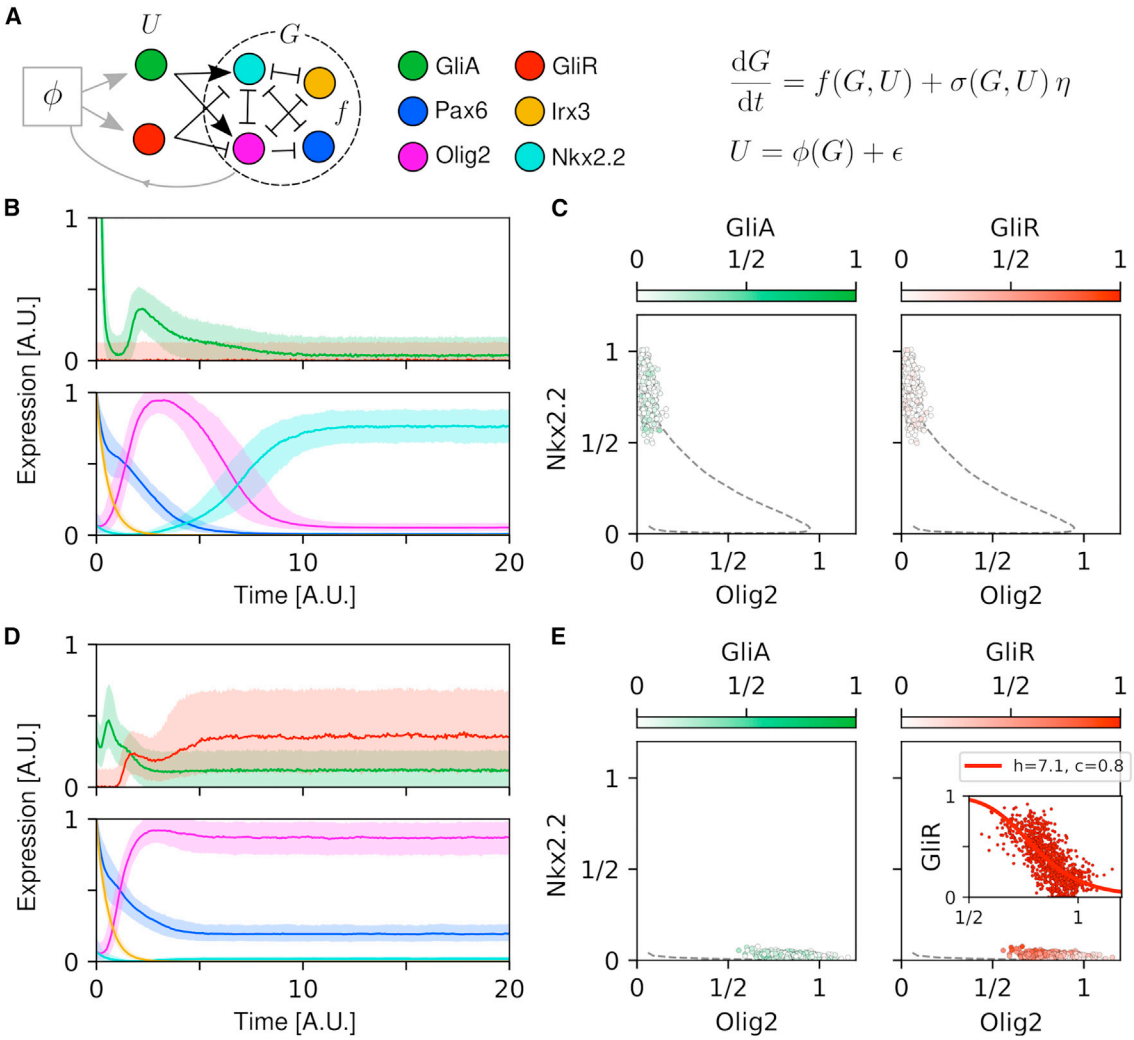


Figure 4. Reinforcement learning solution for the optimal control of the ventral neural tube GRN

(A) Schematic representation of the closed-loop control: the activator and repressor form of Gli proteins are given by a function φ of the neural progenitor markers Pax6, Olig2, Nkx2.2, and Irx3, which evolve according to the stochastic dynamics defined in STAR Methods section ventral neural-progenitor GRN model (PONI network). The colors in the legend are maintained in all plots. The optimal control problem is solved separately for two different targets: (A and C) the Nkx2.2+ target; (D and E) the Olig2+ target. For each problem, a different feedback function φ is found through deep-RL—φ is perturbed by a noise ϵ, as prescribed by the algorithm used (see algorithm 1 in STAR Methods for details about the TD3 algorithm).

(B) Samples of the controlled dynamics for the Nkx2.2+ target (solid lines are the medians, and the shaded areas the 10th to 90th percentile): the control u^*, comprising activator and repressor Gli (top) and the gene expression dynamics (bottom). In this solution, although the repressor Gli remains always low, the activator Gli exhibits an initial transient characterized by high values, to then reduce drastically at steady state, reminiscent of the adaptive dynamics observed in experiments.[16]

(C) Activator Gli (left) and repressor Gli (right) as a function of Olig2 and Nkx2.2 levels at steady state (colored points) together with the corresponding average trajectory (dashed gray line). For the parameters used no feedback φ could be found for the dynamics to "avoid" the transient Olig2+ state.

(D) A similar transient appears in the activator Gli for the Olig2+ target.

(E) In addition, a negative feedback from Olig2 onto the repressor appears to be required to maintain cells in the Olig2+ state, which also explains the high variability of repressor Gli at steady state—(E), right; the inset shows Gli repressor as a function of Olig2 (points), with a repressive Hill function fit, with Hill coefficient h and threshold c. One possibility is that this prevents the activator driving the state toward Nkx2.2+ state (the optimally controlled trajectories of C are overlaid as gray lines—the dashed-gray line is the average).

signaling pathway with its feedback loops—has to be the same for all cells. The target pattern, implemented through the running cost q, has two boundaries that divide the tissue into three equal parts, with ventral, middle, and dorsal fates corresponding to Nkx2.2, Olig2, and Pax6+/Irx3 expressing, respectively (Figure 5B). We adapt the TD3 algorithm for the patterning task and test it on the patterning of the ventral neural tube (see algorithm 2 in STAR Methods).

The morphogen dynamics (Figure 5B) are given by stochastic simulations of a diffusion process of independent Shh particles, whereas the GRN model is the same as in the previous section (details in STAR Methods section environment dynamics). We

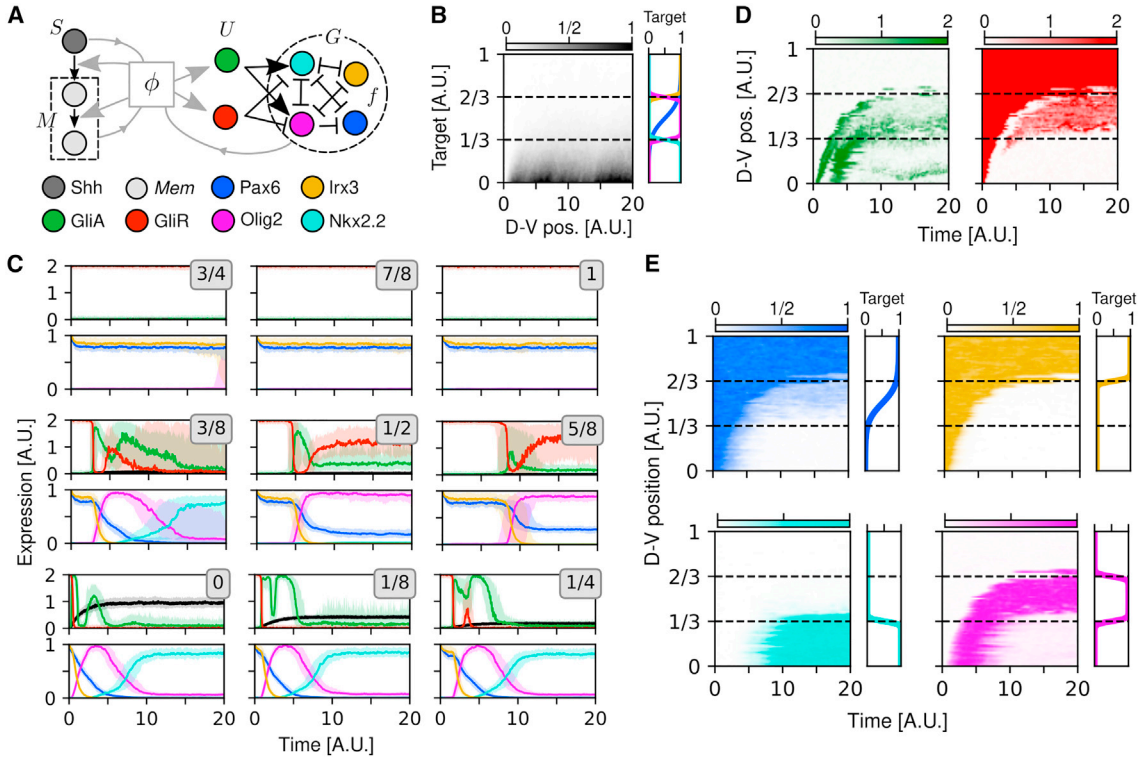


Figure 5. Reinforcement learning solution for the morphogen-driven patterning task

(A) Compared with the single-cell case in Figure 4, in addition to the target genes *G*, the controller φ also receives as input the morphogen signal *S* and the memory variables *M* and returns the maximal production rates for the memory variables in addition to the signaling effectors *U*. Also, a unique feedback function φ needs to be found across all cells, which we seek for by means of multi-agent reinforcement learning algorithm, in a mean-field approximation (see STAR Methods section multi-agent control and algorithm 2 for details).

(B) Driven by a stochastically diffusing morphogen *S* (one realization shown in the heatmap, left), the goal is to minimize a trade-off between the distance from a target gene expression profile (right) and the magnitude of the control over time (see main text). The dashed lines at 1/3 and 2/3 of the total D-V extension indicate the positions of the boundaries between target differential expression regions.

(C) The cell-by-cell view of the dynamics averaged over 100 simulations (solid lines are the medians, shaded areas the 10^{th} to 90^{th} percentile; individual panels are labeled by the D-V position of the selected cells) reveals the control strategy for each position. Similar features shown in Figure 4 are also found here, highlighting the potential functional role of Gli repression by Olig2 and Nkx2.2 in the patterning process.

