

UNIVERSITÉ PARIS CITÉ

École doctorale de Sciences Mathématiques de Paris Centre (ED 386)

Laboratoire MAP5, UMR 8145 CNRS

**Au delà de la diffusion : inférence d'un processus de
différenciation basée sur un modèle de gène réaliste et la
résolution du problème de Schrödinger**

par CLÉMENCE FOURNIÉ

Thèse de doctorat de Mathématiques Appliquées

Dirigée par FABIEN CRAUSTE et OLIVIER GANDRILLON

Soutenue publiquement le 29 septembre 2026 devant un jury composé de :

FABIEN CRAUSTE	DR CNRS, Université Paris Cité	Directeur
CORALIE FRITSCH	CR HDR, Inria Nancy	Rapportrice
OLIVIER GANDRILLON	DR CNRS, ENS de Lyon	Directeur
SILVIA GRIGOLON	CR CNRS, Sorbonne Université	Examinatrice
GURVAN HERMANGE	MCF, CentraleSupélec	Examineur
NATALIYA SOKOLOVSKA	PR, Sorbonne Université	Rapportrice
AMANDINE VEBER	DR CNRS, Université Paris Cité	Présidente

Université Paris Cité - MAP5

45 rue des Saints-Pères

75006 Paris



Abstract

Cell differentiation is a dynamic biological process that leads a cell to acquire a specific cellular identity. Recently, single-cell RNA sequencing has enabled simultaneous measurement of gene expression at specific points in this process across a large number of cells and genes. Repeating these measurements at different times provides insight into the temporal variation of a distribution in the gene expression space. Since most measurement techniques are destructive, the goal of inferring cell differentiation trajectories is to reconstruct the differentiation process at the level of each cell in the population. Existing work assumes that cells evolve, within the gene expression space, via diffusion. However, the actual dynamics of gene expression are far more complex.

In my PhD thesis, I developed a method for inferring a cell differentiation trajectory from single-cell RNA sequencing data, based on the solution to the Schrödinger problem and a mechanistic, stochastic gene model that generates realistic dynamics. The method relies on constructing a reference process for the Schrödinger problem based on the gene model applied to a gene regulatory network, known as the Bursty model.

By comparing the solutions to the Schrödinger problem obtained with a diffusive reference process and a reference process based on the Bursty model, I established the validity of this new method, first using simulated data and a simple gene model, and then using experimental data and a complex gene network. This work thus illustrates the value of integrating mechanistic knowledge of the biological system into the inference method to enable a more realistic and, ultimately, more relevant inference of cell differentiation.

Keywords: Trajectory inference, gene expression model, gene regulatory network, Schrödinger problem, cell differentiation, single-cell data.

Résumé

La différenciation cellulaire est un processus biologique dynamique qui conduit une cellule à acquérir une identité cellulaire particulière. Récemment, le séquençage d'ARN en cellules uniques a permis de mesurer simultanément les niveaux d'expression génique à des instants précis de ce processus pour un grand nombre de cellules et de gènes. La répétition de ces mesures à différents moments permet d'accéder à la variation temporelle d'une distribution dans l'espace d'expression génique. La plupart des techniques de mesure étant destructives, l'inférence de trajectoires de différenciation cellulaire vise à reconstituer le processus de différenciation au sein de chaque cellule de la population. Les travaux existants supposent que les cellules évoluent dans l'espace d'expression génique par diffusion. Or, la dynamique réelle des gènes est bien plus complexe.

Dans mon travail de thèse, j'ai développé une méthode pour inférer une trajectoire de différenciation cellulaire à partir de données issues de séquençage d'ARN en cellules uniques, en m'appuyant sur la résolution du problème de Schrödinger et un modèle de gènes mécaniste et stochastique générant des dynamiques réalistes. La méthode repose sur la construction d'un processus de référence pour le problème de Schrödinger à partir du modèle de gènes appliqué à un réseau de régulation génique, appelé modèle Bursty.

En comparant les solutions du problème de Schrödinger obtenues avec un processus de référence diffusif et un processus de référence basé sur le modèle Bursty, j'ai établi la pertinence de cette nouvelle méthode, d'abord à partir de données simulées et d'un modèle de gènes simple, puis à partir de données expérimentales et d'un réseau de gènes complexe. Ce travail illustre donc l'intérêt d'intégrer une connaissance mécaniste du système biologique à la méthode d'inférence pour permettre une inférence plus réaliste et, *in fine*, plus pertinente de la différenciation cellulaire.

Mots-clés : Inférence de trajectoire, modèle d'expression génique, réseau de régulation génique, problème de Schrödinger, différenciation cellulaire, données en cellules uniques.

Remerciements

Merci aux rapportrices, Coralie FRITSCH et Nataliya SOKOLOVSKA, d'avoir accepté de rapporter ma thèse pendant leurs vacances.

Merci aux membres du jury d'avoir accepté : Amandine VEBER, Silvia GRIGOLON, Gurvan HERMANGE, Paul VILLOUTREIX, Coralie FRITSCH et Nataliya SOKOLOVSKA.

Merci à mes directeurs de thèse, Olivier GANDRILLON, pour ta patience, tes conseils et ta franchise, et je suis honorée d'avoir été ta dernière doctorante, et à Fabien CRAUSTE, pour ton écoute, ton temps, ton accompagnement, ton honnêteté et ton soutien.

Je suis tellement reconnaissante d'avoir été encadrée par vous : votre aide et votre optimisme dans les moments difficiles, ainsi que votre capacité à valoriser les moments joyeux. Je ne pourrai jamais vous remercier assez.

Merci à Laura CANTINI et Julie DELON d'avoir fait partie de mon CSI. Merci pour votre bienveillance et votre accompagnement.

Merci au MAP5 dans son entièreté de m'avoir accueillie, suivie et soutenue pendant ces trois ans. Merci à Anne ESTRADÉ et Antoine CHAMBAZ, qui se sont succédé.e.s à la direction, pour votre accompagnement, vos gourmandises et la gazette. Merci au personnel administratif du MAP5 et de l'UFR, en particulier à Gladys BERNARI pour son aide dans les batailles administratives du CNRS. Merci à l'équipe Modélisation ou MAS, la meilleure équipe en toute objectivité. Merci au comité parité pour les échanges, les débats et tous ces super projets. Merci à tous les habitué.e.s du déjeuner dans la cuisine/terrasse. Merci à la terrasse.

Merci évidemment à tou.te.s les "éphémères", les "ancien.ne.s" parti.e.s trop tôt : Zoé, Ariane, Sonia, Mariem, Mehdi, Diala, Yacine, Keanu, Eloi et Ivan ; la promo 2023 : Adélie, Béatriz, Lucie, Rayane et Ariel ; les membres très selectif.ve.s du bureau 814 (et une lettre dont je ne me rappelle jamais) : Béatriz, Perla, Eyal, César, Timothé et Valentina ; et tous les autres : Tom, Baptiste, Guillaume, Pierre, Hugo, ... Merci à Discord, au tarot, aux sorties, aux goûters...

Merci à l'équipe de l'IUT informatique de l'IUT Paris Rives de Seine de m'avoir permis de vivre cette formidable expérience d'enseignement : Véronique BEDOS, Mourad OUZIRI, Sourat GHELAB et, surtout, l'équipe de maths : Nathalie RODRIGUEZ, Berenice GREC, Anne ESTRADÉ, Sébastien MARTIN et Guillaume GAVRA.

Merci pour votre accueil, votre communication, votre soutien et votre bonne humeur.

Merci au LBMC pour son accueil chaleureux et ses discussions intéressantes.

Merci à mes collaborateurs Elias VENTRE, Ulysse HERBACH et Aymeric BARADAT.

Merci à la FP2M et ses organisateurs.

Merci à tou.te.s mes collègues devenu.e.s ami.e.s.

Merci aux organisateurs des conférences auxquelles j'ai pu assister, notamment les JMBS $\times 3$, le *Waddington Workshop*, la *Bonn Conference on Mathematical Life Sciences* et l'ECMTB. Merci à ma partenaire de conférence, Chloé.

Merci à Schrödinger, à Schiebinger et à tou.te.s les autres scientifiques qui me permettent de mener cette thèse aujourd'hui.

Merci à Alessandro MASIERO et à l'équipe de bioinformatique de Sanofi pour la confiance qu'ils m'ont accordée, ainsi que pour les connaissances et les opportunités.

Thank to Lauren WEISS Lab for the opportunities. Thank you Misha, Bhavika and Zoe.

Merci à Gurvan HERMANGE pour m'avoir encadrée et initiée à la recherche en biomathématiques. Merci de m'avoir donné ma chance, de m'avoir appris tant de choses et de m'avoir soutenue.

Merci à l'équipe du MICS. Merci, Karmen.

Merci à mes professeurs de collège et de lycée de maths : M. JADOT, M. CLASTRES et Mme SAINT-JEAN ; de SVT : Mme RINGHAUSEN et Mme NAHON ; et de grec : Mme LUDVIG.

Merci à Alexis FREITAS.

Merci à Charlotte DION, Marie DOUMIC (pour ses TD d'EDO) et Marie POSTEL (pour mon premier projet maths-bio), mes premières exemples de chercheuses en maths, qui plus est en maths appliquées à la biologie.

Merci au chargé de TP de la mineure biologie.

Merci à l'équipe de L3 et du master de bioinformatique de Lyon 1, surtout Anne-Béatrice DU-FOUR, Christelle LOPES, Laurent PUJO-MENJOUET et Arnaud MARIE, pour leur humanité et leurs cours.

Merci à Laurence RODER et Florence HUBERT pour l'existence du master CMB et pour m'avoir permis de mêler la biologie et les mathématiques dans tous leurs projets, dans ce cadre beau et chaleureux. Merci à Charlotte PERRIN.

Merci au cours du Collège de France d'Édith HEARD sur l'épigénétique.

Merci à ma famille de la "Côte de Castelferrus".

Merci à ma marraine Marion et Jean-Paul.

Merci à la famille TOULLIER.

Merci à mon Papi et Mami pour TOUT : l'amour, l'accueil, les légumes, les anecdotes, les petits déjs, les tentatives de compréhension de tout ce que je fais...

Merci à mes parents d'amour.

Merci, Antoine, d'être le meilleur frère de l'univers.

Merci aux Z-woks et à mes tantes et tontons de Roquemaure de m'avoir suivie et épaulée dans la joie, sous le soleil comme sous la neige. Un mot particulier pour Lucille, qui m'a aidée à me lancer dans la recherche avec nos discussions after-work. Merci.

Merci, Manon. Que et qui serais-je sans toi ? Merci d'être ma meilleure amie. Je sais que je peux toujours compter sur ton soutien et ta franchise. Merci de m'accompagner depuis si longtemps, au fil de tant d'émotions.

Merci à Charlotte, ma meilleure amie, ma boss, ma championne, pour nos réflexions ivres, pour le sport, pour le soutien incontestable, pour les fous rires et pour les vocaux/podcasts interminables.

Manon et Charlotte, la vie est tellement plus belle à vos côtés.

Merci à mon amie Jeanne pour les parties pyjama, pour son amitié et aussi pour le graphisme de cette thèse.

Merci à mon amie Lisa pour les sorties découvertes.

Pour Vincent : mon premier est une grande étendue d'eau, mon second est "oui" en espagnol et mon tout a bien sa place dans cette section.

Merci à Igor, Léo et Seki.

Merci Celine et Ambre.

Merci à la team BISM : Ravy, Martial, Victor, Thomas, Anna et les autres pour les cours en visio à la BU, les couvre-feux, les jeux, la réouverture des bars et les autres aventures. Un merci particulier à Ravy, mon pote depuis mercredi, pour nos *dates* parisiens.

Merci, Isis et Morgane.

Merci à mes collègues de CMB, surtout Léa, Émilie, Piyush et Louise, pour l'Idéfix, les randos et les éclats de rire.

Merci à la colo de Nantes, Aloïs, Louis, Marion, Chloé et Émilie, pour notre escapade maritime.

Merci aux copaines de VV+ et à l'asso, en tant qu'entité, d'exister et de prendre si bien soin des humain.e.s, en particulier : Thalia, Alice, Clémence C., Ariane, Yarong, Bertille, Léa R., Cindy, Fée, Morgane, Charlie toustes les coaches, toustes les membres des bureaux et les voisines du 14, Clem ×2 et Diane.

Merci à la famille POUX au complet et élargie, (coucou Pascaline), cœur cœur paillette sur vous.

Merci au sport : surtout mes chaussures orange et mes genouillères.

Merci aux activités manuelles de tout genre.

Merci au ventilateur (il fait actuellement 8000 degrés).

Merci à ma psy, évidemment.

Merci à mon amoureux : pas assez de pages pour exprimer ma gratitude et mon amour pour toi ; je ne veux pas étaler notre amour comme ça. Merci surtout de m'avoir supportée, dans les deux sens du terme.

Bref, grâce à vous tou.te.s, j'en suis là aujourd'hui ; je n'y serais pas arrivée sans vous, du moins sûrement pas dans le même état. Vous êtes nécessaires et suffisants <3.

PS1 : navrée si je t'ai oublié, il fait actuellement 8000 degrés.

PS2 : navrée si j'ai écorné ton prénom et/ou ton nom, je ne sais pas écrire, c'est même pas la faute de la canicule.

Déclaration d'utilisation d'outils IA

Les outils d'IA, ChatGPT, Claude et le Chat ont été utilisés parfois afin d'améliorer le texte, les figures ou des parties de code. Ce manuscrit n'a pas été rédigé par une IA et les résultats n'ont pas non plus été générés par une IA.

Abréviations et Notations

Abréviations

A	- Adénine
ACP ou PCA	- Analyse en composantes principales
ADN	- Acide désoxyribonucléique
ADNc	- Acide désoxyribonucléique complémentaire
ARN	- Acide ribonucléique
ARNm	- Acide ribonucléique messager
C	- Cytosine
EDO	- Équation différentielle ordinaire
EDS	- Équation différentielle stochastique
EMD	- Earth Mover's Distance ou distance de Wasserstein
GRN	- <i>Gene Regulatory Network</i> ou réseau de régulation génique
G	- Guanine
NK	- Natural killer
OT	- <i>Optimal Transport</i> ou Transport Optimal
PCR	- <i>Polymerase chain reaction</i>
PDMP	- <i>Piecewise deterministic Markov chain</i> ou processus de Markov déterministe par morceaux
RA	- Stimulus du GRN
RT-qPCR	- <i>Reverse transcription quantitative polymerase chain reaction</i>
ScRNA-seq	- <i>Single Cell RNA sequencing</i> ou séquençage en cellule unique
T	- Thymine
U	- Uracile

Notations

β_i	- taux de l'activité basal du gène $i \in \mathbb{N}^G$ dans le modèle Bursty
$d_{0,i}$	- taux de dégradation de l'ARNm du gène $i \in \mathbb{N}^G$ dans le modèle Bursty
$d_{1,i}$	- taux de dégradation de la protéines du gène $i \in \mathbb{N}^G$ dans le modèle Bursty
G	- nombre de gènes étudiés
H	- entropie relative ou divergence de Kullback-Leibler
$k_{on,i}$	- taux d'activation du saut stochastique pour le gène $i \in \mathbb{N}^G$ dans le modèle Bursty
$k_{0,i}$	- fréquence minimale du saut stochastique pour le gène $i \in \mathbb{N}^G$ dans le modèle Bursty
$k_{1,i}$	- fréquence maximale du saut stochastique pour le gène $i \in \mathbb{N}^G$ dans le modèle Bursty
μ	- distribution initiale des cellules
μ_{exp}	- distribution initiale expérimentale des cellules
m	- nombre de cellules étudiées dans ν
$M_i(t)$	- quantité d'ARNm du gène $i \in \mathbb{N}^G$ dans le modèle Bursty
Π	- un espace discret et fini
Ψ	- un espace discret et fini
ν	- distribution finale des cellules
ν_{exp}	- distribution finale expérimentale des cellules
$\tilde{\nu}$	- distribution finale simulée, avec en condition initiale les cellules de la distribution μ
n	- nombre de cellules étudiées dans μ
$\mathbb{N}$	- ensemble des entiers naturels
$\mathcal{P}(\dots)$	- ensemble des mesures de probabilité sur un espace
$P_i(t)$	- quantité de protéines du gène $i \in \mathbb{N}^G$ dans le modèle Bursty
Q	- un processus dans le problème de Schrödinger
R	- un processus de référence dans le problème de Schrödinger
r	- coefficient de corrélation du test de Pearson
ρ_{est}	- distribution estimée par l'interpolation à un temps intermédiaire
$s_{1,i}$	- taux de transcription de l'ARNm en protéines du gène $i \in \mathbb{N}^G$ dans le modèle Bursty
t	- variable temporelle
(t_1, t_2)	- temps de la première et deuxième expériences
(t_1, t_2, t_3)	- temps de trois expériences pour l'interpolation
τ	- variable représentant un retard temporel
θ	- matrice d'interaction du GRN
u, v	- vecteurs uniformes représentant les contraintes dans le problème de Schrödinger
x_i	- coordonnée de la cellule $i \in \llbracket 1, n \rrbracket$ de la distribution μ dans l'espace transcriptomique
y_i	- coordonnée de la cellule $i \in \llbracket 1, m \rrbracket$ de la distribution ν dans l'espace transcriptomique

Table des matières

1	Introduction	3
1.1	Différenciation cellulaire	4
1.2	Gènes, cycle cellulaire, et stochasticité	8
1.3	Les mesures biologiques	11
1.4	Modélisation de la différenciation	13
1.5	Inférence pour un processus de différenciation	17
1.6	Modélisation d'un gène	24
1.7	Modélisation d'un réseau de régulation de gènes	26
1.8	Méthodologie de la thèse	28
2	Cell Trajectory Inference	37
2.1	Introduction	38
2.2	Results	41
2.3	Discussion	54
2.4	Methods	56
2.5	Data Availability	67
2.6	Code Availability	67
2.7	Acknowledgments	67
2.8	Author Contributions	68
2.9	Competing Interests	68
2.10	Supplementary information	69
3	Discussion	77
3.1	Résumé des contributions	77
3.2	Prétraitement des données	78
3.3	Modélisation et paramétrisation du processus diffusif	79
3.4	Perspectives	80
	Bibliographie	89

1

Introduction

1.1	Différenciation cellulaire	4
1.1.1	L'embryogenèse	6
1.1.2	L'hématopoïèse	6
1.2	Gènes, cycle cellulaire, et stochasticité	8
1.3	Les mesures biologiques	11
1.4	Modélisation de la différenciation	13
1.4.1	Modélisation déterministe de la différenciation	14
1.4.2	Modélisation stochastique de la différenciation	14
1.4.3	Modélisation continue de la différenciation	16
1.5	Inférence pour un processus de différenciation	17
1.5.1	État de l'art des méthodes d'inférence	18
1.5.2	Les pionniers de l'inférence avec le transport optimal (Waddington-OT) et ceux qui s'en sont inspirés	19
1.5.3	Problème de Schrödinger	21
1.6	Modélisation d'un gène	24
1.7	Modélisation d'un réseau de régulation de gènes	26
1.8	Méthodologie de la thèse	28
1.8.1	Processus de référence basé sur la diffusion	28
1.8.2	Le modèle Bursty pour construire le processus de référence	30
1.8.3	Application sur données simulées	32
1.8.4	Résolution du problème de Schrödinger	33
1.8.5	Application sur des données expérimentales (Semrau et al., 2017)	35

Dans ce manuscrit, je m'intéresse au développement d'une méthode pour inférer une trajectoire de différenciation cellulaire à partir de données en cellules uniques. L'objectif est d'intégrer un modèle de la dynamique des gènes dans la méthode d'inférence. Pour cela, je construis un processus à partir d'un modèle mécaniste et stochastique de l'expression d'un réseau de régulation de gènes, appelé le modèle Bursty. J'utilise une théorie mathématique, le problème de Schrödinger, en incluant le processus ci-dessus comme un processus de référence qui fournit l'information de la dynamique des gènes. À partir de données ARN en cellules uniques ou de données protéomiques, la résolution du problème de Schrödinger avec le modèle Bursty permet une inférence plus réaliste et, *in fine*, plus pertinente de la différenciation cellulaire.

Avant de rentrer dans le cœur du travail, je définirai dans un premier temps, la différenciation cellulaire pour comprendre l'intérêt biologique d'inférer ces trajectoires. Ensuite, je présenterai des mesures biologiques couramment utilisées pour l'étudier, ce qui me permettra de faire le lien entre l'évolution des technologies de mesures biologiques et l'évolution de la modélisation. Je détaillerai dans une troisième partie les outils mathématiques utilisés dans l'état de l'art pour modéliser la différenciation cellulaire. L'inférence est au centre de mon travail de thèse, je lui dédis une section où je la définis, je présente l'état de l'art et j'introduis la méthode que j'utilise : la résolution du problème de Schrödinger. Mon objectif étant de résoudre le problème de Schrödinger en incluant la dynamique des gènes, je présenterai alors l'état de l'art de la modélisation d'un gène et d'un réseau de régulation de gènes.

1.1 Différenciation cellulaire

La différenciation cellulaire est un processus biologique qui mène une cellule à acquérir une ou des fonctions qu'elle ne possédait pas. À force de divisions successives, les cellules souches et les cellules non-différenciées vont progressivement s'engager vers un destin cellulaire dédié à une fonction précise, donnant naissance à une diversité de types cellulaires, caractérisés par des cellules dites différenciées. Les cellules différenciées (comme par exemple les globules rouges, cellules T, cellules musculaires, cellules épithéliales) remplissent ainsi des rôles distincts et spécifiques (Evans & Kaufman, 1981; Martin, 1981; Morrison et al., 1997; Thomson et al., 1998). La différenciation cellulaire est induite par des facteurs intrinsèques et extrinsèques. Les facteurs extrinsèques sont la communication entre les cellules, les signaux chimiques et mécaniques de l'environnement (Engler et al., 2006; Morrison & Spradling, 2008). Les facteurs intrinsèques portent principalement sur l'activité des gènes, leurs interactions (Graf & Enver, 2009) et l'épigénétique (Reik, 2007) (section 1.2). Dans cette thèse, je me concentre sur les facteurs intrinsèques de la différenciation ; je mentionnerai, toutefois, quelques travaux de l'état de

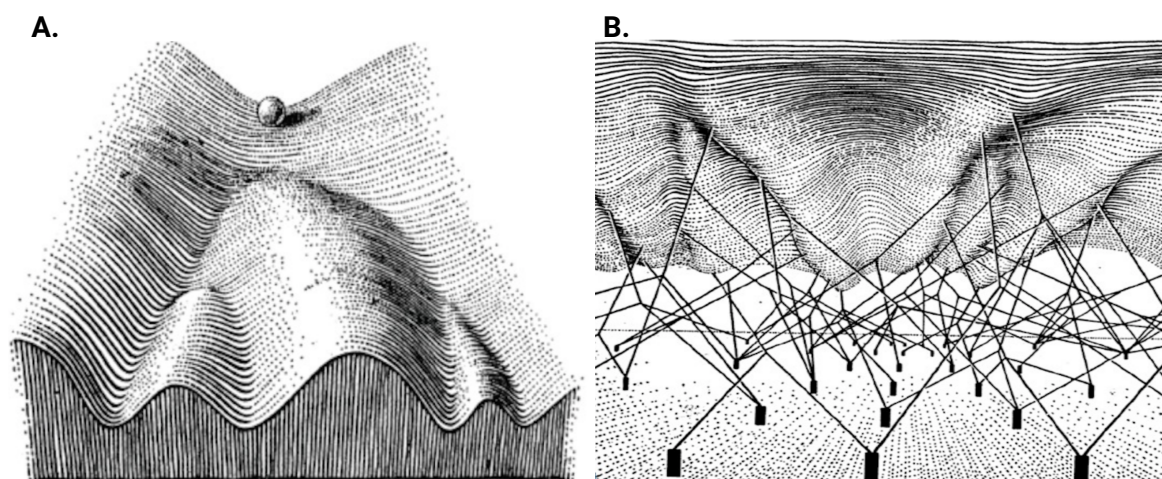


Figure 1.1. Paysage de Waddington. **A.** Paysage de Waddington où le processus de différenciation cellulaire est expliqué comme le mouvement d'une bille le long d'un paysage dans lequel les creux des vallées correspondent à des types cellulaires différenciés. **B.** Les gènes connectés dessinent un paysage que les cellules parcourent. (D'après les figures de Waddington (1957))

l'art qui prennent en compte les facteurs extrinsèques ainsi que dans la discussion (Chapitre 3).

Waddington (1957) a proposé en 1957 un modèle de la différenciation cellulaire : celle d'un paysage où les cellules roulent du haut d'une vallée vers le bas. Le haut des vallées représente le stade le moins différencié des cellules, et le bas des vallées le stade différencié (Figure 1.1A.). Le chemin parcouru par la cellule dans ce paysage s'identifie par une trajectoire cellulaire de différenciation. L'analogie va plus loin et imagine que ce sont les gènes qui tirent et poussent le paysage pour créer les montagnes et les vallées (Figure 1.1B.). La différenciation cellulaire est un terme générique, il peut s'appliquer à plusieurs types cellulaires comme la différenciation embryonnaire, la différenciation des cellules du sang et du système immunitaire, la différenciation des cellules épithéliales (Krasnyakov, 2024) et la différenciation des cellules endothéliales (Gerike et al., 1998; Germano & Osborne, 2021). Dans ce chapitre, l'accent est mis sur l'évolution des connaissances relatives à deux processus de différenciation spécifiques : la différenciation embryonnaire et l'hématopoïèse. Sans chercher à être exhaustive, je m'y intéresse car dans ces deux cadres la vision du processus a beaucoup évolué avec le développement de mesures biologiques modernes (section 1.3). Je présente dans les deux sous-sections qui suivent la différenciation embryonnaire et l'hématopoïèse à travers l'évolution des découvertes biologiques et je m'intéresse à l'évolution de la modélisation mathématique de ces deux phénomènes dans la section 1.4.

1.1.1 L'embryogenèse

La différenciation embryonnaire ou embryogenèse, qui n'intervient qu'une seule fois dans la vie d'un individu, permet, à partir d'une cellule-œuf unique, d'obtenir l'ensemble des précurseurs des composants de l'organisme (comme expliqué dans les livres de biologie du développement, par exemple : Barresi et Gilbert (2019)).

À la fin du XIXe siècle, à la suite d'expériences sur un embryon d'oursin, l'idée d'une spécialisation précoce des cellules et la notion de totipotence apparaissent (Driesch, 1891), c'est à dire que chaque cellule conserve le potentiel de se différencier menant au développement complet de l'individu. Soixante-dix ans plus tard, Gurdon (1962) transfère le noyau d'une cellule intestinale de têtard dans un ovocyte sans noyau. Il obtient un têtard viable ; cela prouve que le noyau d'une cellule différenciée contient toute l'information génétique nécessaire à la génération d'un organisme entier. Cette expérience fonde la notion de reprogrammation, ce qui lui vaudra le prix Nobel de médecine en 2012. Pour les souris (Evans & Kaufman, 1981; Martin, 1981) puis pour les humains (Thomson et al., 1998), respectivement, dans les années 1980 et 1990, les biologistes isolent les premières cellules souches embryonnaires, capables de se différencier en tous les types cellulaires.

Yamanaka, l'autre lauréat du prix Nobel de médecine en 2012, et ses collaborateurs ont montré que la différenciation n'est pas irréversible. En introduisant seulement quelques facteurs de transcription chez des fibroblastes de souris, ils obtiennent des cellules pluripotentes¹ induites, identiques aux cellules embryonnaires de souris (Takahashi & Yamanaka, 2006; Takahashi et al., 2007). Enfin, le rôle de l'expression des gènes et celui de l'épigénétique², sont identifiés comme régulateurs de la pluripotence et de la différenciation vers un destin cellulaire à partir de 2014 (Boland et al., 2014; Gökbuget & Blelloch, 2019). Avant l'arrivée de la technologie de séquençage en cellules uniques, que je présente dans la section 1.3, la différenciation était vue comme un processus discret entre des stades de différenciation clairement identifiés. L'arrivée de cette nouvelle technologie apporte de nouvelles informations sur l'évolution de l'expression des gènes durant la différenciation (Girard et al., 2026) et permet de mettre en valeur une trajectoire continue entre ces stades (Pijuan-Sala et al., 2019).

1.1.2 L'hématopoïèse

L'hématopoïèse est un processus continu tout au long de la vie, qui permet la formation des cellules sanguines et participe à la réponse immunitaire. Ce phénomène est particulièrement étudié, notamment en raison de son implication dans des pathologies du système immunitaire

¹ Les cellules pluripotentes sont des cellules avec la capacité de se différencier en presque tous les types cellulaires.

² L'épigénétique regroupe l'ensemble des mécanismes qui ne sont pas directement liés à l'expression des gènes (section 1.2)

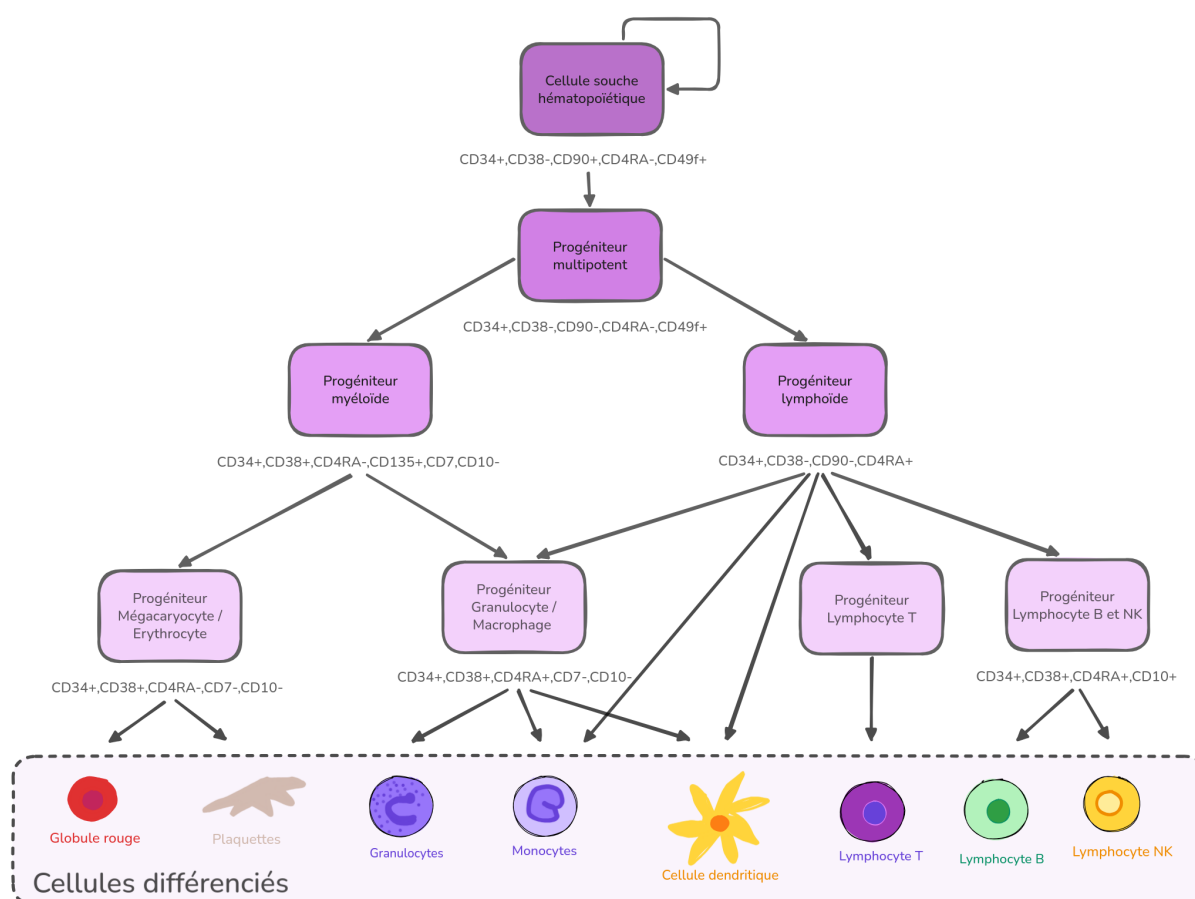


Figure 1.2. Une représentation en arbre de l'hématopoïèse humaine avec les marqueurs de surface identifiés dans Tajer et al. (2019). Les cellules souches hématopoïétiques sont définies par l'expression de CD34 (+), CD38 (-), CD45RA (-), CD90 (+) et CD49f (+). Les cellules souches sont représentées en haut de l'arbre et les cellules différenciées en bas de l'arbre. Les étapes de différenciation discrètes sont représentées par les stades intermédiaires dans l'arbre.

ou du système sanguin avec par exemple un dérèglement du nombre de cellules sanguines, tel que les leucémies.