(D and E) A single realization of the optimally controlled dynamics with the morphogen field as in (B)—the target expression for each gene is shown at the right of each panel for comparison.

derive the optimality equation for this, in the ansatz of independent cells, in STAR Methods section multi-agent control. This ansatz can only be an approximation to the optimal solution because the (stochastic) morphogen dynamics exhibit spatio-temporal correlations. Indeed, it works for a deterministic and static gradient—where the ansatz is exact (Figure S5)—and can be a good approximation when the steady state of the morphogen is reached fast compared with the GRN. A naive implementation of the independence ansatz for a "slow" morphogen fails to reproduce the target pattern, due to the increasing effect of the correlations between morphogen signals at different locations in the tissue. Nevertheless, the (ensemble) average of the morphogen signal experienced by individual cells can be expressed with independent but non-autonomous dynamics (see STAR Methods section dynamics of a stochastic gradient).

This suggests that the introduction of memory variables into the decision-making may help to solve the problem by "extracting" temporal features of the morphogen (Figures 5A and S6; STAR Methods section memory in signal interpretation). These variables can be thought to represent the intermediate components in the signaling cascade, such as the Shh receptor Ptch1, the transmembrane protein Smo, etc. The activity of these components in response to Shh introduces delays and persistence to the transmission of the instantaneous changes in the morphogen. The control model we introduce features more general feedback mechanisms within the signaling cascade and from the GRN species. With this extension, the algorithm is able to find strategies that lead to the target pattern (Figure 5B, right), which we were not able to achieve without the memory variables.

In Figures 5C and 5E, we see the dynamics of the tissue patterning process under the control strategy found by the RL algorithm. At the beginning of the morphogen spread, all cells are in the initial pre-pattern (dorsal) condition. As morphogen spreads into the tissue, Olig2 and Nkx2.2, are sequentially induced ventrally (Figures 5D and 5E, dynamics in space and time, per gene), resulting in a kinematic wave of gene expression spreading from ventral to dorsal until the target pattern is reached. The

pattern is then maintained. The dynamics of the effectors in individual cells (Figure 5C) share some features with those found for the single-cell control (Figures 4A and 4C). Because the initial conditions are the same for all cells in the tissue (Pax6+/Irx3+, vanishing morphogen signal and memory variables—see STAR Methods section memory in signal interpretation), the signal levels are also the same, corresponding to the values needed to maintain cells in the dorsal state, i.e., high levels of repressor together with low levels of activator (Figure 5C, top row). For cells that are assigned to an Olig2+ fate, after an initial delay set by the spread of the Shh morphogen, the dynamics are similar to those found for the Olig2 target in a single cell (Figures 4D and 5E): levels of repressor negatively correlated with Olig2 concentration and low levels of activator at steady state (Figure 5C, middle). In cells acquiring an Nkx2.2+ fate, we also observe a negative correlation of Gli repressor levels with Nkx2.2 (Figure 5C, bottom). Thus, the learned control strategy recovers the repressive feedback from both Olig2 and Nkx2.2 on Gli, which results in adaptive dynamics of the signaling effectors. Both of these features are supported by experimental data.[15,16,42,43]

DISCUSSION

Here, we used optimal control theory to develop a framework to analyze morphogen signaling strategies and identify mechanisms that produce rapid, precise, and reproducible cell-fate decisions during tissue patterning in embryo development. We demonstrate that this framework can be combined with dynamical—Waddington-like—landscape models of cell-fate decisions to provide an optimal control representation in the form of a new landscape. RL can be used to solve optimal control problems associated with signaling and cell-fate decisions, and we formulate the patterning problem as a multi-agent cooperative optimal control task, in which the objective function is a measure of the performance of all the cells in the tissue. By using these approaches to analyze the morphogen patterning of neural progenitors, we highlight how the mechanisms obtained from the optimization are consistent with experimental data.

In the celebrated French Flag model of morphogen patterning, cell fates are proposed to be instructed by morphogen concentration such that the concentration is read out directly by cells.[2] Information-theoretic approaches have built on this view of morphogen activity to develop quantitative measures of positional information based on measuring the local concentrations of patterning molecules at a specific developmental time point.[44–46] This has been applied to patterning of the anterior-posterior axis of the *Drosophila* blastoderm, leading to the idea that precise cellular identities are available directly from the level of morphogen.[46] In this example, the morphogen signal is read out within nuclei, without signal transduction or feedback from the downstream GRN, since the developing tissue is a syncytium, and the precision of patterning is considered in terms of the statistical properties of the morphogen signal steady state. Applying a similar approach to other morphogen-patterned tissues—in which the dynamics of morphogen gradient formation, signaling, and the GRN are crucial to pattern formation, and the morphogen level is not a direct correlate of position—is challenging. From this perspective, the French Flag model does not explain the complex cellular signaling dynamics often observed experimentally. Moreover, it subordinates the role of the GRN to that of the extracellular signals.

The optimal control perspective provides an alternative paradigm that accommodates the dynamics in signal interpretation and establishes a relationship between the control signal and the system. Our analysis revealed that for both individual cell fate decisions and morphogen-driven tissue patterning, adaptive signaling dynamics, which are observed experimentally *in vivo*,[47] emerge as an optimal strategy in the presence of multi-stability. This suggests that signaling pathways may have evolved to take advantage of the dynamic landscape that arises from the GRN.

The objective function includes a notion of "timing" through exponential discounting. This can be regarded as representing the tempo of development and the rate of differentiation in a tissue, which limits the amount of time that is available to the cell to integrate the signal and make a decision. We set this time to be comparable with differentiation rates and the degradation rates of the key transcription factors in the GRN.[48]

Importantly, when a Waddington landscape offers a good phenomenological model of cell-fate decision, the optimal control framework provides analytical tools to "isolate" the contribution of morphogen signaling to the GRN dynamics. Practically, this could be achieved via the comparison of experimentally measured landscapes under different genetic or pharmacologic manipulations of signaling pathways.[20]

There are limitations to our approach that will need to be addressed in future work. In the current formulation, the control input to the system is selected in a "reactive" way, as a function of the target genes. This rules out possible hysteresis effects in feedback mechanisms. This is partially addressed via the addition of memory variables in the morphogen-driven tissue patterning example. However, the signaling effectors—as a function of components of the GRN—still retain a memory-less component. This could be tackled by introducing production-degradation dynamics, where the control defines the production rates, rather than the levels. This would have the benefit of allowing the inclusion of known kinetic properties of the effectors, such as degradation rates.[48] Also, the degradation rate has been assumed independent of the cell state. The control problem solved here can be extended to cases where the terminal-time statistics depends on the state and control variables and include optimal stopping time problems (see, e.g., Sorger[49]).

From the RL perspective, the introduction of memory variables is analogous to the use of recurrent networks for modeling systems with memory,[50] e.g., in partially observable environments.[51,52] Examining this problem in the broader context of decision-making in non-Markovian or non-stationary environments[53] could highlight general design principles that optimally deal with memory. It is interesting to note that the morphogen-driven patterning task can be formally regarded as a classification of signal time series: hidden in the optimally controlled dynamics are the features of the temporal profile of the signal, which can be utilized by the cell in order to make decisions. Hence, the optimal control perspective provides a link between the complex computational problem of morphogen interpretation and the biological hardware available for its solution.

Although we addressed the function of cell-autonomous feedback, such as that mediated by transcriptional targets of the

morphogen-controlled GRN on the expression of components of the morphogen signal transduction pathway,[16] we did not address all possible feedback mechanisms that could be exploited by the system. For example, Shh signaling controls the expression of Shh binding proteins, such as Ptch1, Scube2, and Hhip1, that alter transport of the morphogen through the tissue.[12,14,54] Feedback on morphogen spread could be incorporated into the model. Indeed, the framework could be used to investigate virtually any aspect of the system. This could include, for example, control of diffusivity of signals, degradation rates of system components, or the accessibility of *cis*-regulatory elements and the effect of chromatin remodeling, all of which have been implicated in the interpretation of morphogen signaling.[1,8,14]

The patterning example dealt with in this study is one in which positional information is provided by a signal external to the tissue. In other cases, symmetry is broken, and patterning is controlled by internally generated signals, such as in the case of organoids patterned by Turing-like mechanisms[55] Patterning, in these contexts, poses a problem of coordination by means of signaling that can be cast into a multi-agent decision-making task. This, in turn, can be tackled numerically with multi-agent RL (MARL) algorithms[56,57] or analytically via, e.g., mean-field approximation in the limit of large numbers of cells.[58,59] Therefore, optimal control provides a framework in which to analyze these systems to investigate functional explanations for the observed signaling strategies, proportions of cell types, and self-organization of patterning.

The optimal control approach, with its focus on linking mechanisms with control, is ideally suited for the analysis of *in vitro* and synthetic systems. This could be used to design and refine signaling regimes for the directed differentiation of stem cells *in vitro* and the production of specific sets of cell types in defined proportions. An understanding of the control principles operating in biological systems may provide insights and inspiration for the construction of artificial systems as well as support the use of stem cells in disease modeling and regenerative medicine.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Overview of optimal stochastic control and its solution
  - Optimal control in a potential
  - Environment dynamics
  - Multi-Agent control
  - RL solution

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.10.004.

ACKNOWLEDGMENTS

We are grateful to Rubèn Perez-Carrasco and Zena Hadjivasiliou and members of the lab for their constructive comments. A.P. thanks Antonio Celani for insightful discussions. A.P. was funded by an EMBO Long Term Fellowship (ALTF 860-2019). This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK, the UK Medical Research Council, and Wellcome Trust (all under CC001051). Work in the Briscoe lab is funded by the European Research Council under European Union (EU) Horizon 2020 research and innovation program grant 742138.

AUTHOR CONTRIBUTIONS

A.P. and J.B. conceptualized the research. A.P. conducted the formal analysis and performed simulations. A.P. and J.B. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 17, 2022
Revised: June 6, 2023
Accepted: October 10, 2023
Published: November 15, 2023

REFERENCES

1. Stapornwongkul, K.S., and Vincent, J.-P. (2021). Generation of extracellular morphogen gradients: the case for diffusion. Nat. Rev. Genet. *22*, 393–411. https://doi.org/10.1038/s41576-021-00342-y.
2. Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. J. Theor. Biol. *25*, 1–47. https://doi.org/10.1016/S0022-5193(69)80016-0.
3. Balaskas, N., Ribeiro, A., Panovska, J., Dessaud, E., Sasai, N., Page, K.M., Briscoe, J., and Ribes, V. (2012). Gene regulatory logic for reading the sonic hedgehog signaling gradient in the vertebrate neural tube. Cell *148*, 273–284. https://doi.org/10.1016/j.cell.2011.10.047.
4. Briscoe, J., and Small, S. (2015). Morphogen rules: design principles of gradient-mediated embryo patterning. Development *142*, 3996–4009. https://doi.org/10.1242/dev.129452.
5. Green, J.B.A., and Sharpe, J. (2015). Positional information and reaction-diffusion: two big ideas in developmental biology combine. Development *142*, 1203–1211. https://doi.org/10.1242/dev.114991.
6. Manu, S., Surkova, S., Spirov, A.V., Gursky, V.V., Janssens, H., Kim, A.R., Radulescu, O., Vanario-Alonso, C.E., Sharp, D.H., Samsonova, M., et al. (2009). Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. PLoS Biol. *7*, e1000049. https://doi.org/10.1371/journal.pbio.1000049.
7. Manu, Surkova, S., Spirov, A.V., Gursky, V.V., Janssens, H., Kim, A.-R., Radulescu, O., Vanario-Alonso, C.E., Sharp, D.H., Samsonova, M., and Reinitz, J. (2009). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. PLOS Comp. Biol. *5*, e1000303. https://doi.org/10.1371/journal.pcbi.1000303.
8. Exelby, K., Herrera-Delgado, E., Perez, L.G., Perez-Carrasco, R., Sagner, A., Metzis, V., Sollich, P., and Briscoe, J. (2021). Precision of tissue patterning is controlled by dynamical properties of gene regulatory networks. Development *148*, dev197566. https://doi.org/10.1242/dev.197566.
9. Zagorski, M., Tabata, Y., Brandenberg, N., Lutolf, M.P., Tkačik, G., Bollenbach, T., Briscoe, J., and Kicheva, A. (2017). Decoding of position in the developing neural tube from antiparallel morphogen gradients. Science *356*, 1379–1383. https://doi.org/10.1126/science.aam5887.
10. Lander, A.D. (2013). How cells know where they are. Science *339*, 923–927. https://doi.org/10.1126/science.1224186.

11. Briscoe, J., and Ericson, J. (2001). Specification of neuronal fates in the ventral neural tube. Curr. Opin. Neurobiol. *11*, 43–49. https://doi.org/10.1016/S0959-4388(00)00172-0.

12. Ribes, V., and Briscoe, J. (2009). Establishing and interpreting graded sonic hedgehog signaling during vertebrate neural tube patterning: the role of negative feedback. Cold Spring Harbor Perspect. Biol. *1*, a002014. https://doi.org/10.1101/cshperspect.a002014.

13. Cohen, M., Page, K.M., Perez-Carrasco, R., Barnes, C.P., and Briscoe, J. (2014). A theoretical framework for the regulation of Shh morphogen-controlled gene expression. Development *141*, 3868–3878. https://doi.org/10.1242/dev.112573.

14. Jeong, J., and McMahon, A.P. (2005). Growth and pattern of the mammalian neural tube are governed by partially overlapping feedback activities of the hedgehog antagonists patched 1 and Hhip1. Development *132*, 143–154. https://doi.org/10.1242/dev.01566.

15. Lek, M., Dias, J.M., Marklund, U., Uhde, C.W., Kurdija, S., Lei, Q., Sussel, L., Rubenstein, J.L., Matise, M.P., Arnold, H.-H., et al. (2010). A homeodomain feedback circuit underlies step-function interpretation of a shh morphogen gradient during ventral neural patterning. Development *137*, 4051–4060. https://doi.org/10.1242/dev.054288.

16. Cohen, M., Kicheva, A., Ribeiro, A., Blassberg, R., Page, K.M., Barnes, C.P., and Briscoe, J. (2015). Ptch1 and Gli regulate Shh signalling dynamics via multiple mechanisms. Nat. Commun. *6*, 6709. https://doi.org/10.1038/ncomms7709.

17. Waddington, C.H. (1957). The Strategy of the Genes (Routledge).

18. Corson, F., and Siggia, E.D. (2012). Geometry, epistasis, and developmental patterning. Proc. Natl. Acad. Sci. USA *109*, 5568–5575. https://doi.org/10.1073/pnas.1201505109.

19. Corson, F., and Siggia, E.D. (2017). Gene-free methodology for cell fate dynamics during development. eLife *6*, e30743. https://doi.org/10.7554/eLife.30743.

20. Sáez, M., Blassberg, R., Camacho-Aguilar, E., Siggia, E.D., Rand, D.A., and Briscoe, J. (2022). Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions. Cell Syst. *13*, 12–28.e3. https://doi.org/10.1016/j.cels.2021.08.013.

21. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: an Introduction (MIT Press).

22. Marr, D.C., and Poggio, T. (1976). From Understanding Computation to Understanding Neural Circuitry (Massachusetts Institute of Technology Artificial Intelligence Laboratory), pp. 1–22.

23. Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (MIT Press). https://doi.org/10.7551/mitpress/9780262514620.001.0001.

24. Willshaw, D.J., Dayan, P., and Morris, R.G.M. (2015). Memory, modelling and Marr: a commentary on Marr (1971) 'Simple memory: a theory of archicortex'. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140383. https://doi.org/10.1098/rstb.2014.0383.

25. Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R. (2005). Transcriptional regulation by the numbers: applications. Curr. Opin. Genet. Dev. *15*, 125–135. https://doi.org/10.1016/j.gde.2005.02.006.

26. Marr, D. (1975). Approaches to biological information processing: physics and Mathematics of the Nervous System. Proceedings of a summer school, Trieste, Italy, Aug. 1973. M. Conrad, W. Güttinger, and M. Dal Cin, Eds. Springer-Verlag, New York, 1974. xiv, 584 pp., illus. Paper, $18.50. Lecture Notes in Mathematics, vol. 4. Science *190*, 875–876. https://doi.org/10.1126/science.190.4217.875.

27. Anderson, P.W. (1972). More is different. Science *177*, 393–396. https://doi.org/10.1126/science.177.4047.393.

28. Bohm, D. (1984). Causality and Chance in Modern Physics (Routledge).

29. Levins, R., and Lewontin, R. (1985). The Dialectical Biologist (Harvard University Press).

30. Woods, A., and Grant, T. (2012). Reason in Revolt (Wellred, London).

31. Gillespie, D.T. (2000). The chemical Langevin equation. J. Chem. Phys. *113*, 297–306. https://doi.org/10.1063/1.481811.

32. Van Kampen, N.G. (2007). Stochastic Processes in Physics and Chemistry, Third Edition (Elsevier). https://doi.org/10.1016/B978-0-444-52965-7.X5000-4.

33. Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. USA *99*, 12795–12800. https://doi.org/10.1073/pnas.162041399.

34. Coomer, M.A., Ham, L., and Stumpf, M.P.H. (2022). Noise distorts the epigenetic landscape and shapes cell-fate decisions. Cell Syst. *13*, 83–102.e6. https://doi.org/10.1016/j.cels.2021.09.002.

35. Seifert, U. (2008). Stochastic thermodynamics: principles and perspectives. Eur. Phys. J. B *64*, 423–431. https://doi.org/10.1140/epjb/e2008-00001-9.

36. Bo, S., and Celani, A. (2014). Entropy production in stochastic systems with fast and slow time-scales. J. Stat. Phys. *154*, 1325–1351. https://doi.org/10.1007/s10955-014-0922-1.

37. Bo, S., and Celani, A. (2017). Multiple-scale stochastic processes: decimation, averaging and beyond. Phys. Rep. *670*, 1–59. https://doi.org/10.1016/j.physrep.2016.12.003.

38. Bertsekas, D.P. (2017). Dynamic programming and optimal control*1* (Athena scientific Belmont, MA).

39. Bryson, A.E., and Ho, Y.-C. (2018). Applied Optimal Control, First Edition (Routledge). https://doi.org/10.1201/9781315137667.

40. Perez-Carrasco, R., Guerrero, P., Briscoe, J., and Page, K.M. (2016). Intrinsic noise profoundly alters the dynamics and steady state of morphogen-controlled bistable genetic switches. PLoS Comp. Biol. *12*, e1005154. https://doi.org/10.1371/journal.pcbi.1005154.

41. Fujimoto, S., van Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.09477.

42. Junker, J.P., Peterson, K.A., Nishi, Y., Mao, J., McMahon, A.P., and van Oudenaarden, A. (2014). A predictive model of bifunctional transcription factor signaling during embryonic tissue patterning. Dev. Cell *31*, 448–460. https://doi.org/10.1016/j.devcel.2014.10.017.

43. Nishi, Y., Zhang, X., Jeong, J., Peterson, K.A., Vedenko, A., Bulyk, M.L., Hide, W.A., and McMahon, A.P. (2015). A direct fate exclusion mechanism by Sonic hedgehog-regulated transcriptional repressors. Development *142*, 3286–3293. https://doi.org/10.1242/dev.124636.

44. Gregor, T., Tank, D.W., Wieschaus, E.F., and Bialek, W. (2007). Probing the limits to positional information. Cell *130*, 153–164. https://doi.org/10.1016/j.cell.2007.05.025.

45. Tkačik, G., Dubuis, J.O., Petkova, M.D., and Gregor, T. (2015). Positional information, positional error, and readout precision in morphogenesis: A mathematical framework. Genetics *199*, 39–59. https://doi.org/10.1534/genetics.114.171850.

46. Petkova, M.D., Tkačik, G., Bialek, W., Wieschaus, E.F., and Gregor, T. (2019). Optimal decoding of cellular identities in a genetic network. Cell *176*, 844–855.e15. https://doi.org/10.1016/j.cell.2019.01.007.

47. Dessaud, E., Yang, L.L., Hill, K., Cox, B., Ulloa, F., Ribeiro, A., Mynett, A., Novitch, B.G., and Briscoe, J. (2007). Interpretation of the sonic hedgehog morphogen gradient by a temporal adaptation mechanism. Nature *450*, 717–720. https://doi.org/10.1038/nature06347.

48. Rayon, T., Stamataki, D., Perez-Carrasco, R., Garcia-Perez, L., Barrington, C., Melchionda, M., Exelby, K., Lazaro, J., Tybulewicz, V.L.J., Fisher, E.M.C., and Briscoe, J. (2020). Species-specific pace of development is associated with differences in protein stability. Science *369*, eaba7667. https://doi.org/10.1126/science.aba7667.

49. Sorger, G. (1991). Maximum principle for control problems with uncertain horizon and variable discount rate. J. Optim. Theor. Appl. *70*, 607–618. https://doi.org/10.1007/BF00941305.