Au XIX^e siècle, la notion de cellules souches pour la lignée des cellules du sang apparaît (Ramalho-Santos & Willenbring, 2007). Ehrlich (1879) décrit pour la première fois en 1879 les différents types de cellules du sang et introduit des techniques de coloration pour les distinguer. Trente ans plus tard, Maximow (1909) postule l'existence d'une cellule souche hématopoïétique unique, capable de donner naissance à toutes les cellules du sang. Les preuves expérimentales n'arrivent qu'une cinquantaine d'années plus tard, avec l'expérience menée par Till et McCulloch (1961), qui irradiant des souris et leur injectent des cellules de moëlle osseuse. Ils observent la formation de colonies comportant plusieurs types de cellules sanguines, ce qui prouve l'existence de cellules souches multipotentes.

Dans les années 1990, les cellules souches hématopoïétiques sont identifiées grâce à des marqueurs de surface (Civin et al., 1984; Spangrude et al., 1988). Ces marqueurs de surface permettront avec la cytométrie en flux (Bonner et al., 1972; De Rosa et al., 2001; Hulett et al., 1969) (décrite en section 1.3), la classification des cellules et l'identification des étapes de

différenciation (Lacombe et al., 2017; Perfetto et al., 2004) (Figure 1.2). Passegué et son équipe (2005) étudient l'activité des protéines des cellules souches hématopoïétiques et montrent que leurs activités diffèrent selon le devenir cellulaire de leur lignée, ce qui suggère que le destin se joue au cours de la vie de la cellule. L'identification de signaux dans la régulation, la prolifération et la différenciation est établie dans les années 2000 et 2010 (Bruns et al., 2014; Orkin & Zon, 2008). En parallèle, au milieu des années 2010, le rôle de l'épigenétisme est identifié dans le maintien de la pluripotence et son rôle dans l'engagement vers un destin cellulaire (Papp & Plath, 2013), pour l'hématopoïèse.

Pour conclure, ces deux processus suscitent un intérêt biologique et médical important (Orkin & Zon, 2008; Rossant & Tam, 2017). L'embryogénèse et l'hématopoïèse présentent une particularité supplémentaire que je ne prends pas en compte dans ce manuscrit, mais j'en discute dans le chapitre 3 : la dimension spatiale. Le point commun qui m'intéresse, entre ces deux processus est le changement de vision survenu avec l'arrivée de la technologie de séquençage en cellules uniques (scRNA-seq, voir section 1.3) : on passe d'une vision discrète de la différenciation à un continuum de différenciation (Girard et al., 2026; Laurenti & Göttgens, 2018). Le séquençage ARN en cellules uniques donne accès à des informations sur l'activité génétique de chaque cellule pour un grand nombre de gènes. Dans la section suivante, je présente les gènes et leur expression, le cycle cellulaire et la stochasticité qui résulte de l'expression des gènes.

1.2 Gènes, cycle cellulaire, et stochasticité

Comme décrit dans les livres de biologie moléculaire (Lodish et al., 2022)³, toutes les cellules eucaryotes possèdent un noyau et, dans ce noyau, l'intégralité de l'information génétique est contenue dans les chromosomes (Figure 1.3), mais les cellules ne l'expriment pas toutes de la même manière et au même moment.

Chaque cellule a son propre cycle de croissance ; une nouvelle cellule doit traverser différentes phases appelées G1, S, G2, M et G0. Les quatre premières phases correspondent à la phase de prolifération, tandis que la dernière correspond à la phase de repos. C'est ce qu'on appelle le cycle cellulaire. Au cours de sa vie, une cellule peut mourir par apoptose⁴ à tout moment de son cycle, ou se diviser en deux cellules filles à la fin de sa phase de prolifération. Lors de la mitose, les chromosomes se dupliquent puis se séparent en deux, donnant une copie de l'information génétique à chacune des deux cellules filles.

³Pour n'en citer qu'un

⁴L'apoptose est la mort cellulaire programmée

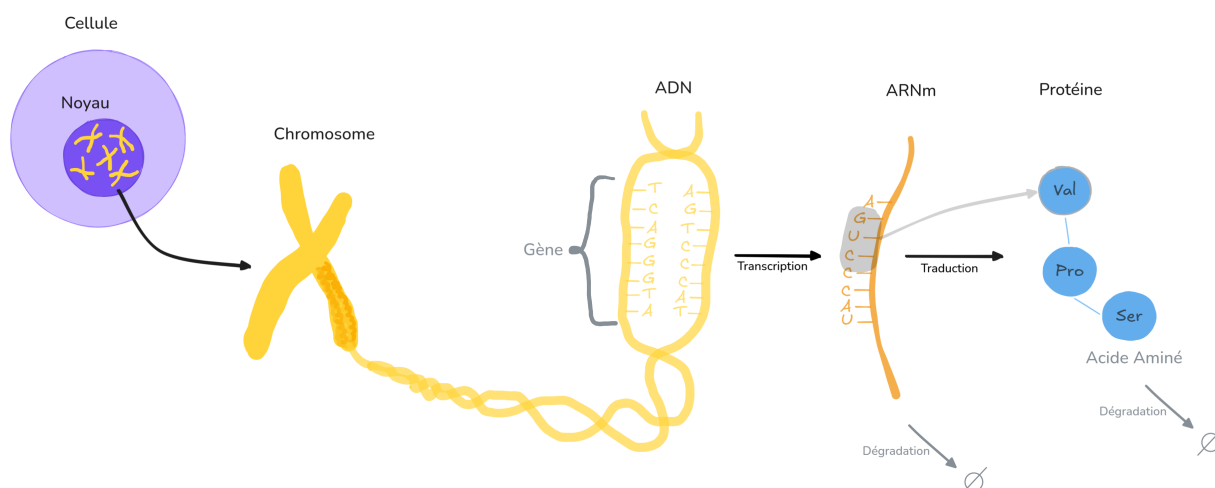


Figure 1.3. Dessin schématique d'un noyau dans une cellule, montrant la condensation de l'ADN en chromatine, composant des chromosomes. L'ADN est composé de quatre types de nucléotides : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). L'ADN est transcrit en ARNm. L'ARNm est aussi composé de quatre types de nucléotides : l'adénine (A), la cytosine (C), la guanine (G) et l'uracile (U). L'ARNm avec des triplés de nucléotides codant pour un acide aminé, est traduit en protéines. Dans l'exemple, j'illustre que la Valine (Val) est codée par le trio de nucléotides GUC, la Proline (Pro) est codée par CCA et la Serine (Ser) est codée par un trio qui commence par U (par exemple UCU). L'ARNm et les protéines sont ensuite dégradés.

La Figure 1.3 illustre un aspect des chromosomes, le fait qu'ils soient une condensation de la chromatine. La chromatine est une structure où l'acide désoxyribonucléique (ADN) en forme ellipsoïdale s'enroule avec d'autres molécules (par exemple les histones). L'ADN est composé de quatre types de nucléotides : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T), et est contenu dans le noyau. L'ADN est transcrit en acide ribonucléique messager (ARNm). Chez les cellules eucaryotes, qui possèdent un noyau contenant l'ADN, l'ARNm est ensuite épissé, c'est-à-dire qu'il ne conserve que les parties dites codantes pour les **protéines**. L'ARN épissé est alors transporté du noyau vers le cytoplasme où il est traduit en protéines. Un triplet de nucléotides code un acide aminé, élément constitutif des protéines. Ainsi, une suite de nucléotides précise peut coder une protéine précise. On peut définir cela comme un **gène**. Si un gène est transcrit, on dit qu'il est exprimé, ainsi pour estimer l'expression d'un gène, on mesure la quantité d'ARNm ou de protéines.

Les gènes sont interconnectés entre eux, dans ce qu'on appelle un "**réseau de régulation génique**" (**GRN** pour *Gene Regulatory Network*). Les gènes peuvent exercer une action inhibitrice ou activatrice sur eux-mêmes ou sur d'autres gènes. La protéine traduite peut être un facteur de transcription et se fixer sur le promoteur d'autres gènes pour en activer ou inhiber l'expression (Lambert et al., 2018; Latchman, 1996). Les facteurs de transcription peuvent jouer un rôle sur la structure de la chromatine. En effet, la condensation de la chromatine est plus ou moins ouverte, ce qui permet plus ou moins d'accéder à l'ADN et aux nucléotides (Bintu et al., 2016; Cremer et al., 2015). La conformation spatiale de l'ADN et l'état de la chromatine participent

également à la régulation génique (Fourel et al., 2004; Haddad et al., 2017; Rao et al., 2014) ; c'est ce qu'on appelle l'**épigénétique**. L'expression des gènes varie selon les cellules. Cette variation fut initialement décrite comme du "bruit" et sa première description a eu lieu au cours de la seconde moitié du XXe siècle (Novick & Weiner, 1957; Spudich & Koshland Jr, 1976). Longtemps considérée comme ne faisant que du bruit dans les cellules (Kaern et al., 2005), l'expression des gènes est en fait stochastique (Friedman et al., 2006; McAdams & Arkin, 1997; Paulsson et al., 2000) et joue un rôle clé dans la prise de décision cellulaire (Balázsi et al., 2011).

Eldar et Elowitz (2010) apportent une preuve expérimentale à l'expression stochastique des gènes. En suivant des cellules humaines transplantées avec un gène qui, au contact de l'ATP, produit de la lumière, on voit clairement que les gènes s'expriment de manière aléatoire. Cela décrit comment la synthèse des ARNm est générée de façon brève et stochastique. Cette stochasticité joue un rôle central (Chalancon et al., 2012; Eldar & Elowitz, 2010; Huang, 2009). Dans le livre "Le hasard au cœur de la cellule" (Kupiec et al., 2011), les auteurs invitent à repenser la biologie en intégrant la dimension probabiliste et stochastique dans la compréhension des mécanismes du vivant. En effet, la stochasticité à l'échelle moléculaire n'est pas incompatible avec l'organisation des structures à l'échelle macroscopique (Laforge et al., 2005). Une vision "darwinienne" est que la variation est au fondement du vivant et que les processus de hasard-sélection ne se restreignent pas qu'aux organismes eux-mêmes (Kupiec et al., 2011).

Viñuelas et al. (2012) présentent certaines explications à la stochasticité inhérente au processus d'expression génique : par des phénomènes biophysiques, comme la répartition aléatoire lors de la mitose (Huh & Paulsson, 2011), ou par des phénomènes biochimiques, comme la dynamique du réseau intracellulaire (Çağatay et al., 2009) et par la dynamique de la chromatine (Kupiec, 1997)⁵. Chalancon et al. (2012) examinent comment les GRN permettent d'atténuer, d'amplifier ou de tolérer le bruit d'expression des gènes. En effet, le GRN permet que le bruit dans l'expression d'un gène puisse se propager et affecter le bruit dans l'expression d'un gène connecté (Raser & O'Shea, 2005). La rétroaction positive ou négative a un impact sur l'amplification ou la diminution du bruit par exemple (Becskei & Serrano, 2000; Raser & O'Shea, 2005).

Pour conclure, l'expression génique est un processus stochastique, régi par des mécanismes tels que la structure de la chromatine, le réseau de régulation génique, ou les propriétés biophysiques du cycle cellulaire. Cette variabilité est clef dans la prise de décision cellulaire dans le processus de différenciation. Comprendre comment elle s'orchestre nécessite des outils de mesure adaptés. Dans la section suivante, je présente plusieurs mesures biologiques qui permettent d'analyser l'expression des gènes au cours du processus de différenciation cellulaire.

⁵ce dernier point est aussi étudié par Viñuelas et al. un an plus tard dans (Viñuelas et al., 2013)

1.3 Les mesures biologiques

Pour étudier les cellules lors du processus de différenciation, une analyse génétique est possible. Dans cette section, je présente plusieurs mesures biologiques qui permettent cette analyse. L'évolution des techniques repose principalement sur la résolution : augmentation du nombre de gènes étudiés et/ou diminution du nombre de cellules étudiées (de la population à la cellule individuelle). Je présente, dans un premier temps, des approches sur des populations de cellules, dites en *bulks*, et dans un second temps des approches à l'échelle de la cellule individuelle, dites en "cellules uniques" (Wang et al., 2009).

Le **Northern blot** (Alwine et al., 1977), introduit en 1977, est la première technique standardisée pour l'analyse quantitative de l'expression des ARNm en bulks. Son principe repose sur la séparation des ARNm grâce à leur charge magnétique (électrophorèse). Le signal obtenu permet à la fois de détecter et de quantifier un ARNm spécifique, et d'en déterminer la taille.

Plus tard, en 1996, une autre technique plus précise, la **RT-qPCR** (*reverse transcription quantitative polymerase chain reaction*) (Heid et al., 1996) est proposée pour analyser en bulks. Elle repose sur la transcription inverse (*reverse transcription*, RT) de l'ARNm en ADN complémentaire (ADNc), suivie d'une amplification par PCR (des réactions polymérase⁶ en chaîne) et d'une quantification à l'aide de la fluorescence. Elle permet de détecter des transcrits rares à partir de quantités infimes d'ARNm, mais elle reste limitée à un petit nombre de gènes analysés simultanément.

En parallèle, la méthode des **puces à ADN** (*microarrays*) (Schena et al., 1995) s'applique à plus grande échelle et permet une analyse simultanée de l'expression de milliers de gènes. Le principe consiste à transcrire de façon inverse des ARNm cibles en ADNc marqués par fluorescence. Les ADNc marqués sont ensuite mis en contact sur une lame de verre avec des milliers de sondes, c'est ce qu'on appelle l'hybridation. Les ADNc et les sondes créent un complexe qui est scanné à très haute résolution pour le signal des longueurs d'ondes des marqueurs fluorescents. L'intensité du signal à chaque spot reflète le niveau d'expression du gène correspondant.

Dans ce manuscrit, je présente deux des méthodes en cellule unique qui, comme leur nom l'indique, donnent des informations pour chaque cellule individuellement. Car l'effet de population peut brouiller les événements individuels, voire induire un paradoxe de Simpson (Simpson, 1951), c'est-à-dire une tendance inverse à la tendance réelle. Comme le montrent Richard et al. (2016), contrairement aux données d'expression génique populationnelles, les données en cellule unique ont révélé une forte variabilité intercellulaire, masquée par l'effet de masse.

La **cytométrie en flux** (Bonner et al., 1972; Hulett et al., 1969) est une méthode utilisée depuis la fin des années 60 qui permet de trier les cellules. Le principe consiste à introduire un échan-

⁶Les polymérases sont des enzymes actrices de la transcription, qui assurent la synthèse d'un brin d'ADN ou ARN.

tillon de cellules dans un tube étroit, ce qui permet l'isolement individuel des cellules. Un laser traverse le tube, atteignant chaque cellule. La réflexion de la lumière laser est étudiée en fonction de la longueur d'onde (Shapiro, 2003). Les marqueurs de surface ne réfléchissent pas la même longueur d'onde. La fonction biologique des marqueurs de surface n'est pas centrale, dans ce cadre, ils servent uniquement à identifier les différents types cellulaires (Goodwin et al., 2020) ; par exemple, le marqueur CD34 est couramment utilisé pour isoler les cellules souches hématopoïétiques (Ikuta & Weissman, 1992; Kiel et al., 2005; Oguro et al., 2013; Okada et al., 1992; Osawa et al., 1996) (Figure 1.2). Cette méthode permet ainsi de classer les cellules et d'identifier les étapes de différenciation, caractérisées par l'expression d'un marqueur de surface ou d'une combinaison de marqueurs (De Rosa et al., 2001; Lacombe et al., 2017; Perfetto et al., 2004). Les expériences détaillées ci-dessus renvoient l'image d'un processus de différenciation discret, comme représenté pour l'hématopoïèse dans la Figure 1.2, où les cellules passent d'un état discret de différenciation à un autre.