50. Graves, A., Mohamed, A., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on

Acoustics, Speech and Signal Processing, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947.

51. Hausknecht, M., and Stone, P. (2017). Deep recurrent Q-learning for partially observable MDPs. Preprint at arXiv. https://doi.org/10.48550/arXiv.1507.06527.
52. Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., Rae, J., Mirowski, P., Leibo, J.Z., Santoro, A., et al. (2018). Unsupervised predictive memory in a goal-directed agent. Preprint at arXiv. https://doi.org/10.48550/arXiv.1803.10760.
53. Gajane, P., Ortner, R., and Auer, P. (2019). Variational regret bounds for reinforcement learning. Preprint at arXiv. https://doi.org/10.48550/arxiv.1905.05857.
54. Collins, Z.M., Cha, A., Qin, A., Ishimatsu, K., Tsai, T.Y.C., Swinburne, I.A., Li, P., and Megason, S.G. (2018). A Scube2-Shh feedback loop links morphogen release and spread to morphogen signaling to enable scale invariant patterning of the ventral neural tube. Preprint at bioRxiv. https://doi.org/10.1101/469239.
55. Ishihara, K., and Tanaka, E.M. (2018). Spontaneous symmetry breaking and pattern formation of organoids. Curr. Opin. Syst. Biol. *11*, 123–128. https://doi.org/10.1016/j.coisb.2018.06.002.
56. Littman, M.L. (1994). Markov games as a framework for multi-agent reinforcement learning. In Machine Learning Proceedings, *1994* (Elsevier), pp. 157–163. https://doi.org/10.1016/B978-1-55860-335-6.50027-1.
57. Canese, L., Cardarilli, G.C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., and Spanò, S. (2021). Multi-agent reinforcement learning: a review of challenges and applications. Appl. Sci. *11*, 4948. https://doi.org/10.3390/app11114948.
58. Lasry, J.-M., and Lions, P.-L. (2007). Mean field games. Jpn. J. Math. *2*, 229–260. https://doi.org/10.1007/s11537-007-0657-8.
59. Pezzotta, A., Adorisio, M., and Celani, A. (2018). Chemotaxis emerges as the optimal solution to cooperative search games. Phys. Rev. E *98*, 42401. https://doi.org/10.1103/PhysRevE.98.042401.
60. Bellman, R. (1952). On the theory of dynamic programming. Proc. Natl. Acad. Sci. USA *38*, 716–719. https://doi.org/10.1073/pnas.38.8.716.
61. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature *518*, 529–533. https://doi.org/10.1038/nature14236.
62. Todorov, E. (2009). Efficient computation of optimal actions. Proc. Natl. Acad. Sci. USA *106*, 11478–11483. https://doi.org/10.1073/pnas.0710743106.
63. Dvijotham, K., and Todorov, E. (2011). A unified theory of linearly solvable optimal control.. Preprint at arXiv. https://doi.org/10.48550/arXiv.1202.3715.
64. Gardiner, C. (2009). Stochastic Methods – A Handbook for the Natural and Social Sciences (Springer).
65. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1509.02971.
66. Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. Preprint at arXiv. https://doi.org/10.48550/arXiv.1801.01290.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
GRN-control	This paper	https://doi.org/10.5281/zenodo.8321764

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Alberto Pezzotta (a.pezzotta@ucl.ac.uk).

Materials availability

This study did not generate new materials.

Data and code availability

- This paper did not generate new data and does not analyze existing data.
- All original code has been deposited at https://doi.org/10.5281/zenodo.8321763 and is publicly available as of the date of publication.
- Any additional information is available from the lead contact upon request.

METHOD DETAILS

Overview of optimal stochastic control and its solution

A system with state variables x satisfies the controlled stochastic dynamics

$$\frac{dx}{dt} = f(x,u) + \sigma(x,u)\,\eta(t), \qquad \text{(Equation 6)}$$

where f is a deterministic drift, σ – multiplying the standard Gaussian white noise η – is the magnitude of the multiplicative noise (interpreted here according to the Itô convention) and u represent a set of control variables. We ask what is the optimal choice of the control variables u over time in that minimizes the mean of a cost function

$$C = \int_0^\infty dt\, e^{-t/\tau} m(x(t),u(t)), \qquad \text{(Equation 7)}$$

where m is a cost per unit time (also termed running cost) associated with the instantaneous state and control at a given time, and τ sets the time-scale for the exponential discount factor – defining the "far-sightedness" of the decision-maker in the estimation of the cost that is expected to be paid in the future. As we show in STAR Methods section terminal cost and discounting, optimal-control problems with terminal-state cost and uncertain terminal time can be cast in the minimisation of a cost function of the form Equation 7. Throughout this study, the running cost has the form $m(x,u) = q(x) + \epsilon \| u\|^2/2$, that is a trade-off between the squared magnitude of the control and a state-dependent cost measuring the squared distance from a target ξ, $q(x) = \| x - \xi\|^2/2$.

For the class of cost functions in the form of Equation 7, it is possible to solve the optimal control problem via dynamic programming. This is achieved by maximising, at every state x, the value function J_u, defined as the negative of the cost-to-go function

$$J_u(x) = -\mathbb{E}_{u(\cdot)}[C|x(0) = x] \qquad \text{(Equation 8)}$$

i.e., the cost to be paid conditioned on the initial state, averaged over all the realisations dynamics in Equation 6, with control function u. Via Itô rule, it is possible to show that this satisfies

$$f\cdot\nabla J_u + D\nabla^2 J_u - m = 0 \qquad \text{(Equation 9)}$$

where $D = \sigma\sigma^\top/2$ and ∇ is the gradient with respect to the state variables x. Denoting the component α of the state x by x^α, the differential operators in Equation 9 are expressed as $f\cdot\nabla = \sum_\alpha f^\alpha \partial/\partial x^\alpha$ and $D\nabla^2 = \sum_{\alpha,\beta} D^{\alpha\beta}\partial^2/\partial x^\alpha \partial x^\beta$.

The value function corresponding to the optimal control u^*, denoted $J^* \equiv J_{u^*}$, therefore satisfies

$$\max_u \{f \cdot \nabla J^* + D\,\nabla^2 J^* - m\} = 0\,. \quad \text{(Equation 10)}$$

This equation, known as the dynamic programming (or Bellman) equation,[38,60] yields the optimal cost as well as the optimal control as a function u^* of the state x. The non-linearity introduced by the max operator, along with the infinite number of states (for continuous states and actions), makes the exact solution of Equation 10 generally impossible. Note that here we did not make any restrictive assumption on the type of noise, i.e. additive vs multiplicative.

Numerical techniques can be employed to find approximate solutions: reinforcement learning (RL)[21] with function approximation through deep neural networks[41,61] is the numerical scheme used in this work for the solution of Equation 10 for the optimal control of the ventral neural tube GRN. However, the case where σ is constant while f and m have, respectively, linear and quadratic dependence on u (as in the case of the control in a landscape dealt with in the main text), falls into a general class of linearly solvable control problems,[62,63] in that Equation 10 can be cast into a linear form through a change of variables (as detailed in STAR Methods section optimal control in a potential).

Optimal control in a potential

Let us consider the Langevin dynamics

$$\mathrm{d}x = (-\nabla V + u)\mathrm{d}t + \sqrt{2D}\,\mathrm{d}W_t \quad \text{(Equation 11)}$$

where V is a confining potential, u is an additional control drift, W_t is a standard Wiener process with $\mathbb{E}[\mathrm{d}W_t \mathrm{d}W_{t'}] = \delta(t - t')\mathrm{d}t$, and D is the diffusion constant, setting the strength of the noise. Everywhere in the manuscript, the noise is interpreted à la Itô. In this section the noise is additive (we assume that D does not depend on x or t) and the noise interpretation does not pose a problem.

The control u is chosen to minimize a given cost functional, as detailed in the following. We choose the potential V in such a way that the uncontrolled dynamics has two stable fixed points (i.e. minima of V) at $x = \pm 1$: $V(x) = x^4/4 - x^2/2$.

Stationary-state optimization

We introduce the cost function

$$C_u = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}t \left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t))\right) \quad \text{(Equation 12)}$$

with

$$q(x,u) = \frac{1}{2}|x - \xi|^2, \quad \text{(Equation 13)}$$

measuring an overall distance from the target state along the trajectory. The quadratic cost for control can be interpreted as the Kullback-Leibler (KL) divergence of the measure of controlled paths from that of uncontrolled paths, i.e. generated by the dynamics with or without u; this KL divergence takes the form of an integral in time of the KL divergence between the controlled and uncontrolled (infinitesimal time increment $\mathrm{d}t$) propagators: both propagators are Gaussian distributions with the same variance $4D\mathrm{d}t$, but with means differing by $u\mathrm{d}t$, which yields the $|u|^2$ term. See also Todorov[62] and Dvijotham and Todorov.[63]

We seek to find the control strategy u that minimizes the expectation value of C_u over all realisations of the stochastic dynamics Equation 11. If the system is ergodic, $\mathbb{E}[C_u|X_0 = x]$ is a constant, i.e. it does not depend on the initial condition. In particular, this average is equivalent to that of the running cost at the stationary state:

$$\mathbb{E}[C_u|X_0 = x] = \mu = \int dx\, \rho_{\mathrm{eq}}(x) \left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t))\right) \quad \text{(Equation 14)}$$

We can introduce the value function

$$J(x) = -\lim_{T\to\infty} \mathbb{E}\left[\int_0^T \mathrm{d}t' \left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t)) - \mu\right) \middle| x_0 = x\right] \quad \text{(Equation 15)}$$

that is (minus) the excess cumulated cost from a given state relative to the steady state average. We can use the Feynman-Kac formula,[64] to show that this satisfies

$$-D\nabla^2 J - (u - \nabla V)\cdot\nabla J + q + \frac{\epsilon}{2}u^2 = \mu. \quad \text{(Equation 16)}$$

It can be verified by multiplying by the steady state (equilibrium) distribution ρ_{eq}, satisfying $(u - \nabla V)\rho_{\mathrm{eq}} = D\nabla\rho_{\mathrm{eq}}$, and integrating over all states. The principle of dynamic programming holds that in order to minimize μ, it is sufficient to minimize $J(x)$ for every x. We therefore see that the minimum condition for J yields

$$u^* = \frac{1}{\epsilon}\nabla J^* \quad \text{(Equation 17)}$$

and that the optimal value function J^* satisfies the Bellman equation

$$-D\nabla^2 J^* - \frac{1}{2\epsilon}|\nabla J^*|^2 + \nabla V\cdot\nabla J^* + q = \mu^*. \quad \text{(Equation 18)}$$

The constant μ^* is the minimum average cost at the stationary state.
By replacing $J^* = \epsilon(V + 2D\log\psi)$ this rewrites

$$-D\nabla^2\psi + \left(\frac{q}{2D\epsilon} + \frac{|\nabla V|^2}{4D} - \frac{\nabla^2 V}{2}\right)\psi = \frac{\mu^*}{2D\epsilon}\psi \quad \text{(Equation 19)}$$

This is formally equivalent to the ground-state problem of a quantum particle of mass $m = 2D/\hbar^2$ in the potential

$$V_S = \frac{q}{2D\epsilon} + \frac{|\nabla V|^2}{4D} - \frac{\nabla^2 V}{2}. \quad \text{(Equation 20)}$$

The change of variables implies that the optimally controlled dynamics is given by

$$dx = 2D\nabla\log\psi dt + \sqrt{2D}\, dW_t. \quad \text{(Equation 21)}$$

From the Fokker-Planck equation associated to Equation 21:

$$\partial_t\rho + \nabla\cdot(2D\,\rho\,\nabla\log\psi - D\nabla\rho) = 0 \quad \text{(Equation 22)}$$

we see that the function ψ is related to the equilibrium steady-state distribution, $\rho_{eq}\propto\psi^2$.