En 2009, le **séquençage ARN en cellules uniques** (*single Cell RNA sequencing*, **scRNA-seq**) (Tang et al., 2009) est développé. Cette expérience fournit des données à haute résolution permettant de caractériser les états cellulaires et d'observer les transitions entre eux (Trapnell, 2015) (Figure 1.4). Le principe est que, pour chaque cellule individuellement, on mesure l'expression des gènes ou le transcriptome à un temps donné, avec du séquençage. Pour cela, les cellules sont extraites d'un milieu et isolées dans des compartiments individuels à l'aide de la microfluidique à base de gouttelettes (Klein et al., 2015; Macosko et al., 2015). D'un point de vue moléculaire, les ARNm sont extraits et transcrits inversement (*reverse transcript*) en ADNc par une rétrotranscriptase. Ils sont ensuite amplifiés (Huang et al., 2018; Wang & Navin, 2015), puis séquencés en *reads* (Macosko et al., 2015). Les *reads* sont alignés sur un génome de référence (Dobin et al., 2013). La quantité d'ARNm d'un gène séquencé correspond à son niveau d'expression. On obtient alors une matrice de $\mathbb{N}^{n \times G}$ de transcrits, où n est le nombre de cellules et G le nombre de gènes. Cette matrice est un échantillon discret de la distribution à un instant donné, appelé un *snapshot* (Weinreb et al., 2018). De plus, les lignes de la matrice sont indépendantes (les cellules sont indépendantes entre elles par hypothèse).

Contrairement à la cytométrie en flux (Shapiro, 2003), l'extraction d'information nécessite la lyse des cellules, ce qui rend impossible de suivre ce processus au cours du temps, dans l'écrasante majorité des cas⁷. Le scRNA-seq nous donne, à un moment donné, pour des milliers de cellules individuellement, l'information sur la quantité de molécules d'ARNm pour des milliers de gènes (Klein et al., 2015; Macosko et al., 2015; Tanay & Regev, 2017; Wilbrey-Clark et al., 2020). L'expérience est largement utilisée depuis une quinzaine d'années et a démontré sa précision (Jovic et al., 2022; Svensson et al., 2018).

⁷Il existe une technologie spectrale qui ne lyse pas les cellules et que je discuterai dans le Chapitre 3 (Kobayashi-Kirschvink et al., 2024).

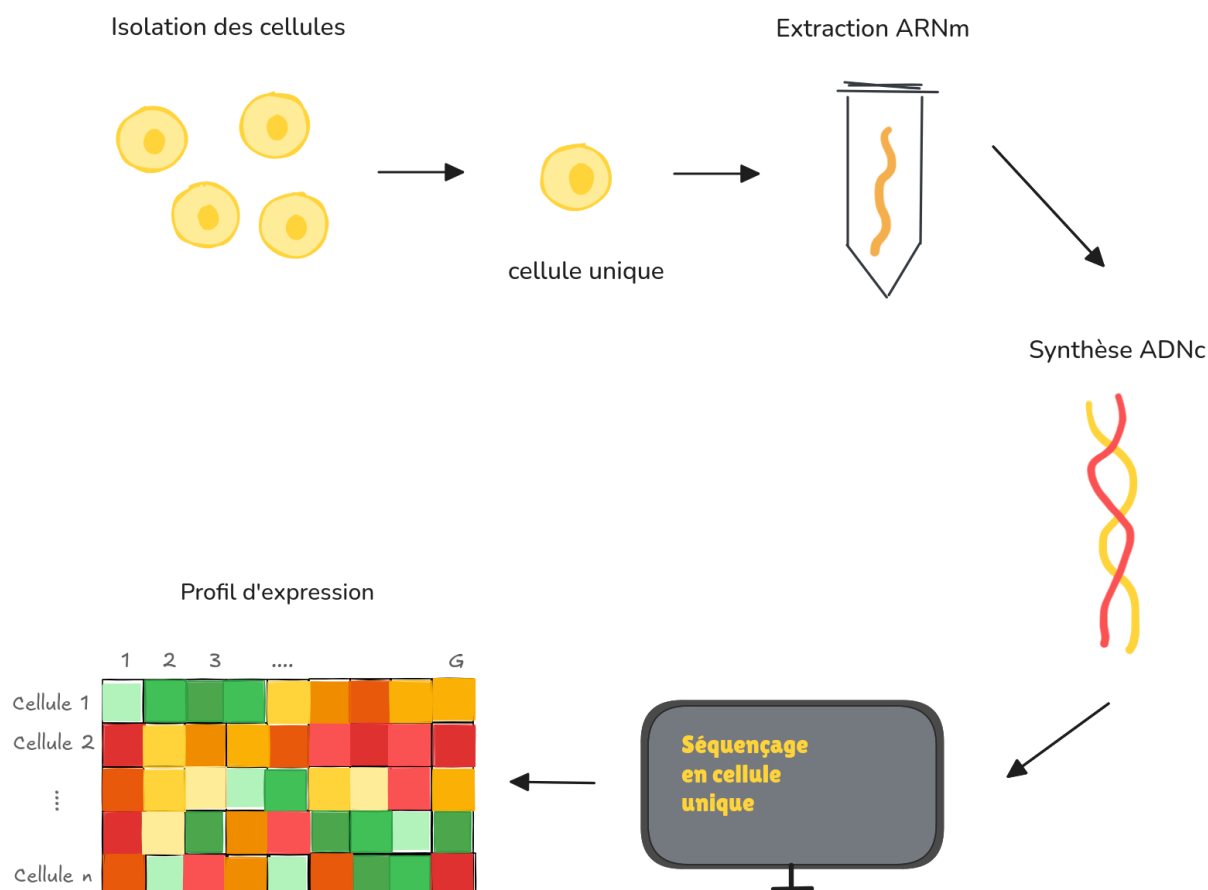


Figure 1.4. Schéma du protocole de l'expérience du scRNA-seq, d'après la figure de Bawa et al. (2022). Ce protocole comprend l'isolation cellulaire, l'extraction d'ARN, la synthèse d'ADNc par transcriptase inverse, le séquençage de cellules uniques et l'établissement du profil d'expression.

1.4 Modélisation de la différenciation

Comme l'explique Pujo-Menjouet (2016), en parallèle des découvertes biologiques sur les connaissances sur la différenciation, les biologistes ouvrent ou renforcent les collaborations, notamment avec les mathématicien.ne.s. Dans ce manuscrit, je me concentre sur la modélisation de deux processus de différenciation, l'hématopoïèse et l'embryogenèse "normales", dans le sens non-pathologique, mais il existe beaucoup de modèles dans le cas pathologique (Hermange et al., 2022; Iwasa, 2023).

L'histoire de la modélisation mathématique de la différenciation cellulaire commence dans les années 1960, avec par exemple la modélisation cinétique des globules rouges (Kirk et al., 1968) ou un modèle de régulation génique booléen pour étudier le phénomène de différenciation (Kauffman, 1969). Je regarde l'histoire des modèles sous plusieurs prismes : leur caractère déterministe, la prise en compte de la stochasticité et la prise en compte de la continuité du processus de différenciation. Je cherche à présenter l'éventail de la modélisation de la différenciation, sans chercher à être exhaustive.

1.4.1 Modélisation déterministe de la différenciation

Les premières formalisations mathématiques reposent sur des **équations différentielles ordinaires** (EDO). Avec l'exemple de l'hématopoïèse, la représentation en arbre de différenciation est courante, ce qui suggère un processus déterministe et discret où les cellules passent d'un stade de différenciation à l'autre par étapes bien définies (Figure 1.2). Ainsi, des modèles compartimentaux ont été proposés, où chaque compartiment homogène représente un stade de différenciation et le passage d'un compartiment à l'autre représente la différenciation. Les dynamiques cellulaires sont modélisées par des EDO, comme par exemple le modèle non linéaire de Marciniak-Czochra et al. (2008).

Des modèles proposés à la fin des années 70 (Mackey & Glass, 1977; Wążewska-Czyżewska & Lasota, 1976), et qui ont inspiré d'autres travaux plus récents (Crauste, 2009), reposent sur une **équation différentielle à retard**, c'est-à-dire une équation reliant les variations au cours du temps t d'une certaine variable, ici respectivement la quantité de globules rouges et de globules blancs en circulation, à ses valeurs prises à un instant antérieur $t - \tau$. Les équations différentielles à retard permettent de prendre en compte le temps nécessaire aux cellules souches pour se différencier en cellules matures (Bélair et al., 1995) ou le temps de la division cellulaire (Adimy et al., 2015).

Une troisième approche déterministe utilise des **équations aux dérivées partielles** (EDP), formalisées par exemple, pour l'hématopoïèse par Doumic et al. (2011). Ces modèles peuvent permettre d'incorporer une dimension spatiale (Eymard et al., 2014), pouvant être parfois couplée à des méthodes de contrôle optimal (Dupuis, 2014). La modélisation de la morphogenèse embryonnaire par des systèmes EDP dits de réaction-diffusion, hérités des travaux de Turing (1953), constitue également un pan important de la modélisation déterministe du développement (Maini & Baker, 2012).

1.4.2 Modélisation stochastique de la différenciation

Comme introduit dans la section 1.2, la stochasticité de l'expression des gènes a pu être considérée comme du bruit et non une information (Kaern et al., 2005). Dans cette sous-section, je décris des modèles qui prennent en compte la stochasticité du processus de différenciation. Je présente des exemples de processus de Markov, de modèle non markovien, de modèle de Markov caché, de modèle à bases d'agents, de modèles couplés et de modèle d'équations différentielles stochastiques (EDS).

La première approche stochastique pour décrire la prolifération des cellules souches est introduite par Till et al. (1964) en 1964, avec un **processus de Markov** de type naissance-mort ; on constate que cette approche apparaît au même moment, voire un peu plus tôt, que les modèles déterministes mentionnés dans la section précédente. Armond et al. (2014) modélisent

les transitions entre états cellulaires, les auteurs proposent un cadre mathématique s'appuyant sur les chaînes de Markov pour décrire les probabilités de passage d'un sous-type à l'autre. Les chaînes de Markov sont un ensemble d'états finis et discrets et une matrice de transition qui donne la probabilité de transition d'un état à un autre. La propriété qui rend un processus markovien est que la probabilité de transition ne dépend que de l'état actuel, et non de l'historique global.

Udomlumlert et al. (2021) ont montré quant à eux que les lignées de cellules souches embryonnaires présentent des transitions d'état **non markoviennes**, c'est-à-dire que la probabilité de transition ne dépend pas seulement de l'état présent mais de tout l'historique cellulaire antérieur. Cette propriété biologique est connue sous le nom de la mémoire cellulaire.

Stumpf et al. (2017) ont proposé un modèle non markovien et stochastique pour la différenciation des cellules souches embryonnaires le long de la lignée neuronale : en profilant l'expression génique de cellules individuelles, ils identifient une chaîne d'états moléculaires cachés que chaque cellule traverse séquentiellement, conférant ainsi à la cellule une forme de mémoire de sa progression le long de la trajectoire de différenciation. Ces états cachés sont formalisés par **un modèle de Markov caché**, dont les dynamiques observables sont non markoviennes. La même logique d'états cachés est employée différemment par Catlin et al. (2001) en 2021, qui développent un processus de Markov caché pour l'hématopoïèse où la dynamique interne des compartiments de cellules souches demeure inobservable : seule la production de cellules matures est accessible, et les paramètres du modèle sont inférés à partir de ces seules variables observables.

Dans les **modèles à base d'agents et les modèles multi-agents**, une cellule peut être modélisée comme un agent qui évolue au sein d'une population (Bernard et al., 2024; Bessonov et al., 2006; d'Inverno & Saunders, 2004; Nguyen et al., 2024; Roeder et al., 2008). Ces modèles sont un mélange d'équations déterministes et stochastiques, le modèle résultant étant généralement stochastique, visant à décrire soit le mouvement d'une cellule, soit ses interactions avec l'environnement cellulaire interne et externe, soit les interactions entre un groupe de cellules dans la moelle osseuse. Cette approche a été appliquée à l'embryogenèse, en simulant les interactions mécaniques et moléculaires (Glen et al., 2019; Nissen et al., 2017). Une approche multi-agents sur la différenciation, prolifération, mort des cellules T CD8 est proposée avec le logiciel Simuscale qui considère que les agents sont des cellules et au sein de chaque agent le GRN est implémenté avec un processus de Markov déterministe par morceaux (PDMP) (Bernard et al., 2024; Nguyen et al., 2024) ; je présente en détail ce PDMP dans la section 1.7.

Le **couplage de modèles déterministes et stochastiques** est largement utilisé pour l'étude de lignées cellulaires et porte principalement sur le déclenchement des transitions entre les états cellulaires et l'intégration de la dynamique moléculaire au sein de la population, pour l'hématopoïèse (Glauche et al., 2007; Roeder & Glauche, 2006; Roeder et al., 2005) et pour l'em-

bryogénèse (Task et al., 2012). Certains modèles considèrent que la transition entre les états de différenciation est induite par la stochasticité ou forcée entre des états attracteurs (Chang et al., 2008; Furusawa & Kaneko, 2012; Richard et al., 2016). Fisher et al. (2012) couplent des approches déterministes (EDO et EDP) avec des approches discrètes et stochastiques, notamment en intégrant l'aspect spatiale de l'érythropoïèse.

Enfin, je mentionne des modèles mathématiques stochastiques où la différenciation est orchestrée par un GRN (Bessonard et al., 2014; Dunn et al., 2014; MacArthur et al., 2008; 2009; Tosenberger et al., 2017). MacArthur et al. (2009) ont utilisé la stochasticité dans un modèle de GRN décrit par un **modèle d'équations différentielles stochastiques** pour modéliser les transitions entre les états de différenciation. Tosenberger et al. (2017) ont identifié des gènes déterminants dans la différenciation embryonnaire chez la souris. Ils ont modélisé des cellules en différenciation par l'activité des gènes avec un système d'EDO auquel ils ajoutent un saut stochastique qui implémente l'amplification aléatoire d'un signal, donnant lieu, entre autre, à la division.

Certains modèles stochastiques modélisent la stochasticité de l'expression des gènes mais ne prennent toutefois pas en compte la différenciation cellulaire comme un processus continu. La sous-section suivante est dédiée aux modélisations continues de la différenciation cellulaire.

1.4.3 Modélisation continue de la différenciation

Avec l'arrivée de la technologie du scRNA-seq, on observe un passage à une vision continue de la différenciation (Girard et al., 2026; Laurenti & Göttgens, 2018). Dans cette section, je présente des modèles qui peuvent être stochastiques ou déterministes et je me concentre sur leur capacité à modéliser un phénomène continu. La métaphore du paysage de Waddington (Figure 1.1) propose que le processus de différenciation soit considéré comme un phénomène dynamique, continu et aléatoire (Waddington, 1957). Dans cette représentation, la différenciation n'est plus un saut discret entre états pré-définis mais une trajectoire continue dans un espace de (grande) dimension (Laurenti & Göttgens, 2018).

Laurenti et Gottgens (2018) discutent de la remise en cause du modèle hiérarchique strict au profit d'un processus plus continu, plus en adéquation avec le paysage de Waddington. Ils identifient que la modélisation en arbre reposait largement sur l'analyse de populations cellulaires prédéfinies par des marqueurs de surface, ce qui masquait l'hétérogénéité réelle des cellules individuelles.

Pour modéliser ce continuum pour l'hématopoïèse, des approches ont été proposées telles que : des modèles de marche aléatoire dans l'espace des états cellulaires (Haghverdi et al., 2016) ou des analyses statistiques de type pseudo-temps. Le pseudo-temps est une unité de mesure abstraite qui permet d'ordonner les cellules selon leurs degrés de différenciation (Trap-

nell et al., 2014).

Pour modéliser ce continuum pour l'embryogenèse, conformément au paysage de Waddington, des études cherchent à représenter graphiquement la différenciation des données scRNA-seq. Parmi celles-ci, Fard et al. (2016) appliquent les réseaux de Hopfield ⁸ à des cellules souches embryonnaires humaines en différenciation, interprétant les attracteurs du réseau comme des types cellulaires stables.

Shi et al. (2020) s'appuient sur l'entropie des chaînes de Markov avec entre autre les données de Semrau et al. (2017) que nous avons aussi étudiées (voir section 1.8.5 et Chapitre 2). L'entropie est une notion centrale dans mon travail de thèse que je définis dans la section 1.5.3 ; elle est essentielle dans la méthode d'inférence que je propose et je m'en sers aussi comme d'un outil pour étudier la différence entre deux processus. Shi et al. (2020) reconstruisent le paysage de Waddington en intégrant le profil scRNA-seq avec un réseau d'interaction de protéines, dont les transitions sont estimées avec l'entropie des chaînes de Markov.

Dans le contexte de l'embryogenèse, Saez et al. (2022) mobilisent la théorie des catastrophes pour modéliser les bifurcations de destin cellulaire dans un cadre continu : les singularités du paysage de Waddington correspondent à des bifurcations mathématiques au sens de la théorie de Thom (Arnol'd et al., 1989; Thom, 1989). D'un point de vue dynamique, les vallées du paysage sont assimilées aux bassins d'attraction d'un système d'équations différentielles (Huang, 2010), permettant d'ancrer rigoureusement la métaphore de Waddington dans un formalisme mathématique. Des approches de paysage permettent, en outre, de relier la topologie du réseau de régulation génique aux transitions de lignage observées expérimentalement.

On pourra noter qu'aucune de ces études n'examine directement le mécanisme cellulaire qui génère de la stochasticité à l'échelle du gène, contrairement à la méthode que j'introduis dans ce travail (Chapitre 2).

Dans la sous-section précédente, j'ai présenté des modèles d'EDO et des modèles d'EDS mais qui ne prenaient pas en compte la continuité du processus de différenciation. Pujadas et Feinberg (2012) proposent, dès 2012, un modèle d'EDS pour modéliser l'embryogenèse de façon continue, expliquant que la stochasticité de l'expression est régulée et non subie, comme le suggérait Waddington et illustré avec la Figure 1.1B.

Dans l'idée d'étudier la différenciation comme un phénomène stochastique et continu, une branche se développe : l'inférence, présentée dans la section 1.5.

1.5 Inférence pour un processus de différenciation

Je rappelle que la technologie de séquençage ARN en cellules uniques est destructive, il est donc impossible de suivre l'évolution temporelle de l'expression des gènes d'une même cellule

⁸Un réseau avec un formalisme emprunté à la physique statistique des systèmes magnétiques (Hopfield, 1982).

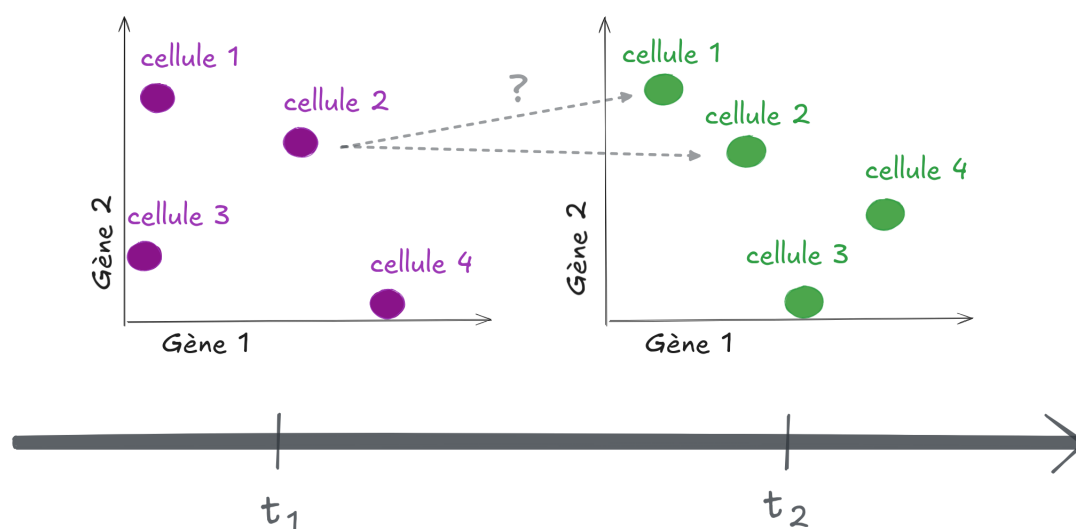


Figure 1.5. Représentation de quatre cellules dans l'espace d'expression de deux gènes lors de deux expériences à deux temps, t_1 et t_2 . Les indices des cellules sont propres à chaque temps. On se demande comment les cellules vont évoluer entre les deux temps.

lors d'une expérience utilisant cette technologie. Or je cherche précisément à la reconstruire. La méthode consistant à estimer des propriétés d'une population ou d'un tissu à partir de l'observation d'un échantillon (de petite taille) s'appelle l'inférence. Dans ce travail, j'utilise une méthode d'inférence pour estimer une évolution temporelle à partir de mesures indépendantes.

L'inférence à laquelle je m'intéresse se base sur l'étude des *snapshots* de scRNA-seq pour reconstituer l'évolution continue du phénomène de différenciation (Saelens et al., 2019; Trapnell, 2015). C'est comme si on voulait reconstituer le film à partir d'une série de photos ordonnées, mais avec des acteurs différents, et peut-être même des décors différents (Figure 1.5).

1.5.1 État de l'art des méthodes d'inférence

L'inférence ne prend pas toujours l'aspect stochastique de la différenciation en compte. En 2014, Bendall et al. (2014) proposent un algorithme d'inférence de trajectoires, basé sur la construction de graphes des k plus proches voisins, qui, à partir de l'expression d'une vingtaine de marqueurs en cytométrie, reconstruit la trajectoire de développement d'une lignée hématopoïétique. Plus récemment, Saelens et al. (2019) ont répertorié des méthodes d'inférence à partir de données scRNA-seq. Il existe des méthodes basées sur des graphes, comme Monocle (Trapnell et al., 2014), TSCAN (Ji & Ji, 2016) ou Slingshot (Street et al., 2018), d'autres sur des modèles probabilistes comme GPfates (Lönnberg et al., 2017) et CellRouter (Lummertz da Rocha et al., 2018), ou se basant sur des méthodes de *machine learning* (Hashimoto et al., 2016).

Monocle (Trapnell et al., 2014) combine une réduction de dimension non-linéaire et un ordonnancement des cellules le long d'une trajectoire. Cette méthode introduit la notion de pseudo-temps : une organisation abstraite des cellules qui reconstitue le processus de différenciation.

Ce concept est revisité avec de la marche aléatoire pour développer la “diffusion pseudo-time” (Haghverdi et al., 2016), ou amélioré avec Slingshot (Street et al., 2018) et TSCAN (Ji & Ji, 2016) qui utilisent le clustering cellulaire et connectent les clusters. Les méthodes de pseudo-temps sont aujourd’hui remplacées par des méthodes de “RNA velocity” qui mesurent l’abondance des ARNm épissés *versus* non épissés, ce qui peut permettre de prédire l’état transcriptionnel des cellules (La Manno et al., 2018), avec notamment ScVelo (Bergen et al., 2020).

Parmi les méthodes probabilistes, GPfates (Lönnerberg et al., 2017) modélise des trajectoires cellulaires dans un espace réduit par une analyse en composantes principales (ACP), en considérant un processus gaussien latent et en supposant que les cellules évoluent de manière continue. CellRouter (Lummertz da Rocha et al., 2018) modélise un réseau de flux ; la méthode cherche le chemin le plus probable entre un état de départ et un état final.

1.5.2 Les pionniers de l’inférence avec le transport optimal (Waddington-OT) et ceux qui s’en sont inspirés

Le Transport Optimal (OT) est utilisé pour la première fois dans l’inférence de trajectoires par Schiebinger et al. (2019).

Le transport optimal a été à l’origine développé par Monge (1781) pour redistribuer la terre entre des déblais et des remblais, qu’il imagine comme des distributions de terre à transporter. Plus tard, en 1942, Kantorovich généralise le problème en introduisant la minimisation du coût du transport. On parle alors du problème de Monge-Kantorovich ou du transport optimal quadratique, car le coût utilisé est généralement la distance euclidienne au carré (Kantorovich, 1942; Peyré & Cuturi, 2019a).

Le problème de Monge-Kantorovich admet également une formulation dynamique, proposée par McCann (1997), qui prend en compte la trajectoire complète du système et non seulement ses positions initiales et finales. On peut modéliser l’évolution d’une mesure au cours du temps comme un chemin minimal sur un champ vectoriel dépendant du temps, ou selon le gradient d’un potentiel convexe, appelé plan de Brenier ou *Brenier map* (Benamou & Brenier, 2000).

Ensuite, une variante a été proposée : le transport optimal entropique, dans lequel le coût à minimiser inclut un terme de régulation entropique (Cuturi, 2013).

Le transport optimal entropique peut alors être résolu grâce à l’algorithme de Sinkhorn (Sinkhorn, 1967). L’algorithme de Sinkhorn a été introduit à l’origine pour résoudre le problème de Schrödinger ; je présente le problème de Schrödinger et l’algorithme de Sinkhorn dans la section 1.5.3.

Schiebinger et al. (2019) proposent l’algorithme Waddington-OT (Waddington Optimal Transport). Leurs travaux amènent une vision probabiliste, qui permet de capturer des dynamiques, telles que des bifurcations dans les trajectoires. Ils considèrent le problème à l’échelle de la population mais en tenant compte de l’individualité des cellules. Ils étudient des fibroblastes

d'embryons de souris reprogrammés en cellules souches pluripotentes, analysés par scRNA-seq pendant 18 jours. Les cellules sont représentées comme des points dans un espace multidimensionnel, appelé "espace d'expression des gènes", où le nombre de dimensions correspond au nombre de gènes étudiés (un exemple pour deux gènes est illustré en Figure 1.5) (Huang, 2010). Les cellules sont positionnées dans cet espace d'expression des gènes en fonction de l'expression de chaque gène. Ainsi, l'idée est, pour chaque cellule et à chaque temps, de dessiner une distribution de probabilité dans l'espace des gènes, puis de reconstruire, avec la théorie du transport optimal, la distribution la plus probable d'origine et de destin cellulaires.

Schiebinger et al. (2019) développent Waddington-OT pour étudier un processus de différenciation, à partir de 315 000 cellules. Ils utilisent le transport optimal quadratique régularisé introduit par Chizat et al. (2018), dans un espace d'expression des gènes réduit par une analyse en composantes principales (ACP). Dans cet espace réduit, ils supposent que les cellules ne se déplacent pas beaucoup entre deux temps proches. Waddington-OT suppose que les cellules se déplacent de façon diffusive dans l'espace d'expression des gènes réduit. Ces travaux considèrent 1 492 gènes initialement, après la réduction en ACP, ils se placent généralement en 30 dimensions.

D'autres équipes vont s'inspirer de ces travaux et appliquer le transport optimal aux données scRNA-seq (Bunne et al., 2024; Jiang et al., 2022; Lange et al., 2024; Lavenant et al., 2024; Liu et al., 2020; Lotfollahi et al., 2023; Prasad et al., 2022; Sha et al., 2023; Tong et al., 2020; Tronstad et al., 2025; Yachimura et al., 2024; Zhu et al., 2025). La plupart vont aussi utiliser la distance euclidienne au carré pour le transport optimal, équivalent du processus diffusif dans le problème de Schrödinger. Intuitivement, on peut imaginer qu'un processus diffusif va avoir du mal à capter des déplacements dans l'espace d'expression des gènes quand ceux-ci obéissent à des modèles Bursty. Je vais vérifier cette intuition dans cette thèse.

Avec gWOT, Lavenant et al. (2024) ont adopté un processus de drift-diffusion pour inférer les trajectoires. gWOT reconstruit les trajectoires cellulaires en modélisant le transport optimal régularisé par une EDS, ce qui peut être interprété comme le processus de référence du problème de Schrödinger. Même si une critique du simple processus diffusif est formulée, le drift diffusif ne tient toujours pas compte de la dynamique des gènes.

Certains travaux s'éloignent de la diffusion et de la distance euclidienne, mais surtout pour ajouter d'autres informations aux données scRNA-seq, telles que la lignée cellulaire (Lange et al., 2024; Majima et al., 2024), la co-expression de certains gènes (Du et al., 2019), la spatialité (Klein et al., 2025) ou, plus largement, considèrent des données multi-omiques (Samaran et al., 2026). Par exemple, l'algorithme moslin (Lange et al., 2024) se penche sur le transport dans l'espace des phénotypes tout en respectant les relations de lignée entre les cellules. Pour cela, il minimise le coût entre deux matrices de coût : l'une, basée sur les distances euclidiennes au

carré des données scRNA-seq, et l'autre, basée sur les distances de Hamming⁹ pour traiter les informations des lignées cellulaires.

Néanmoins, certaines méthodes prennent en compte un autre coût dans le transport optimal. Par exemple, Liu et al. (2020) ont critiqué l'utilisation de la distance euclidienne et ont proposé une implémentation d'une fonction de coût apprise, basée sur l'algorithme de Sinkhorn (Sinkhorn, 1967). Cependant, ils n'utilisent pas la théorie du transport optimal pour inférer les trajectoires cellulaires ; ils se concentrent plutôt sur l'alignement de jeux de données scRNA-seq. De plus, le coût appris n'est pas interprétable, une propriété partagée par les méthodes de calcul fondées sur le *machine learning*.

Afin de réduire le coût de calcul et d'améliorer l'interprétabilité, scEGOT (Yachimura et al., 2024) combine le transport optimal avec un graphe d'états cellulaires. En ajoutant une distribution de mélange gaussien aux coordonnées de chaque cellule, scEGOT permet de considérer des groupes de cellules plutôt que des cellules individuelles. Le mélange gaussien rend les résultats interprétables, mais ses objectifs diffèrent clairement des nôtres, car les auteurs utilisent la distance euclidienne entre distributions gaussiennes pour inférer les trajectoires, s'appuyant une fois de plus sur une approche standard, c'est à dire des méthodes de transport optimal avec la distance euclidienne.

Il est important de mentionner d'autres méthodes qui recourent au *machine learning* ou au *deep learning*. Leur objectif est d'apporter des améliorations aux méthodes existantes, par exemple : sur le temps de calcul et la précision (Prasad et al., 2022), sur la complexité (Huguet et al., 2023), sur les paramètres de croissance et de mort des cellules (Sha et al., 2023), ou sur une réponse aux perturbations (Bunne et al., 2023; Lotfollahi et al., 2023), mais qui s'éloignent de nos enjeux.

Dans la prochaine sous section, je présente le problème de Schrödinger qui est très proche de la théorie du transport optimal. Mon objectif est de résoudre le problème de Schrödinger en prenant en compte une dynamique d'expression des gènes réaliste.

1.5.3 Problème de Schrödinger

Introduit en 1932, le problème de Schrödinger est un problème de minimisation de l'entropie avec des contraintes marginales et un processus de référence fixe (Schrödinger, 1932). L'objectif est de déterminer un couplage optimal entre deux distributions : une distribution initiale, notée μ , et une distribution finale, notée ν . Pour reprendre la métaphore de la section 1.5 : l'inférence de trajectoire peut être métaphorisée comme la reconstitution d'un film à partir d'une série de photos ordonnées. Dans cette métaphore, le processus de référence est le scénario de référence, et la résolution du problème est d'identifier un scénario qui est le plus proche du

⁹La distance de Hamming permet de quantifier la différence entre deux séquences de symboles (Hamming, 1950)

scénario de référence en prenant en compte les photos.

Le couplage nous dit comment associer les distributions de départ et d'arrivée, alors que le processus décrit comment les distributions ont évolué entre le départ et l'arrivée.

Je détaille les éléments de la définition du problème de Schrödinger. Soient $(x_1, \dots, x_N) \in \mu$ et $(y_1, \dots, y_N) \in \nu$ des niveaux d'expression des gènes, respectivement aux instants de mesure t_1 et t_2 , mesurés par la quantité d'ARNm ou de protéines dans la suite. On note Π et Ψ , deux espaces discrets et finis tels que $\Pi = (x_1, \dots, x_N)$ et $\Psi = (y_1, \dots, y_N)$. Le problème de Schrödinger s'appuie sur:

- Un **processus de référence**, noté R , tel que $R \in \mathcal{P}(\Pi \times \Psi)$ où $\mathcal{P}(\Pi \times \Psi)$ est l'ensemble des mesures de probabilité sur $\Pi \times \Psi$. Il joue le rôle de référence pour la minimisation, tout comme la fonction de coût dans le transport optimal. Il permet d'orienter la minimisation parmi toutes les trajectoires ayant des marginales initiales et finales, μ et ν .
- L'**entropie** relative, notée H , d'un processus Q avec un processus de référence R , définie par :

$$H(Q|R) = \sum_{ij} Q_{ij} \log \left(\frac{Q_{ij}}{R_{ij}} \right), \quad (1.1)$$

que j'appellerai juste "entropie" par la suite. Elle peut aussi se nommer divergence de Kullback–Leibler de Q à partir de R . L'entropie est un outil dont je me servirai pour sa propriété de divergence dans les Résultats du Chapitre 2.2 ;

- Des **contraintes marginales**, qui sont des vecteurs, notés $u \in \mathcal{P}(\Pi)$ et $v \in \mathcal{P}(\Psi)$, qui permettent d'attribuer un poids aux éléments des distributions μ et ν .