This ground-state problem can be solved by introducing a fictitious dynamics in imaginary time,

$$\partial_s\tilde{\psi} = -\hat{H}\,\tilde{\psi} \quad \text{(Equation 23)}$$

with the Hermitian operator $\hat{H} = -D\nabla^2 + V_S$. The ground state ψ_0 of the Hamiltonian $\hat{H}$ is the slowest mode in the imaginary time evolution, and in the long-time limit, Equation 23 is solved by

$$\tilde{\psi}\to e^{-E_0 s}\,\psi_0 \quad \text{(Equation 24)}$$

The solution of the HJB equation, ψ, then identifies with $\tilde{\psi}$, up to a scaling factor which depends solely on time. From the rate of change of the norm of $\tilde{\psi}$ we can infer the minimum average cost:

$$\mu^* = 2D\epsilon E_0 = -2D\epsilon\lim_{s\to\infty}\partial_s\log\|\tilde{\psi}\|_2. \quad \text{(Equation 25)}$$

Exponential discounting

The control can also be chosen to minimize a cost over a shorter window of time, rather than at the steady-state. This can be done by introducing an exponential discount factor over time, as in

$$C_u = \int_0^\infty dt\, e^{-t/\tau}\left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t))\right) \quad \text{(Equation 26)}$$

where τ sets a typical time scale over which rewards are accumulated in the future. As in the above case, we seek u that minimizes the expectation value $\mathbb{E}[C_u]$ over the stochastic dynamics.

We can introduce the value function as (minus) the expected discounted cost-to-go from a given state at a given time

$$J(x,t) = -\lim_{T\to\infty}\mathbb{E}\left[\int_t^T dt'\, e^{-(t'-t)/\tau}\left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t))\right)\,\middle|\, x_t = x\right] \quad \text{(Equation 27)}$$

We see that this satisfies

$$-D\nabla^2 J - (u - \nabla V)\cdot\nabla J + \tau^{-1}J + q + \frac{\epsilon}{2}u^2 = 0\,. \quad \text{(Equation 28)}$$

The optimality condition requires the control to be given by $u^* = \epsilon^{-1}\nabla J$, and optimality Bellman equation writes

$$-D\nabla^2 J^* - \frac{1}{2\epsilon}|\nabla J^*|^2 + \tau^{-1}J^* + \nabla V\cdot\nabla J^* + q = 0\,. \quad \text{(Equation 29)}$$

Analogously to the above case, with the transformation $J^* = \epsilon(V + 2D\log\psi)$, the Bellman equation takes the form

$$\hat{H}\psi \equiv -D\nabla^2\psi + \left(\frac{q}{2D\epsilon} + \frac{|\nabla V|^2}{4D} - \frac{\nabla^2 V}{2} + \tau^{-1}\left(\frac{V}{2D} + \log\psi\right)\right)\psi = 0 \quad \text{(Equation 30)}$$

This non-linear Schrödinger equation can be solved numerically in a similar way as above, by introducing a fictitious dynamics in imaginary time, Equation 23, and solving it until convergence to the stationary state $\hat{H}\psi = 0$.

Terminal cost and discounting

For a process that terminates with a probability per unit time τ^{-1} (or, in other terms, the probability density function for the terminal time is exponential, with mean τ), the exponential discount factor corresponds to the probability that a process that started at time t has not yet terminated at time t':

$$\text{Prob}\{\text{not yet terminated after } \Delta t\} = \int_{\Delta t}^{\infty} \frac{dt}{\tau}\, e^{-t/\tau} = e^{-\Delta t/\tau} \quad \text{(Equation 31)}$$

Therefore, the average of the cost C_u in Equation 26 is equivalent to that of

$$\tilde{C}_u = \int_0^T dt \left(\frac{\epsilon}{2}|u(t)|^2 + q(x(t))\right) \quad \text{(Equation 32)}$$

where T is the exponentially-distributed terminal time with mean τ.

For the dynamics with a terminal state (time), we can include a terminal cost at the time T, $Q(x(T))$. This is particularly relevant in the case of the cell-fate decision or the patterning example considered in the main text.

We can change the definition of the value function in Equation 27 by subtracting the contribution from the terminal cost. This can be written as

$$\mathbb{E}[Q(x(T))|x_t = x] = \int_t^{\infty} dT \tau^{-1}\, e^{-(T-t)/\tau} \mathbb{E}_{x_T = x'}[Q(x') \mid x_t = x] \quad \text{(Equation 33)}$$

Together with the expression in Equation 27, the value for the task including the terminal cost can be expressed as

$$J(x,t) = -\lim_{T\to\infty} \mathbb{E}\left[\int_t^T dt'\, e^{-(t'-t)/\tau}\left(\frac{\epsilon}{2}|u(t')|^2 + q(x(t')) + \tau^{-1} Q(x(t'))\right) \middle| x_t = x\right]. \quad \text{(Equation 34)}$$

Therefore, we recognise that the addition of the terminal cost is equivalent to the replacement of the state-dependent running cost q by $\tilde{q} = q + \tau^{-1}Q$ in Equation 26.

If we choose the terminal cost to be given by the same function q (the dimensions do not match, so we understand that Q is equal to q multiplied by a unit time constant), then $\tilde{q} = (1+\tau^{-1})\, q$. Since the optimal solution is invariant upon multiplications of the cost function by a global constant (see Equation 17), the problem is equivalent to the one where q is kept the same, but τ enters as a re-scaling of the trade-off parameter ϵ, replaced by $\tilde{\epsilon} = \epsilon/(1+\tau^{-1})$.

First passage time near target

The mean first passage time (MFPT) at a given point $\bar{x}$, $T_{\bar{x}}$ for a process starting at a point $x < \bar{x}$, is expressed as

$$T_{\bar{x}}(x) = \mathbb{E}\left[\int_0^{\infty} dt' 1 \,\middle|\, x_t = x\right], \quad \text{(Equation 35)}$$

where the region $x \geq \bar{x}$ is replaced by absorbing states (viceversa if $x > \bar{x}$). For the optimally control dynamics given in Equation 21, this satisfies[64]

$$2D \frac{d}{dx}\log\psi \cdot \frac{d}{dx} T_{\bar{x}}(x) + D \frac{d^2}{dx^2} T_{\bar{x}}(x) = -1\,. \quad \text{(Equation 36)}$$

Its solution can be found by explicit quadratures, with the boundary conditions $T_{\bar{x}}(\bar{x}) = 0$ and $T_{\bar{x}}(x \to -\infty) = \infty$:

$$T_{\bar{x}}(x) = \frac{1}{D}\int_x^{\bar{x}} dx' \int_{-\infty}^{x'} dx'' \frac{\psi(x'')^2}{\psi(x')^2} \quad \text{(Equation 37)}$$

By interpreting $\psi^2 = \exp(-V_{\text{eff}}/D)$, we have

$$T_{\bar{x}}(x) = \frac{1}{D}\int_x^{\bar{x}} dx' \int_{-\infty}^{x'} dx'' \exp\left(-\frac{V_{\text{eff}}(x'') - V_{\text{eff}}(x')}{D}\right) \quad \text{(Equation 38)}$$

When V_{eff} has two minima, in the small-D limit, Equation 38 recovers the Freidlin-Wentzel theory of stochastic transitions via the saddle-point approximation.[64]

Low control and diffusion limit

For small values of u, the controlled potential $V(x,u)$ still has two minima, corresponding to the stable fixed points of the controlled dynamics. If D is also small, the transitions between the two fixed points are rare, while typical realisations of the noise will produce small fluctuations around these: in this limit, Equation 38 gives the Freidlin-Wentzel theory of stochastic transitions,[64] where the MFPT from the left minimum x_- to the right minimum x_+ is therefore approximated as

$$T_{x_+}(x_-) \simeq \frac{1}{D} e^{\Delta V_{\text{eff}}/D} \quad \text{(Equation 39)}$$

where $\Delta V_{\text{eff}} = V_{\text{eff}}(x_0) - V_{\text{eff}}(x_-)$, with x_0 denoting the local maximum of the potential (or saddle) between the two minima. The rate for the opposite transition is analogously given by swapping $x_- \leftrightarrow x_+$.

The steady-state probability to be near one or the other fixed point is given by the average exit time from the fixed point attractor. In the present example, this can be calculated as the MFPT from $x_- \simeq -1$ to $x_+ \simeq 1$, and vice versa.

First of all, we need to solve for the stationary points at a given value of u. In the linear approximation in u, these are

$$x_\pm \simeq \pm 1 + u/2 \ \text{(stable)} \quad \text{and} \quad x_0 \simeq -u \ \text{(unstable)} \tag{Equation 40}$$

The value of the potential at these points is

$$V(x_\pm, u) \simeq -1/4 \mp u, V(x_0, u) \simeq 0 \tag{Equation 41}$$

The MFPT for the "reverse" transition, $T_{x_-}(x_+)$, and the MFPT for the "forward" one, $T_{x_+}(x_-)$, are given by Equation 39, and their ratio gives the relative probability to be in the right or the left attractor at steady state:

$$\frac{\rho_+}{\rho_-} \simeq \frac{T_{x_-}(x_+)}{T_{x_+}(x_-)} \simeq e^{2u/D}. \tag{Equation 42}$$

Therefore, we see that when $D \ll 1$, for a range of control in the regime $D \ll |u| \ll 1$, the probability distribution is highly skewed towards one of the two attractors.