On écrit le problème de Schrödinger comme :

$$Sch(R; u, v) = \underset{Q \in \mathcal{P}(\Pi \times \Psi)}{\operatorname{argmin}} \left\{ H(Q|R); \forall i, u_i = \sum_j Q_{ij}, \forall j, v_j = \sum_i Q_{ij} \right\}. \quad (1.2)$$

L'équation (1.2) est un problème de minimisation entropique par rapport au processus de référence R en respectant les contraintes marginales u et v .

Le problème de Schrödinger peut être résolu grâce **l'algorithme de Sinkhorn**. Il s'agit d'un algorithme itératif qui fournit rapidement une solution approchée. Pour cela, l'algorithme est initialisé avec une matrice issue d'un processus de référence ainsi que deux vecteurs marginaux. De manière itérative, en partant de la matrice, l'algorithme cherche à se rapprocher de l'une des deux marginales, puis de l'autre, jusqu'à convergence vers une solution, comme illustré dans la figure 1.6. Je le décris en détail sous forme de pseudo-code dans l'encadré Algorithme 1.1. À partir d'une matrice de couplage de référence R , et deux marginales, les vecteurs u et v , l'algorithme de Sinkhorn trouve le couplage optimal avec une précision ε .

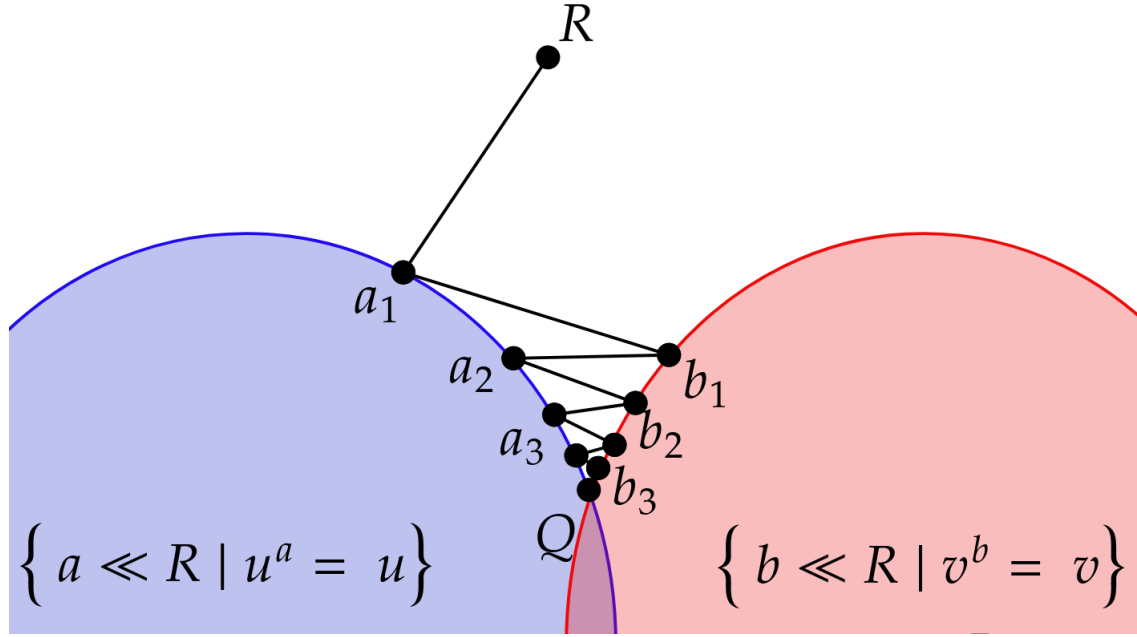


Figure 1.6. Représentation schématique de la résolution du problème de Schrödinger avec l'algorithme de Sinkhorn, d'après la figure de Baradat et Ventre (2024). L'algorithme de Sinkhorn (Algorithme 1.1) prend en entrée une matrice de couplage de référence R , et deux marginales, les vecteurs u et v . Il résout de façon itérative, en se rapprochant petit à petit des marginales avec les a et les b .

Algorithme 1.1 : Sinkhorn

Input : $u \in \mathbb{R}^n, v \in \mathbb{R}^m, R \in \mathbb{R}^{n \times m}, \varepsilon > 0$

Output : Matrice de couplage optimal Q

```

1  $u \leftarrow u.\text{reshape}(m, 1), v \leftarrow v.\text{reshape}(n, 1)$ 
2  $a \leftarrow \mathbf{1}_m, b \leftarrow \mathbf{1}_n$ 
3  $P \leftarrow \text{diag}(a) * R * \text{diag}(b)$ 
4  $p_{\text{norm}} \leftarrow \text{Tr}(P^\top P)$ 
5 repeat
6    $a \leftarrow \frac{u}{\max(R * b, 10^{-300})}$  // pour éviter la division par zéro
7    $b \leftarrow \frac{v}{\max(R^\top * a, 10^{-300})}$ 
8    $Q \leftarrow \text{diag}(a) R \text{diag}(b)$ 
9   if  $|\text{Tr}(Q^\top Q) - p_{\text{norm}}| / p_{\text{norm}} < \varepsilon$  then
10    break
11  end
12   $p_{\text{norm}} \leftarrow \text{Tr}(Q^\top Q)$ 
13 until convergence
14 return  $Q$ 
```

Le problème de Schrödinger utilise le plus souvent le mouvement brownien ou un processus diffusif comme processus de référence. Historiquement, le problème de Schrödinger a été introduit pour décrire le mouvement physique des particules. Un processus diffusif est souvent utilisé comme processus de référence car il décrit des phénomènes physiques et a, de plus, des propriétés pratiques pour la modélisation comme le fait que sa mesure de référence est absolument continue ou que son noyau de transition est gaussien. Dans mon cadre d'étude, l'utilisation d'un processus diffusif suggérerait que dans l'espace d'expression des gènes, les cellules se déplaceraient de manière diffusive.

Aucun travail n'utilise directement le problème de Schrödinger pour l'inférence de trajectoire, mais le problème de Schrödinger est lié au transport optimal : la solution du problème de Schrödinger converge vers la solution du transport optimal quadratique lorsque le coefficient de diffusion tend vers 0 et que le processus de référence est la loi jointe d'un Brownien réversible avec un coefficient de diffusion donné entre deux temps (Christian, 2013; Mariani, 2018; Ripani, 2017).

Je vais chercher à définir un processus de référence qui prenne en compte la dynamique des cellules dans l'espace d'expression des gènes. De façon plus réaliste qu'un processus diffusif. Pour cela, je vais présenter dans les prochaines sections comment je modélise l'activité d'un gène puis d'un réseau de gènes afin de pouvoir construire un tel processus.

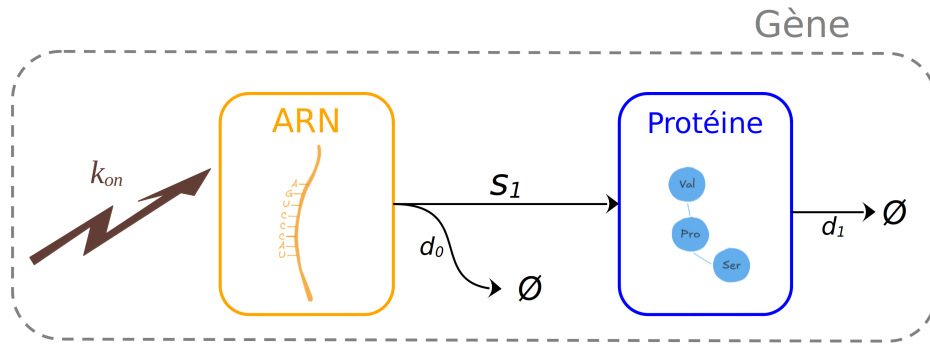
1.6 Modélisation d'un gène

Afin de mieux intégrer l'expression des gènes dans l'inférence des trajectoires des données scRNA-seq, je zoome d'abord sur l'activité d'un gène. Dans les données scRNA-seq, on sait que les snapshots suivent des distributions gamma ou Poisson-gamma (ou un mélange des deux) (Albayrak et al., 2016), je cherche alors un modèle qui reproduit ces distributions caractéristiques.

Le modèle de Peccoud et Ycart (1995) est inspiré de l'équation maîtresse de la chimie (CME) qui décrit les réactions chimiques d'un système. En effet, l'expression d'un gène se résume à un ensemble de réactions chimiques. Dans la revue de Boettiger (2013), de nombreux exemples de modèles de gènes sont cités, dont celui de Sánchez et Kondev (2008), qui généralise le modèle de Peccoud et Ycart (1995). En mathématiques, la dynamique temporelle des molécules, la création et la dégradation, est décrite par des processus markoviens de type naissance-mort (Boettiger, 2013). Dattani et Barahona ont proposé, en 2017, que le processus d'expression génique soit décrit par un modèle à deux réactions : la création d'ARNm par le promoteur et la dégradation de cet ARNm (Dattani & Barahona, 2017).

Le modèle à deux états (Peccoud & Ycart, 1995; Raj et al., 2006; Suter et al., 2011), inspiré du modèle introduit par Ko (1991), implémente la dynamique de l'ARNm. Il a la capacité à modéliser

A.



B.

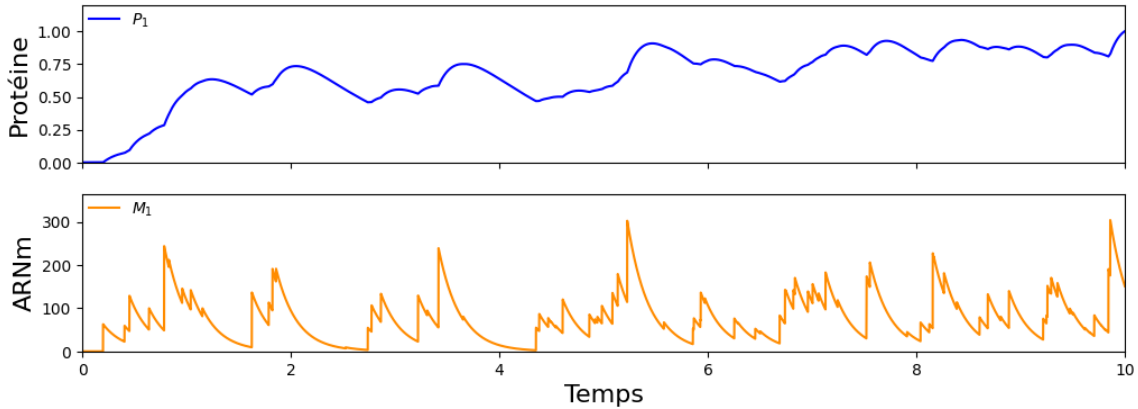


Figure 1.7. Modèle de gène à deux états. **A.** Représentation schématique du modèle Bursty pour un gène. La synthèse d'ARNm est activée de façon aléatoire à un taux k_{on} , défini dans (1.4), et la dynamique de l'ARNm (en orange) active la production des protéines (en bleu) avec un taux s_1 . Les niveaux d'ARNm et protéines dégradent respectivement à des taux d_0 et d_1 , selon le modèle (1.3), avec $d_0 \ll d_1$ (de l'ordre de 5 fois plus important). **B.** Exemple d'évolution temporelle des niveaux de protéines (haut en bleu) et d'ARNm (bas en orange), dans une cellule virtuelle.

l'expression d'un gène à l'échelle d'une cellule (Becskei et al., 2005; Raj et al., 2006; Raser & O'Shea, 2004; Suter et al., 2011). Le modèle à deux états décrit la création de l'ARN qui s'appuie sur un promoteur qui peut être dans deux états, soit actif soit inactif (Coulon et al., 2013). Quand le promoteur est inactif, il n'y a pas de synthèse d'ARNm et c'est son état le plus probable (Chong et al., 2014; Larson, 2011). Le temps passé en état actif est court, ce qui produit des "bursts" de courtes et intenses productions d'ARNm (Herbach et al., 2017). Ce modèle peut être décrit comme un processus de Markov déterministe par morceaux (PDMP). Introduit par Davis (1984), il est considéré comme un processus markovien hybride non-diffusif ; il a une composante discrète et une composante continue et modélise un mouvement déterministe ponctué de sauts aléatoires.

Le régime où le taux d'inactivation du promoteur est largement supérieur aux taux d'activation et de dégradation de l'ARN est celui qui est observé dans les expériences scRNA-seq (Albayrak et al., 2016; Raj et al., 2006; Richard et al., 2016; Suter et al., 2011; Viñuelas et al., 2013), ce qui permet une simplification mathématique (Crudu et al., 2012). Cette simplification, introduite par Herbach et al. (Herbach et al., 2017), aboutit à un modèle dit "Bursty". Un gène est modélisé

par sa quantité d'ARNm et de protéines : l'ARNm est synthétisé par des bursts stochastiques et est dégradé à un taux d_0 et les protéines sont synthétisées à partir de l'ARNm à un taux s_1 et sont dégradées à un taux d_1 (Figure 1.7A.).

Le modèle Bursty est caractérisé par une partie déterministe décrite par un système d'équations différentielles :

$$\begin{cases} \dot{M}(t) &= -d_0 M(t), \\ \dot{P}(t) &= s_1 M(t) - d_1 P(t), \end{cases} \quad (1.3)$$

où $M(t)$ est la quantité d'ARNm et $P(t)$ la quantité de protéines au temps t , et une partie stochastique où le burst de l'ARNm survient à un taux k_{on} , avec une hauteur suivant une loi exponentielle à taux constant (Herbach, 2021; Ventre, 2021), comme illustré dans la figure 1.7B. . Le modèle Bursty est choisi pour trois raisons : sa simplicité, il capture l'essentiel de la dynamique avec peu de paramètres ; son réalisme, il reproduit les distributions observées expérimentalement (Albayrak et al., 2016) ; et sa capacité à se coupler avec d'autres modèles Bursty pour reproduire les réseaux de régulation génique (Section 1.7).

Il existe une variante possible du modèle Bursty qui modélise uniquement les protéines (Friedman et al., 2006; Mackey et al., 2011; Pájaro et al., 2017), qu'on présente dans le chapitre 2.

1.7 Modélisation d'un réseau de régulation de gènes

Comme introduit dans la section 1.2, les gènes sont organisés en réseaux de régulation génique, notés GRN.

Kauffman (1969), dans son article fondateur de 1969, modélise le GRN par un réseau booléen où chaque gène est un nœud binaire, les règles d'interaction sont des fonctions logiques, et les attracteurs du réseau correspondent aux types cellulaires. Davidson et al. (2002) proposent que le GRN soit vu comme un graphe dirigé, où les gènes sont représentés par des nœuds et les arrêtes dirigées représentent l'action d'activation ou d'inhibition d'un gène envers un autre (ou lui même). Dibaëinia et Sinh proposent SERGIO qui intègre la structure du GRN (Dibaëinia & Sinh, 2020), mais ne retrouve pas la distribution gamma ou Poisson-gamma (ou un mélange des deux), caractéristique des distributions de l'expression d'un gène (Albayrak et al., 2016; Chen & Mar, 2018; Mar, 2019).

Ici, je garde ce formalisme des graphes dirigés mais les nœuds ne sont plus binaires et prennent en compte des quantités d'ARNm. Je présente comment le couplage de modèles de gènes, le modèle Bursty (1.3) crée un réseau de gènes dynamique.