Environment dynamics

Ventral neural-progenitor GRN model (PONI network)

We outline here the details of the GRN model first presented in Cohen et al.,[13] with the addition of noise through the chemical Langevin equation approximation.[8,31]

We denote by H^+ the Hill function

$$H^+(x) = \frac{x}{1+x}, \tag{Equation 43}$$

and by the latin letters the concentrations of the transcription factors, i.e. $P \equiv$ [Pax6], $O \equiv$ [Olig2], $N \equiv$ [Nkx2.2], $I \equiv$ [Irx3], $A \equiv$ [GliA], $R \equiv$ [GliR]. The dynamics of gene X, in the ventral neural tube GRN (PONI network), is given by the stochastic differential equation (SDE)

$$dX = f_X(P, O, N, I)dt + \sqrt{2/\Omega}\, g_X(P, O, N, I)\, dW_X \tag{Equation 44}$$

where f_X is the deterministic component of the dynamics (drift), g_X is a multiplicative noise term (intrinsic fluctuations), dW_X is an independent Brownian increment and Ω is an effective system size (copy number) parameter that controls the size of the fluctuations. In Equation 44, the noise must be interpreted according to the Itô convention: only in such convention the CLE can be read as the SDE corresponding to the Fokker-Planck equation that results from expansion in the inverse of the system size Ω of the chemical master equation.[31,32,64] In this context, therefore, the convention is implicit in the modelling framework.

The drift terms for each gene are

$$\begin{aligned}
f_P &= \alpha_{\text{Pax}} H^+\left(\frac{K_{\text{Pax,Pol}}\, c_{\text{Pol}}}{(1+K_{\text{Pax,Oli}}\, O)^2\,(1+K_{\text{Pax,Nkx}}\, N)^2}\right) - \beta_{\text{Pax}}\, P \\
f_O &= \alpha_{\text{Oli}} H^+\left(\frac{K_{\text{Oli,Pol}}\, c_{\text{Pol}}}{(1+K_{\text{Oli,Nkx}}\, N)^2\,(1+K_{\text{Oli,Irx}}\, I)^2}\,\frac{1+f_A\, K_{\text{Oli,Gli}} A}{1+K_{\text{Oli,Gli}}(A+R)}\right) - \beta_{\text{Oli}}\, O \\
f_N &= \alpha_{\text{Nkx}} H^+\left(\frac{K_{\text{Nkx,Pol}}\, c_{\text{Pol}}}{(1+K_{\text{Nkx,Pax}}\, P)^2\,(1+K_{\text{Nkx,Oli}}\, O)^2\,(1+K_{\text{Nkx,Irx}}\, I)^2}\,\frac{1+f_A\, K_{\text{Nkx,Gli}} A}{1+K_{\text{Nkx,Gli}}(A+R)}\right) - \beta_{\text{Nkx}}\, N \\
f_I &= \alpha_{\text{Irx}} H^+\left(\frac{K_{\text{Irx,Pol}}\, c_{\text{Pol}}}{(1+K_{\text{Irx,Oli}}\, O)^2\,(1+K_{\text{Irx,Nkx}}\, N)^2}\right) - \beta_{\text{Irx}}\, I
\end{aligned} \tag{Equation 45}$$

where $K_{X,Y}$ is the binding affinity of the TF/species Y onto its site on gene X, f_A is the binding cooperativity factor for Gli activator, c_{Pol} is the (constant) concentration of RNAp, α_X are the maximum production rates, and β_X the degradation rates.

The multiplicative noise terms, calculated from the CLE approximation, are

$$g_P = \left[\alpha_{\mathrm{Pax}}\, H^+\left(\frac{K_{\mathrm{Pax,Pol}}\, c_{\mathrm{Pol}}}{(1+K_{\mathrm{Pax,Oli}}\, O)^2\,(1+K_{\mathrm{Pax,Nkx}}\, N)^2}\right)+\beta_{\mathrm{Pax}}\, P\right]^{1/2}$$

$$g_O = \left[\alpha_{\mathrm{Oli}}\, H^+\left(\frac{K_{\mathrm{Oli,Pol}}\, c_{\mathrm{Pol}}}{(1+K_{\mathrm{Oli,Nkx}}\, N)^2\,(1+K_{\mathrm{Oli,Irx}}\, I)^2}\,\frac{1+f_A\, K_{\mathrm{Oli,Gli}} A}{1+K_{\mathrm{Oli,Gli}}(A+R)}\right)+\beta_{\mathrm{Oli}}\, O\right]^{1/2}$$

$$g_N = \left[\alpha_{\mathrm{Nkx}}\, H^+\left(\frac{K_{\mathrm{Nkx,Pol}}\, c_{\mathrm{Pol}}}{(1+K_{\mathrm{Nkx,Pax}}\, P)^2\,(1+K_{\mathrm{Nkx,Oli}}\, O)^2\,(1+K_{\mathrm{Nkx,Irx}}\, I)^2}\,\frac{1+f_A\, K_{\mathrm{Nkx,Gli}} A}{1+K_{\mathrm{Nkx,Gli}}(A+R)}\right)+\beta_{\mathrm{Nkx}}\, N\right]^{1/2} \quad \text{(Equation 46)}$$

$$g_I = \left[\alpha_{\mathrm{Irx}}\, H^+\left(\frac{K_{\mathrm{Irx,Pol}}\, c_{\mathrm{Pol}}}{(1+K_{\mathrm{Irx,Oli}}\, O)^2\,(1+K_{\mathrm{Irx,Nkx}}\, N)^2}\right)+\beta_{\mathrm{Irx}}\, I\right]^{1/2}$$

(i.e. the sum of production and degradation rates for the gene of interest, scaled by the inverse system size, under square root) multiplied by a standard Gaussian white noise, independent for each gene. See table below for the parameter values used.

The integration of Equations 44, 45, and 46 is performed via the Euler-Maruyama method, with time step $dt = 0.05$ (10 times smaller than the typical protein life time, given by the inverse degradation rates, $\beta^{-1} = 0.5$). For gene X, this is

$$X_{t+dt} = f_X dt + \sqrt{\Omega^{-1} dt}\, g_X\, w_t^X \quad \text{(Equation 47)}$$

where f_X and g_X are evaluated at time t, and where w_t^X is a normal-distributed random number with mean 0 and covariance $\mathbb{E}[w_t^X\, w_{t'}^Y] = \delta_{X,Y}\, \delta(t - t')$.

Parameters of the GRN model (dimensionality of the constants are indicated in the header to every section)		
Symbol	Meaning	Value
Concentrations ~ [conc]		
c_{Pol}	RNAp concentration	0.8
Binding affinities ~ $[\mathrm{conc}]^{-1}$		
$K_{\mathrm{Pax,Pol}}$	binding affinity of RNAp to Pax6	4.8
$K_{\mathrm{Oli,Pol}}$	binding affinity of RNAp to Olig2	47.8
$K_{\mathrm{Nkx,Pol}}$	binding affinity of RNAp to Nkx2.2	27.4
$K_{\mathrm{Irx,Pol}}$	binding affinity of RNAp to Irx3	23.4
$K_{\mathrm{Oli,Gli}}$	binding affinity of Gli to Olig2	18.0
$K_{\mathrm{Nkx,Gli}}$	binding affinity of Gli to Nkx2.2	373.0
$K_{\mathrm{Pax,Oli}}$	binding affinity of Olig2 to Pax6	1.9
$K_{\mathrm{Nkx,Oli}}$	binding affinity of Olig2 to Nkx2.2	27.1
$K_{\mathrm{Oli,Nkx}}$	binding affinity of Nkx2.2 to Olig2	60.6
$K_{\mathrm{Nkx,Pax}}$	binding affinity of Pax6 to Nkx2.2	4.8
$K_{\mathrm{Pax,Nkx}}$	binding affinity of Nkx2.2 to Pax6	26.7
$K_{\mathrm{Oli,Irx}}$	binding affinity of Irx3 to Olig2	28.4
$K_{\mathrm{Irx,Oli}}$	binding affinity of Olig2 to Irx3	58.8
$K_{\mathrm{Nkx,Irx}}$	binding affinity of Irx3 to Nkx2.2	47.1
$K_{\mathrm{Irx,Nkx}}$	binding affinity of Nkx2.2 to Irx3	76.2
Cooperativity coefficients and noise intensity ~ 1		
f_A	activation constant	10.0
Ω^{-1}	noise intensity	0.005
Degradation rates ~ $[\mathrm{time}]^{-1}$		
β_{Pax}	degradation rate of Pax6	2.0
β_{Oli}	degradation rate of Olig2	2.0
β_{Nkx}	degradation rate of Nkx2.2	2.0

(*Continued on next page*)

Continued

Symbol	Meaning	Value
β_{Irx}	degradation rate of Irx3	2.0
Production rates $\sim$ [conc][time]$^{-1}$		
α_{Pax}	maximum production rate of Pax6	2.0
α_{Oli}	maximum production rate of Olig2	2.0
α_{Nkx}	maximum production rate of Nkx2.2	2.0
α_{Irx}	maximum production rate of Irx3	2.0

Dynamics of a stochastic gradient

In the patterning task, we also include a dynamics for the morphogen gradient. We simulate a non-stationary stochastic field $\widehat{S}_{x,t}$, as the empirical number density field $\widehat{S}_{x,t} = \sum_i \delta(\widehat{X}^i_t - x)$ associated to a stochastic reaction-diffusion with.

$$\mathrm{d}\widehat{X}^i_t = \sqrt{2D}\,\mathrm{d}W^i_t \qquad \text{(Equation 48)}$$

and where particles are removed with independent rates κ and added at x_0 with rate J_0. The SDE in Equation 48 provides an explicit method to simulate the spatio-temporal dynamics of the stochastic field $\widehat{S}_{x,t}$. To do so, we simulate trajectories of Equation 48 via, e.g. Euler-Maruyama method, with time discretisation dt, that is

$$X^i_{t+\mathrm{d}t} = X^i_t + \sqrt{2D\mathrm{d}t}\,w^i_t \qquad \text{(Equation 49)}$$

with w^i_t a normal-distributed random number with mean 0 and covariance $\mathbb{E}[w^i_t\, w^j_{t'}] = \delta_{i,j}\,\delta(t - t')$; in the time step between t and $t + \mathrm{d}t$, each particle is eliminated with probability $\kappa\mathrm{d}t$, and a burst of n_b new particles is added at $x_0 < 0$ with probability $J_0\mathrm{d}t/\,n_b$ (so that J_0 is the overall average production rate, but with burst size n_b). The number density field can be then defined with a spatial resolution dx, as the count of the number of particles within $[x - \mathrm{d}x/2, x + \mathrm{d}x/2]$, divided by d$x$. The resolution d$x$ is chosen to be the single-cell size.

We set the parameters of the dynamics as follows. 81 cells are aligned along one axis within $[0, 1]$, so $\mathrm{d}x = 1/80$. The time discretization dt is chosen as 5 times smaller than that for the PONI network, but configurations are taken every 5 steps. The free parameters of the dynamics must set a time scale, a length scale and a typical number of particles. We set the overall time scale of the process through the degradation rate κ. The length scale is the decay length λ of the average gradient profile at steady state, $\mathbb{E}\left[\widehat{S}_{x,t\to\infty}\right] \propto \exp(-|x - x_0|\lambda)$. This is fixed to 0.15 in all simulations, consistently with experimental measures.[16] This decay length can be derived analytically to be $\lambda = \sqrt{D/\kappa}$, from which we fix the diffusion constant accordingly to be $D = \kappa\,\lambda^2$. The typical density is chosen to be the average number density at $x = 0$ at steady state, which is $N_0 = J_0\,\mathrm{e}^{-|x_0|}/2\kappa\lambda$. With a fixed burst rate $r = J_0/\,n_b = 50$, we modulate the burst size n_b by inverting the expression for N_0.

The ensemble average of the field $S = \mathbb{E}[\widehat{S}]$, satisfies the PDE

$$\partial_t S - D\nabla^2 S + \kappa S = J_0\,\delta(x - x_0) \qquad \text{(Equation 50)}$$

By integrating the spatial part, we can write

$$\partial_t S = J_0\,\frac{\exp\left(-\kappa t - \dfrac{(x - x_0)^2}{4Dt}\right)}{\sqrt{4\pi D t}}. \qquad \text{(Equation 51)}$$

In Equation 51, the spatial variable enters only parametrically and the dynamics can be described as an ODE with time-dependent production rates. Therefore, (ensemble) averages of the signal experienced at different spatial locations can be regarded as "independent", but at the expense of allowing non-autonomous dynamics for the local signal.

Parameters used for the simulations in this work are $\lambda = 0.15$ (in units of D-V axis length), $\kappa = 0.5$ (equal to $\beta/4$ – See the table above, and $N_0 = 5000$.

Multi-Agent control

Here we derive the Bellman equation for the multi-agent (MA) case. The equations are written for the discrete-time and discrete-state case – as it is more transparent for a reinforcement learning implementation – but are easily generalized to continuous space and/or time. The notation is explained in the below table.

Notation for multi-agent reinforcement learning (in parenthesis, the biological interpretation of the variables)

Symbol	Meaning	Value
Dynamics		
N	number of agents (simulated cells)	81
$\bullet$	multi-agent (tissue-level) variable, e.g. $\overline{x} = \{x_i\}_{i=1}^{N}$	–
i	agent (cell) index	1…N
$x_i, \overline{x}$	state variable (gene expression, extracellular signal levels, memory variables)	–
$u_i, \overline{u}$	action variable (intracellular effectors, memory variables production rates)	–
p	single-agent transition probability (single-cell stochastic dynamics)	–
P	multi-agent transition probability (tissue-level stochastic dynamics)	–
π	single-agent policy (single-cell control strategy)	–
Π	multi-agent policy (tissue-level control strategy)	–
Objective function		
τ	exponential discounting time (differentiation rate)	5 (A.U.)
$\xi_i, \overline{\xi}$	target point (gene expression)	–

Full multi-agent case

The multi-agent probability distribution at time t, $\rho_t(\overline{x})$, satisfies the forward Kolmogorov equation

$$\rho_{t+1}(\overline{x}) = \sum_{\overline{x}',\overline{u}'} P(\overline{x}|\overline{x}',\overline{u}')\Pi(\overline{u}'|\overline{x}')\,\rho_t(\overline{x}') \qquad \text{(Equation 52)}$$

The goal of the agents is to maximize the expectation value of the discounted return (in the decision-making and reinforcement learning literature, it is more customary to express the goal in terms of maximisation of rewards, rather than minimisation costs):

$$R_t = \sum_{t'=0}^{\infty} \gamma^{t'}\, r_{t+t'} \qquad \text{(Equation 53)}$$

with

$$r_t = r\left(\overline{x}^t, \overline{u}^t\right) \qquad \text{(Equation 54)}$$

In the end, we will be interested in a reward of the form

$$r(\overline{x},\overline{u}) = -q_{\overline{\xi}}(\overline{x}) - \frac{\epsilon}{2}\,\|\,\overline{u}\|^2 \qquad \text{(Equation 55)}$$

where, e.g. $q_{\overline{\xi}}(\overline{x}) = \|\,\overline{x} - \overline{\xi}\|^2/2$. This negative reward is a cost that penalises certain configurations of the MA system –implementing the requirement to reach the target- and high values of control.

The objective function $\mathcal{J}_\Pi = \mathbb{E}_\Pi[R_0]$, that is the ensemble average of R_0 over the trajectories generated by the policy Π, writes

$$\begin{aligned}\mathcal{J}_\Pi &= \sum_t \gamma^t \sum_{\overline{x},\overline{u},\overline{x}'} P(\overline{x}'|\overline{x},\overline{u})\Pi(\overline{u}|\overline{x})\,\rho_t(\overline{x})\,r(\overline{x},\overline{u}) \\ &= \sum_{\overline{x},\overline{u},\overline{x}'} P(\overline{x}'|\overline{x},\overline{u})\Pi(\overline{u}|\overline{x})\,\eta(\overline{x})\,r(\overline{x},\overline{u})\end{aligned} \qquad \text{(Equation 56)}$$

where η is the discounted occupancy

$$\eta(\overline{x}) = \sum_{t=0}^{\infty} \gamma^t \rho_t(\overline{x}) \qquad \text{(Equation 57)}$$

We can introduce the quality (or state-action value) function, which is the expectation value of the return conditioned on the initial state and action, $Q^t_\Pi(\overline{x},\overline{u}) = \mathbb{E}[R_t|\overline{x}^t = \overline{x},\overline{u}^t = \overline{u}]$. We can write a recursive equation of the value function Q^t_Π, expressing the conditional expectation value $\mathbb{E}[R_t|\overline{x},\overline{u}]$ by making use of Equation 52:

$$Q^t_\Pi(\overline{x},\overline{u}) = \sum_{\overline{x}'} P(\overline{x}'|\overline{x},\overline{u}) \left\{ r(\overline{x},\overline{u}) + \gamma \sum_{\overline{u}'} \Pi(\overline{u}'|\overline{x}')\, Q^{t+1}_\Pi(\overline{x}',\overline{u}') \right\}. \quad \text{(Equation 58)}$$

Since there is no finite horizon and neither rewards nor transition probabilities depend explicitly on time, we can seek for a stationary solution $Q^t_\Pi = Q_\Pi$.

The principle of dynamic programming[38,60] consists in maximizing the expected return –i.e. the objective function $\mathcal{J}_\Pi$– by maximizing its conditional expectation at intermediate times, that is the value function. The optimal policy Π^*, then, is given in terms of the quality function as

$$\Pi^*(\overline{u}|\overline{x}) = \delta_{\overline{u},\overline{u}^*(\overline{x})}, \text{ with } \overline{u}^*(\overline{x}) = \underset{\overline{u}}{\operatorname{argmax}} Q^*(\overline{x},\overline{u}) \quad \text{(Equation 59)}$$

where the optimal quality function satisfies the Bellman equation

$$Q^*(\overline{x},\overline{u}) = \sum_{\overline{x}'} P(\overline{x}'|\overline{x},\overline{u}) \left\{ r(\overline{x},\overline{u}) + \gamma \max_{\overline{u}'} Q^*(\overline{x}',\overline{u}') \right\}. \quad \text{(Equation 60)}$$

Independent agents

To reflect the requirement of each agent individually to reach their own target, we write $q_{\overline{\xi}}(\overline{x}) = \sum_i q_{\xi_i}(x_i)$, where q_ξ is some convex function that has a minimum at ξ. This is true for the cost rate $q_{\overline{\xi}}(\overline{x}) = \| \overline{\xi} - \overline{x} \|_2^2 = \sum_i \| \xi_i - x_i \|_2^2$. So, the instantaneous reward for the MA system is the sum of rewards for the individual agents, c_i, that are functions of the single agent's observations and actions:

$$r_i(x,u) = -q_{\xi_i}(x) - \frac{\epsilon}{2} \| u \|^2 \quad \text{(Equation 61)}$$

As discussed above, the MA policy Π with respect to which we want to optimize the performance is of the form

$$\Pi(\overline{u}|\overline{x}) = \prod_{i=1}^{N} \pi(u_i|x_i) \quad \text{(Equation 62)}$$

that is, actions by individual agents are chosen independently according to the same single-agent policy π. We seek for solutions of the Bellman equation of the form

$$Q^t_\Pi(\overline{x},\overline{u}) = \sum_{i=1}^{N} Q^t_\pi(x_i,u_i). \quad \text{(Equation 63)}$$

By replacing Equations 62 and 63, into the Bellman Equation 58, we have

$$\sum_{\overline{x}'} P(\overline{x}'|\overline{x},\overline{u}) \sum_i \left\{ r(x_i,u_i) + \gamma \sum_{u_i'} \pi(u_i'|x_i')\, Q^{t+1}_\pi(x_i',u_i') - Q^t_\pi(x_i,u_i) \right\} = 0\,. \quad \text{(Equation 64)}$$

Optimality, in this approximation, is

$$\pi^*(\cdot|x) = \delta_{u,u^*(x)}, \text{ with } u^*(x) = \underset{u}{\operatorname{argmax}} Q^*(x,u) \quad \text{(Equation 65)}$$

where Q^* denotes the optimal quality function solving

$$\sum_{\overline{x}'} P(\overline{x}'|\overline{x},\overline{u}) \sum_i \left\{ r(x_i,u_i) + \gamma \max_{u'} Q^*(x_i',u') - Q^*(x_i,u_i) \right\} = 0\,. \quad \text{(Equation 66)}$$