On définit $\theta = \{\theta_{ji}\}_{j,i \in 1, \dots, G}$, où G est le nombre de gènes du GRN, la matrice d'interaction entre les gènes, où θ_{ji} désigne l'intensité de l'action du gène j sur le gène i (Herbach et al., 2017). Les paramètres θ_{ji} peuvent être positifs ou négatifs, selon que le gène j active ou inhibe le gène i . Un "burst" survient pour un certain gène i à un taux $k_{on,i}(P_1, \dots, P_G)$, où $(P_1, \dots, P_G)$ est le

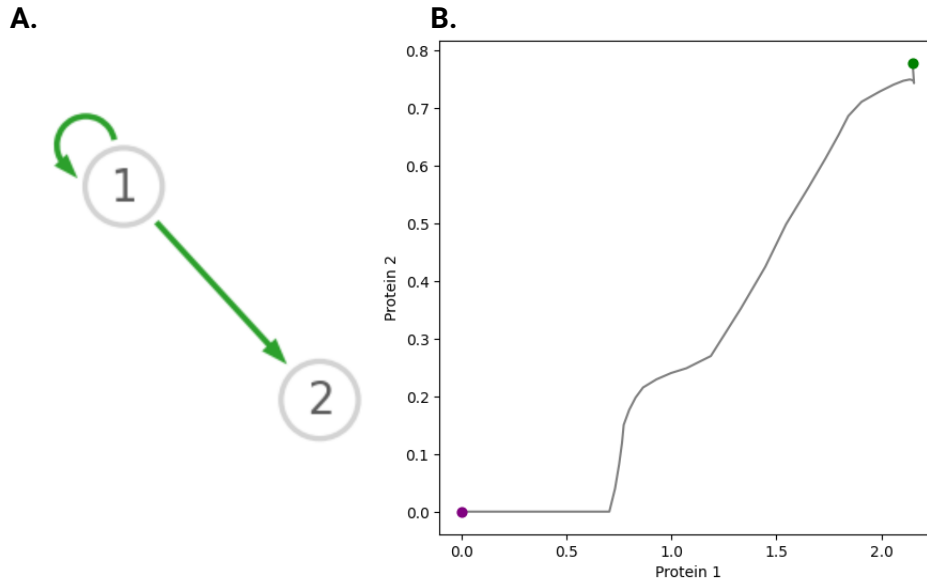


Figure 1.8. Exemple de GRN. **A.** Schéma d'un GRN à deux gènes, 1 et 2, où le gène 1 s'auto-active et active le gène 2 (flèches vertes). **B.** Trajectoire d'une cellule dans l'espace des protéines, qui commence à t_1 en violet et finit à t_2 en vert.

vecteur des niveaux de protéines de tous les gènes du GRN. On définit, pour le gène i (Herbach, 2021) :

$$k_{on,i}(P_1, \dots, P_G) = \frac{k_{0,i} + k_{1,i} \exp(\beta_i + \sum_{j=1}^G \theta_{ji} P_j)}{1 + \exp(\beta_i + \sum_{j=1}^G \theta_{ji} P_j)}, \quad (1.4)$$

où β_i est l'activité basale du gène i , et $k_{0,i}$ et $k_{1,i}$ sont, respectivement, les fréquences minimales et maximales d'un burst du gène i . Le modèle de GRN se compose de PDMPs couplés : chaque gène est modélisé par un PDMP, et les taux k_{on} de chaque PDMP dépendent des niveaux de protéines des autres gènes du réseau, ce qui traduit leurs interactions régulatrices. Le modèle de GRN peut être simulé de façon exacte, pour un GRN donné grâce au paquet Python HARISSA (Gaillard & Herbach, 2025; Herbach, 2021) ; des illustrations de la simulation sont présentées dans la figure 1.7B. et dans le prochain chapitre, dans la figure 2.10.

L'action des gènes les uns sur les autres influe sur les trajectoires des cellules dans l'espace d'expression génique. Je prends pour exemple un GRN à 2 gènes, où le gène 1 active le gène 2 (Figure 1.8A). Dans l'espace d'expression des gènes, une cellule qui exprime le gène 1, se déplace sur l'axe des abscisses. Puis, dû à la structure du GRN, l'expression du gène 1 va activer l'expression du gène 2, et ainsi la cellule va se décoller de l'axe des abscisses. Ses deux gènes sont exprimés et comme on le voit sur la Figure 1.8B., la cellule finit sa trajectoire dans le coin en haut à droite. Cela montre que même des interactions simples peuvent engendrer des trajectoires non linéaires.

1.8 Méthodologie de la thèse

Dans cette section, je présente la construction de la méthode implémentée, appelée *méthode d'inférence basée sur le modèle Bursty*. Le principe est le suivant : à partir de deux *snapshots* de données scRNA-seq, j'infère des trajectoires dans l'espace des protéines ou des gènes en résolvant le problème de Schrödinger avec un processus de référence qui prend en compte la dynamique des gènes.

Dans le reste de cette section, je présente comment je construis le processus de référence pour le problème de Schrödinger à partir de la diffusion et à partir du modèle Bursty. Puis, je présente les deux types de données que j'ai utilisés pour tester la méthode : des données simulées avec des GRN de petite taille (2 et 3 gènes) et des données expérimentales provenant de Semrau et al. (2017). Je détaille enfin la formalisation de la résolution du problème de Schrödinger. L'ensemble des résultats liés à cette méthode est présenté dans la section 2.2 (Figure 1.9).

On considère deux distributions μ et ν , représentant deux *snapshots* de données en scRNA-seq, composées de N cellules $x_1, \dots, x_N$ (distribution μ) et N' cellules $y_1, \dots, y_{N'}$ (distribution ν). Tous les x et y sont des vecteurs de $G \in \mathbb{N}$ éléments, chaque élément g est la quantité d'ARN et/ou de protéines du gène g . La valeur G représente le nombre de dimensions de l'espace dans lequel j'infère les trajectoires des cellules.

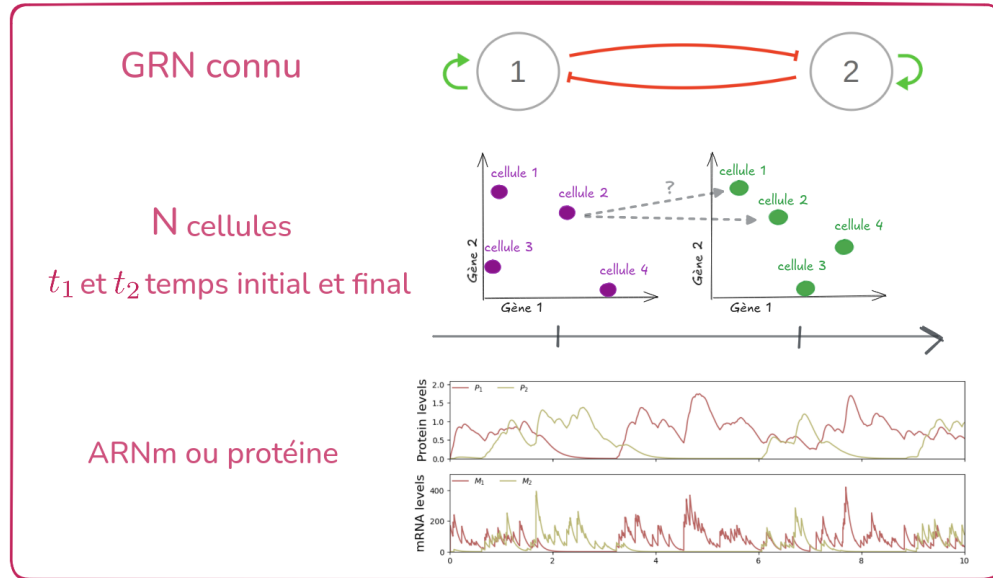
1.8.1 Processus de référence basé sur la diffusion

Avec l'objectif de comparer notre méthode à des méthodes existantes (section 1.5.2), je définis un processus de référence basé sur la diffusion. Dans la littérature, aucun travail n'utilise directement le problème de Schrödinger pour l'inférence de trajectoire ; les travaux de transport ou de couplage de données ScRNAseq (section 1.5.2) utilisent le transport optimal et le coût de transport associé est classiquement la distance euclidienne. Le lien entre la matrice C de coût de transport pour le transport optimal et le processus de référence basé sur la diffusion, noté D_{ref} , pour le problème de Schrödinger est :

$$D_{ref}(x, y) = \frac{1}{K} e^{-\frac{C(x, y)}{2\sigma^2(t_2 - t_1)}}, \quad (1.5)$$

avec K une constante telle que $\sum_{1 \leq i, j \leq N} D_{ref}(x_i, y_j) = 1$, x les éléments de l'espace de départ, y ceux de l'espace d'arrivée, σ^2 le coefficient de diffusion et $C(x, y)$ la matrice des distances euclidiennes entre x et y (Peyré & Cuturi, 2019b).

Données



Problème de Schrödinger

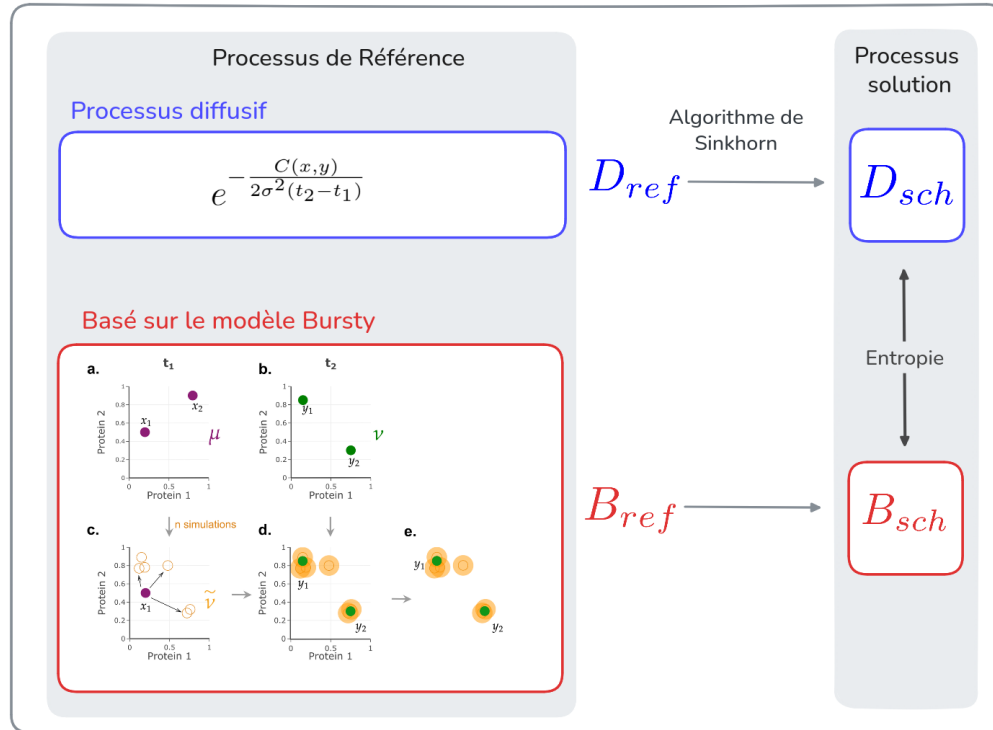


Figure 1.9. Représentation schématique de la méthodologie de cette thèse. La partie supérieure décrit les données nécessaires à la méthode : elles proviennent d'un jeu de données où le réseau de régulation génique (GRN) est connu ; l'exemple ici est le GRN toggle-switch, qui considère deux gènes qui s'inhibent l'un l'autre (flèche rouge) et s'auto-activent (flèche verte). Les données proviennent de mesures en cellule unique, effectuées à au moins deux temps que je représente dans l'espace d'expression des gènes (soit les niveaux d'expression de l'ARNm, soit les niveaux de protéines). Je considère le même nombre N de cellules aux temps t_1 et t_2 : les cellules sont observées dans l'espace transcriptomique et on cherche leur couplage entre t_1 (violet) et t_2 (vert). Avec HARISSA, je simule les niveaux d'expression des protéines (courbes Protein levels) et des ARNm (courbes mRNA levels) pour chaque gène, compte tenue du GRN. La partie inférieure décrit le cadre mathématique utilisé pour coupler les distributions cellulaires avec le problème de Schrödinger. Je présente les deux processus de référence proposés : le processus diffusif (en rouge), calculé à partir de la distance euclidienne d'après Peyré et Cuturi (2019b), D_{ref} , et le processus basé sur le modèle Bursty, approximé à partir de simulations, B_{ref} . Je résous le problème de Schrödinger avec l'algorithme de Sinkhorn et obtiens le couplage optimal de Schrödinger (D_{sch} ou B_{sch}). Les détails sont présentés dans la section 2.4.

1.8.2 Le modèle Bursty pour construire le processus de référence

J'utilise le modèle Bursty pour générer un processus de référence réaliste pour le problème de Schrödinger.

Ma première tentative d'approximation du processus de référence, testée pour 2 gènes, consistait à discrétiser l'espace d'arrivée et à calculer la probabilité qu'une cellule se trouve dans une case au temps t_2 en partant des positions à t_1 . Malheureusement, cette méthode était très coûteuse en temps de calcul et s'avérait difficile à étendre à davantage de gènes. J'ai alors utilisé une autre stratégie et développé une méthode que j'appellerai dans la suite *méthode d'inférence basée sur le modèle Bursty*. Cette méthode a fait l'objet d'une publication (Fournié et al., 2026), présentée au Chapitre 2.

À titre de comparaison, pour deux gènes, la première méthode nécessitait environ une fois et demie plus de temps de calcul que la méthode publiée, et le temps de calcul augmente avec le nombre de gènes. Les deux méthodes donnent toutefois des résultats équivalents car les principes sont proches : identifier le devenir le plus probable des cellules de μ dans la distribution ν , en se servant d'une distribution approchée par le modèle Bursty. La différence entre les méthodes repose sur la manière de lisser cette distribution approchée afin d'éviter qu'elle ne soit trop spécifique. La méthode par discrétisation procède en découpant l'espace en cases : plutôt que de représenter une distribution comme un nuage de points isolés, elle moyenne sur une case en agrégeant les points au sein de régions discrètes de l'espace. La *méthode d'inférence basée sur le modèle Bursty*, quant à elle, effectue un lissage en associant à chaque point une distribution gaussienne, ce qui sera détaillé ci-après. Le principe de la *méthode d'inférence basée sur le modèle Bursty* est illustré figure 2.13. Elle consiste à simuler n_{simu} fois le modèle Bursty avec pour condition initiale les coordonnées d'une cellule x_i de la distribution μ au temps t_1 , pour toutes les cellules i , avec i un entier de 1 à N . On obtient une nouvelle distribution, qu'on note $\tilde{\nu}$, qui représente les potentiels états futurs des cellules x_i . Avec la fonction `kernel()` de Python (Pedregosa et al., 2011), pour chaque point y_i de la distribution ν mesurée au temps t_2 , j'évalue la densité estimée en ce point de $\tilde{\nu}$. Ainsi, on obtient une approximation du processus de référence entre μ et ν . On en déduit quelles sont les chances d'une cellule x_i de devenir une cellule y_i selon le modèle. J'obtiens le processus de référence que je note B_{ref} .