This is approximately solved by minimizing the expectation of the square MA error

$$\Delta_Q(\overline{x}',\overline{x},\overline{u})^2 = \sum_i \left\{ r(x_i,u_i) + \gamma \max_{u'} Q(x_i',u') - Q(x_i,u_i) \right\}^2 \quad \text{(Equation 67)}$$

with respect to the Q,

$$Q^* \simeq \underset{Q}{\operatorname{argmin}} \sum_{\overline{x}'} P(\overline{x}'|\overline{x},\overline{u})\, \Delta_Q(\overline{x}',\overline{x},\overline{u})^2. \quad \text{(Equation 68)}$$

Memory in signal interpretation

The independent-agent ansatz is exact when the transition probabilities $P(\overline{x}'|\overline{x},\overline{u})$ can be factorized into single-agent transition probabilities

$$P(\overline{x}'|\overline{x},\overline{u}) = \prod_{i=1}^{N} p_i(x_i'|x_i,u_i), \quad \text{(Equation 69)}$$

that is, when the dynamics of each agent is independent of other agents. This can be seen intuitively for a static and deterministic gradient, but also for a stochastic gradients such as in Petkova et al.,[46] where no dynamics in the morphogen is taken into account. In such cases, the constant (statistics of the) morphogen signal at the location of a given cell enters as a parameter in the quality function Q: it's role is to "select" the specific single-agent problem for that particular cell. This effectively makes the MA task trivially decomposed into single-agent ones. In general, when the morphogen gradient is modelled as a diffusion-degradation process –as in this case– this approximation is not valid. The limit of a stationary gradient could be recovered only when the morphogen concentration profile reaches a steady state fast enough (high κ) compared to other system variables.

One can show that the average of the concentration field over the noise, $S = \mathbb{E}[\widehat{S}]$, can be calculated as the solution of independent differential equations with local time-dependent rates (see Equation 51). So, even though we may be able to express the average dynamics of the morphogen at individual cells locations as independent, 1) fluctuations will anyway be correlated and 2) we do so at the cost of introducing time dependence.

Here, we assume that it is possible to approximate the transition probability P by a factorized form as in Equation 69, at the expense of introducing auxiliary variables $\{M_h\}_{h=1}^{N_{\text{mem}}}$, included in the "state" of the single cell along with its gene expression G and the local (stochastic) morphogen signal $\widehat{S}$. These memory variables integrate over time the extracellular signal S and model the effective memory. We model these as the species in a signalling cascade, whereby $\widehat{S}$ directly influences the production of M_1, which in turn affects production of M_2 etc.,

$$\begin{aligned} \tau_M \frac{dM_1}{dt} &= r_1 \widehat{S} - M_1 \\ \tau_M \frac{dM_h}{dt} &= r_h M_{h-1} - M_h, \ \text{for } h>1 \end{aligned} \quad \text{(Equation 70)}$$

where r_h are components of the control vector u, therefore functions of the single cell state variables – bound between $\pm$ 1. While the coefficients $\{r_h\}_h^{N_{\text{mem}}}$ contribute to a generally non-linear dynamics in signal interpretation –possibly integrating feedback from any of the GRN components– the linear dependence of the production rate of M_h on M_{h-1} represents a feed-forward backbone architecture. In this work, this architecture is motivated by the experimental knowledge about the Shh signalling pathway

We choose the overall time constant $\tau_M = 1$; this value has been chosen to be comparable with the typical timescales of the GRN (see table in the section ventral neural-progenitor GRN model (PONI network), reflecting the experimental observation about the dynamics of intermediate species in the Shh signalling.[16]

RL solution

The approximate solution of Equation 64 via reinforcement learning (RL) requires the sampling of the tuples $(\vec{x}^t,\vec{u}^t,r^t,\vec{x}^{t+1})$. State-of-the-art deep-RL algorithms — such as DQN,[61] DDPG,[65] TD3,[41] SAC[66] etc— solve the problem of the stability of learning by storing a replay buffer $\mathcal{B}$ with the last N_{replay} tuples visited, and estimating gradients of the loss functions by averaging over a small number N_{batch} (batch size) of them.

Here we use TD3,[41] which is an actor-critic deep-RL algorithm, designed for continuous control problems. Similar to other actor-critic algorithm, it stores function approximators for both the policy (actor), and the value (critic) function. These are represented by deep neural networks with parameters φ and θ, respectively ($\pi \simeq \pi_\varphi$ and $Q \simeq Q_\theta$).

In order to reduce the bias in the estimate of the value function Q, TD3 uses two critics (T for "twin"). In standard Q-learning, the value of the state after the transition is taken to be the maximum over all actions of the Q function evaluated at that state, by boostrapping. This is a problem that is present also in actor-critic algorithms like DDPG, where the "maximization over actions" is implicit in the policy-gradient formula, which typically leads to an overestimation of the value (as demonstrated in the paper), and therefore to sub-optimal policies.

As in other deep-RL AC algorithms, in order to make learning more stable, TD3 stores two copies of each function approximator: the first is updated on-line; the second is used as target and integrates the first at a slow rate, and with delay. TD3 uses a SARSA-like target for the value function, by sampling the next action using the target policy.

We here use the TD3 algorithm for episodic tasks (see Fujimoto et al.[41] for details). We use $\alpha = 10^{-3}$, $\beta = 10^{-3}$. All other details are the same as in the original paper. The discount factor (which is a property of the task!) $\gamma = 0.99$, which for time step $dt = 0.005$ corresponds to the exponential discount time in continuous time $\tau \simeq 5$.

In the following, the notation $\langle\ \rangle_{\text{batch}}$ indicates a sample mean over a batch.

Algorithm 1. Twin Delayed Deep Deterministic (TD3) policy gradient for episodic tasks.

Initialize actor and critic networks with parameters φ, θ_1 and θ_2

Initialize target networks: $\varphi' \leftarrow \varphi$, $\theta_1' \leftarrow \theta_1$ and $\theta_2' \leftarrow \theta_2$

Initialize replay buffer $\mathcal{B}$
Define exploration parameters σ, regularization parameter $\tilde{\sigma}$, target learning rate τ, and optimizers learning rates α and β
for N_{ep} episodes **do**:
  Initialize agent in state $x^0 \sim \rho_0$
  for $t = 0 \ldots T-1$ (T cutoff time) or until terminal state **do**
    Select control, $u^t = \pi_\varphi(x^t) + \epsilon$, with exploration noise $\epsilon \sim \mathcal{N}(0, \sigma)$
    Observe reward r^t and new state x^{t+1}
    Store the tuple (x^t, u^t, r^t, x^{t+1}) in the buffer $\mathcal{B}$
    Sample N_{batch} random tuples (x, u, r, x')
    For each of these, compute target $y \leftarrow r + \gamma \min_{i \in \{1,2\}} Q_{\theta_i'}(x', u')$, where $u' = \pi_{\varphi'}(x') + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \tilde{\sigma})$
    Update the critic networks ("$\leftarrow_\alpha$" indicates gradient-based optimizer step with learning rate α):
      $\theta_i \leftarrow_\alpha \nabla_{\theta_i} \langle (y - Q_{\theta_i}(x, u))^2 \rangle_{\text{batch}}$
    if episode multiple of d (delay) **then**
      Update on-line policy network with deterministic policy gradient:
      $\varphi \leftarrow_\beta \nabla_\varphi \langle \nabla_{u'} Q_{\theta_1}(x, u')|_{u' = \pi_\varphi(x)} \nabla_\varphi \pi_\varphi(x) \rangle_{\text{batch}}$
      Update the target networks:
      $\varphi' \leftarrow (1 - \tau)\varphi' + \tau\, \varphi$
      $\theta_i' \leftarrow (1 - \tau)\, \theta_i' + \tau\, \theta_i$

In the case of the MA problem described above, we need to modify this algorithm by storing transitions of the MA system, defining a target for each individual agent (based on their single-agent rewards, states and actions), and averaging gradients over the agents as well. This is detailed in algorithm 2. The learning rates here are $\alpha = 3 \times 10^{-5}$ and $\beta = 10^{-5}$.

Algorithm 2. Multi-Agent Twin Delayed Deep Deterministic (TD3) policy gradient for episodic tasks
Initialize actor and critic networks with parameters φ, θ_1 and θ_2
Initialize target networks: $\varphi' \leftarrow \varphi$, $\theta_1' \leftarrow \theta_1$ and $\theta_2' \leftarrow \theta_2$
Initialize replay buffer $\mathcal{B}$
Define exploration parameters σ, regularization parameter $\tilde{\sigma}$, target learning rate τ, and optimizers learning rates α and β
for N_{ep} episodes **do**:
  Initialize the N agents in state $\overline{x}^0 \sim \rho_0$
  for $t = 0 \ldots T-1$ (T cutoff time) or until terminal state **do**
    Select control, $\overline{u}^t = \pi_\varphi(\overline{x}^t) + \epsilon$, with exploration noise $\epsilon \sim \mathcal{N}(0, \sigma)$
    Observe reward r^t and new state $\overline{x}^{t+1}$
    Store the tuple $(\overline{x}^t, \overline{u}^t, \overline{r}^t, \overline{x}^{t+1})$ in the buffer $\mathcal{B}$
    Sample N_{batch} random tuples $(\overline{x}, \overline{u}, \overline{r}, \overline{x}')$
    For each of these, and for each agent j,
    compute targets $y_j \leftarrow r_j + \gamma \min_{i \in \{1,2\}} Q_{\theta_i'}(x_j', u_j')$, where $u_j' = \pi_{\varphi'}(x_j') + \epsilon_j$, with $\epsilon_j \sim \mathcal{N}(0, \tilde{\sigma})$
    Update the critic networks
      $\theta_i \leftarrow_\alpha \nabla_{\theta_i} \langle N^{-1} \sum_{j=1}^{N} (y_j - Q_{\theta_i}(x_j, u_j))^2 \rangle_{\text{batch}}$
    if episode multiple of d (delay) **then**
      Update on-line policy network with deterministic policy gradient:
      $\varphi \leftarrow_\beta \nabla_\varphi \langle N^{-1} \sum_{j=1}^{N} \nabla_{u'} Q_{\theta_1}(x_j, u')|_{u' = \pi_\varphi(x_j)} \nabla_\varphi \pi_\varphi(x_j) \rangle_{\text{batch}}$
      Update the target networks:
      $\varphi' \leftarrow (1 - \tau)\varphi' + \tau\, \varphi$
      $\theta_i' \leftarrow (1 - \tau)\, \theta_i' + \tau\, \theta_i$