Dans la Figure 1.10, je représente le temps computationnel pour calculer le processus de référence à partir du modèle Bursty dans deux situations : quand j'augmente le nombre de cellules (Figure 1.10A.) pour un réseau de 2 gènes avec les 4 interactions possibles (le "toggle switch" que je définis dans la sous-section suivante 1.8.3), et quand j'augmente le nombre de gènes et d'interactions (Figure 1.10B.) pour 100 cellules. Pour la Figure 1.10B., je construis les GRNs sur le même principe que le "toggle-switch", quel que soit le nombre de gènes, il y a le maximum d'interactions possibles. On constate que le temps de calcul reste raisonnable lorsque

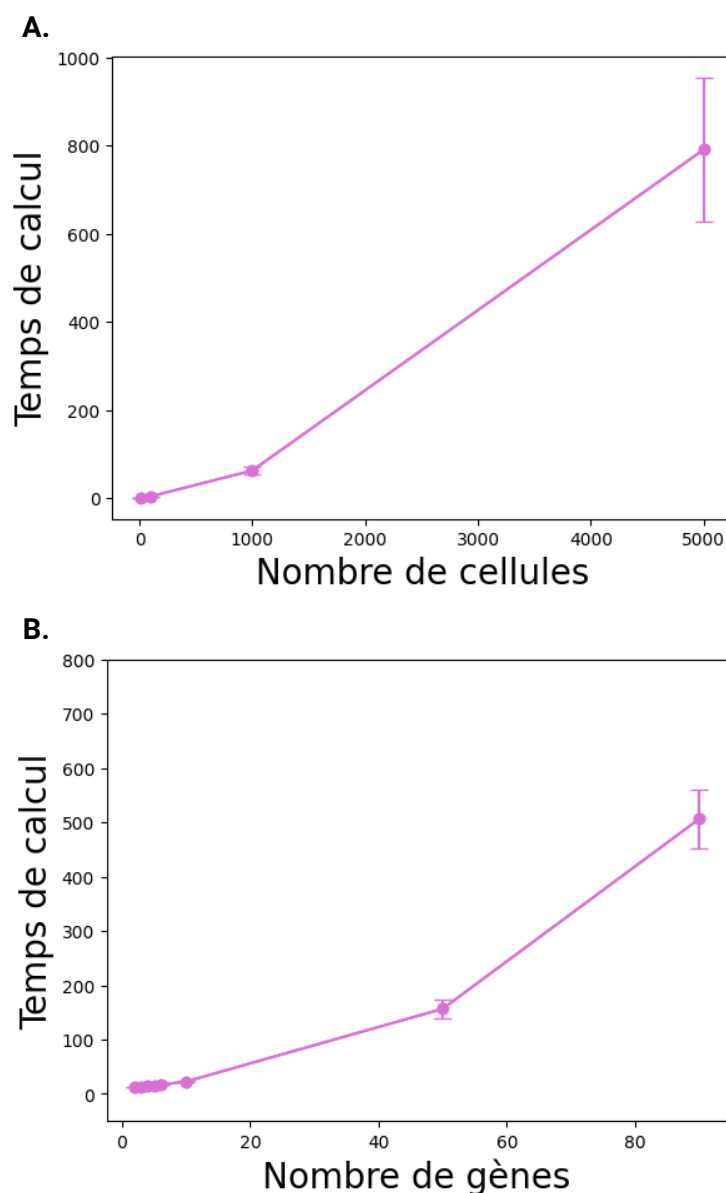


Figure 1.10. Temps de calcul moyen (en secondes) en fonction **A.** du nombre de cellules et **B.** du nombre de gènes. La courbe représente le temps de calcul moyen du processus de référence Bursty. La moyenne est calculée sur 5 simulations. Tous les autres paramètres sont fixés et détaillés dans le tableau 2.1.

le nombre de cellules ou de gènes augmente. Nous n'avons pas cherché à mettre en œuvre une stratégie pour le réduire. Toutefois, il est possible de paralléliser le calcul du processus de référence basé sur le modèle Bursty si le nombre de simulations, de cellules ou de gènes est trop élevé.

En conséquence, le temps de calcul du processus de référence n'apparaît pas comme une limitation de la méthode. On pourrait considérer que la principale "limitation" de cette méthode est qu'il est nécessaire de connaître un GRN sous-jacent au processus observé, mais il s'agit d'une autre question, que je développerai dans la discussion dans le chapitre 3.

1.8.3 Application sur données simulées

Le GRN jouet sur lequel j'ai benchmarké la *méthode d'inférence basée sur le modèle Bursty* est le "toggle-switch" (Figures 2.10 et 2.6). Il s'agit d'un GRN à deux gènes, où les gènes s'autoactivent et s'inhibent mutuellement. Ce modèle jouet a été choisi pour sa simplicité et le fait qu'on ne peut pas savoir quel gène va prendre le dessus sur l'autre. Comme illustré par la Figure 1.11, ce modèle jouet est simulé N fois avec HARISSA (Herbach, 2021) et génère N trajectoires cellulaires temporelles. Dans ce cadre, j'ai considéré le cas $N = N'$, c'est à dire qu'il y a le même nombre de cellules dans la distribution de départ et dans la distribution d'arrivée. J'ai déjà mentionné que le modèle Bursty génère des distributions réalistes (Gaillard & Herbach, 2025; Herbach et al., 2017). En extrayant des données simulées les quantités d'ARNm à un temps donné, on reconstruit l'équivalent d'un snapshot de données de scRNA-seq. On peut alors représenter les cellules dans l'espace d'expression des gènes en 2D, comme illustré dans la figure 1.5.

Le modèle Bursty simule à la fois les quantités d'ARNm et de protéines. La technologie du scRNA-seq fournit des informations sur les quantités d'ARNm. Ce serait logique de se concentrer uniquement sur le nombre d'ARNm. Cependant, il n'est pas possible de réduire le modèle Bursty à un modèle ne décrivant que l'ARNm et de préserver les propriétés mathématiques, car les ARNm sont généralement beaucoup moins stables que les protéines et ne permettraient donc pas de préserver la propriété de mémoire du réseau de régulation génique.

En simulant le modèle Bursty, je m'intéresse aux deux cas de figure : celui où on ne se concentre que sur les données protéines et celui où on ne se concentre que sur les données ARNm. Dans le cas où on étudie les trajectoires à partir des mesures de protéines, on se place dans l'espace d'expression des protéines (Figure 1.8). L'évolution des protéines est plus lente que celle de l'ARNm. Dans le cas où on étudie les trajectoires à partir des données ARNm, on se place dans l'espace d'expression des gènes. Je considère, dans la majorité des cas, le modèle Bursty complet, c'est à dire, je considère la quantité d'ARNm et de protéines, même si je ne traite qu'un seul type de données. Dans la section 2.4, je présente et étudie le modèle réduit qui ne s'appuie que sur les données des protéines.

Pour construire les données simulées, je simule le modèle Bursty avec les paramètres d'un réseau de gènes. Ces paramètres sont donc connus. Ainsi, on peut résoudre le problème de Schrödinger en testant l'impact des paramètres du modèle Bursty sur le résultat de l'inférence.

En fonction du temps choisi pour extraire les deux snapshots, le système peut être dans différents "états" : transitoire ou stationnaire. J'appelle état du modèle Bursty une distribution (stationnaire ou transitoire) dont on ne connaît pas la loi, mais qui existe car le processus sous-jacent est bien défini (c'est un processus de Markov à fréquence de bursts bornée). La dynamique du système est stochastique, on ne peut pas vraiment parler d'état au sens déter-

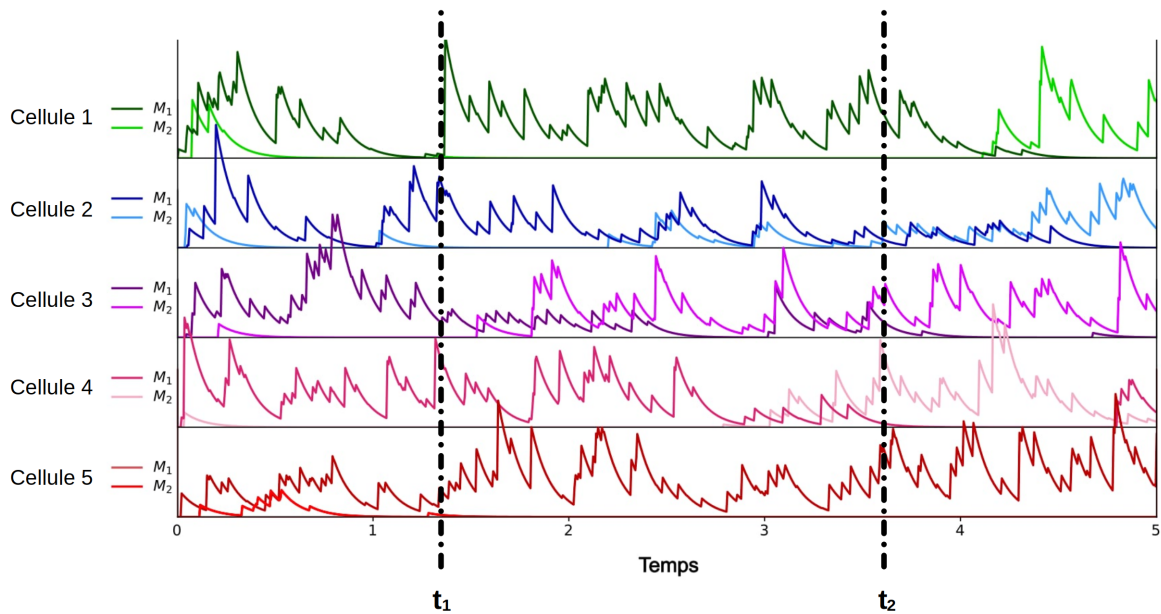


Figure 1.11. Représentation schématique de la construction de données synthétiques grâce à HARRISSA. Je représente $N = 5$ trajectoires de cellules, pour 2 gènes (M_1 et M_2), orchestrés par un toggle-switch et, j'extrais les données ARNm au temps t_1 et t_2 . Les paramètres sont fixés et détaillés dans le tableau 2.1.

ministre (Huang, 2010; Ventre, 2021). Mais si on prend l'exemple du toggle-switch, dans le cas stationnaire, après un certains temps de simulation, on observe deux clusters de cellules : la distribution stationnaire est bimodale. Les cellules évoluent soit au sein de leur cluster (la quantité d'une seule protéine augmente ou diminue, tandis que l'autre est très peu exprimée), soit entre les deux clusters (une protéine prend l'avantage sur l'autre). Dans le cas transitoire, les deux temps choisis pour le snapshot sont très proches l'un de l'autre et proches de 0 et les conditions initiales ont été choisies de sorte à ne pas correspondre à une distribution stationnaire, à $(0, 0)$ et donc la distribution des cellules est assez loin de la distribution stationnaire.

Même si l'essentiel de mon travail sur données simulées a été réalisé à partir du toggle-switch, j'ai également étudié les performances de la *méthode d'inférence basée sur le modèle Bursty* sur d'autres GRN à 2 et 3 gènes, afin notamment de tester sa robustesse et d'en évaluer l'extension en termes de coûts de calcul.

1.8.4 Résolution du problème de Schrödinger

Grâce à l'algorithme de Sinkhorn (1967) (voir section 1.5.3), on résout le problème de Schrödinger avec deux processus de référence : celui basé sur la diffusion (section 1.8.1), D_{ref} et celui basé sur le modèle Bursty (section 1.8.2), B_{ref} , et on obtient deux processus solution, nommés respectivement D_{sch} et B_{sch} . Les deux solutions sont alors comparées en utilisant l'entropie (1.1). L'entropie est un outil classique qui a des propriétés intéressantes pour comparer des processus (Cover & Thomas, 2005; Kullback & Leibler, 1951).

Dans le cas des données simulées, on a accès aux couplages réels entre les distributions de

départ et d'arrivée, qu'on peut utiliser pour déterminer la validité de la solution du problème de Schrödinger. Cependant, le couplage deux à deux n'est pas le plus instructif, car, dans l'optique d'appliquer la méthode à des données expérimentales, les cellules sont détruites lors de l'expérience ; le couplage deux à deux n'est donc pas réaliste. Nous cherchons donc à identifier des tendances de couplage et non des couplages réels, c'est à dire, pour une cellule au temps initial, identifier les cellules couplées de sorte qu'elles soient proches, dans l'espace des gènes, de la cellule réellement couplée.

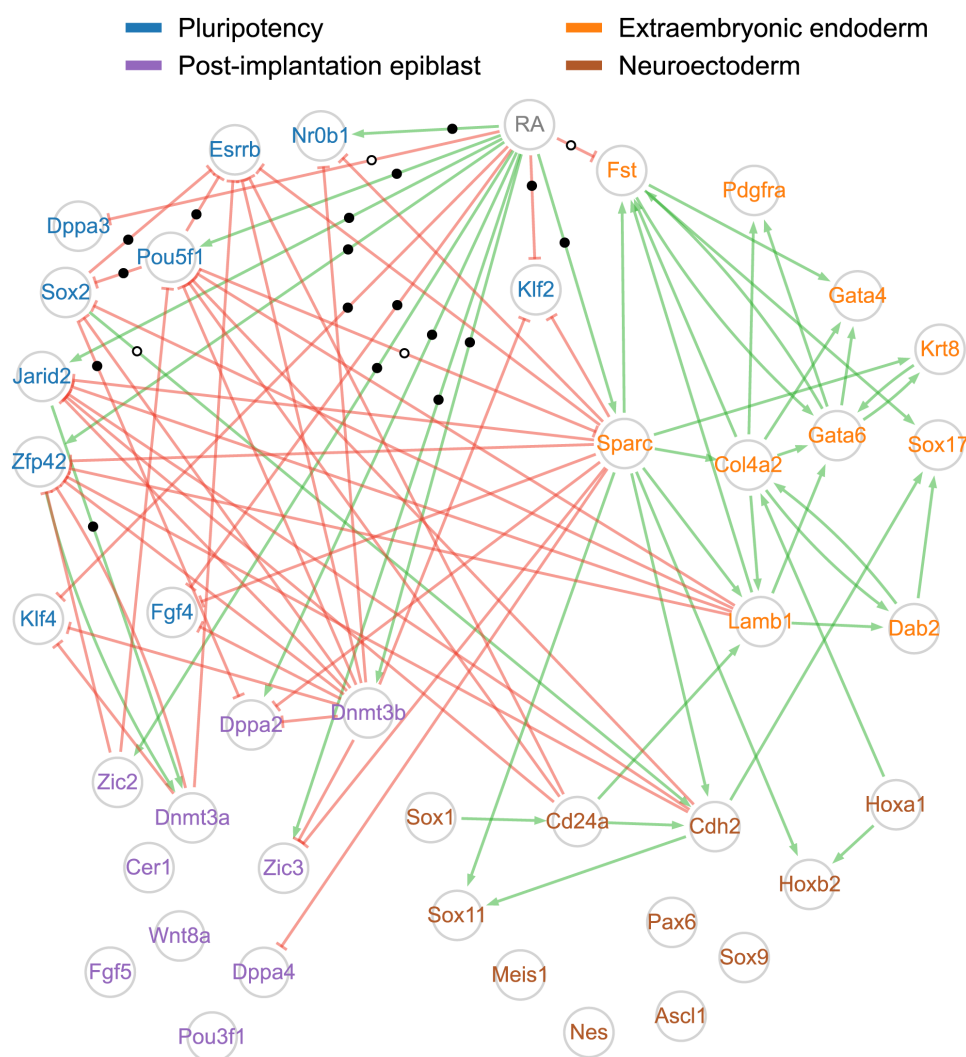


Figure 1.12. GRN inféré avec CARDAMOM (Ventre et al., 2023) à partir des données de Semrau et al. (Semrau et al., 2017). Le GRN est composé de 41 gènes, qui sont colorés selon quatre groupes correspondant à différents états cellulaires (pluripotence, épiblaste post-implantation, neuroectoderme, endoderme extra-embryonnaire) et un stimulus, noté "RA". Les flèches vertes représentent l'activation d'un gène par un autre et les flèches rouges représentent l'inhibition d'un gène par un autre. Les interactions confirmées par expérience biologique sont marquées d'un point noir et celles qui ne le sont pas, sont marquées d'un point blanc. Les interactions pour lesquelles aucune information fiable n'est disponible ne sont pas marquées.

1.8.5 Application sur des données expérimentales (Semrau et al., 2017)

Après avoir testé la *méthode d'inférence basée sur le modèle Bursty* sur plusieurs GRN de petites tailles, je l'ai appliquée aux données expérimentales extraites de Semrau et al. (2017). Les données *in vitro* proviennent d'une analyse scRNA-seq de cellules souches embryonnaires de souris exposées à l'acide rétinoïque, un déclencheur de la différenciation cellulaire. Les données sont restreintes à 41 gènes d'intérêt et regroupent neuf snapshots, répartis entre 0h et 96h. On peut y suivre la sortie de la pluripotence et l'engagement vers des lignées cellulaires : les cellules neuroectodermiques et les cellules endodermiques extra-embryonnaires.

Pour appliquer notre méthode sur les données expérimentales, il est nécessaire d'inférer un GRN sous-jacent (Figure 1.12). On peut, pour cela, utiliser le paquet Python CARDAMOM (Ventre et al., 2023) qui permet d'inférer un GRN à partir de données ScRNAseq.

L'objectif ensuite a été de vérifier la méthode sur les données réelles. Contrairement aux données simulées, nous n'avons ni accès aux vraies trajectoires ni au processus sous-jacent réel. En s'inspirant du protocole de Schiebinger et al. (2019), j'ai cherché à interpoler une distribution à un temps intermédiaire. Comme l'illustre la figure 1.13, j'ai choisi deux distributions à des temps $t_1 = 0h$ et $t_3 = 72h$ et interpolé une distribution à $t_2 = 36h$. Il est ainsi possible de comparer la distribution interpolée à la distribution expérimentale à t_2 , en utilisant la distance de Wasserstein ou *Earth Mover's Distance* (EMD). Je calcule l'EMD entre la distribution expérimentale et respectivement la distribution interpolée basée sur la diffusion, EMD_D et la distribution interpolée basée sur le modèle Bursty, EMD_B . De plus, je compare ces deux distances avec : $\Delta_{EMD} = EMD_D - EMD_B$.

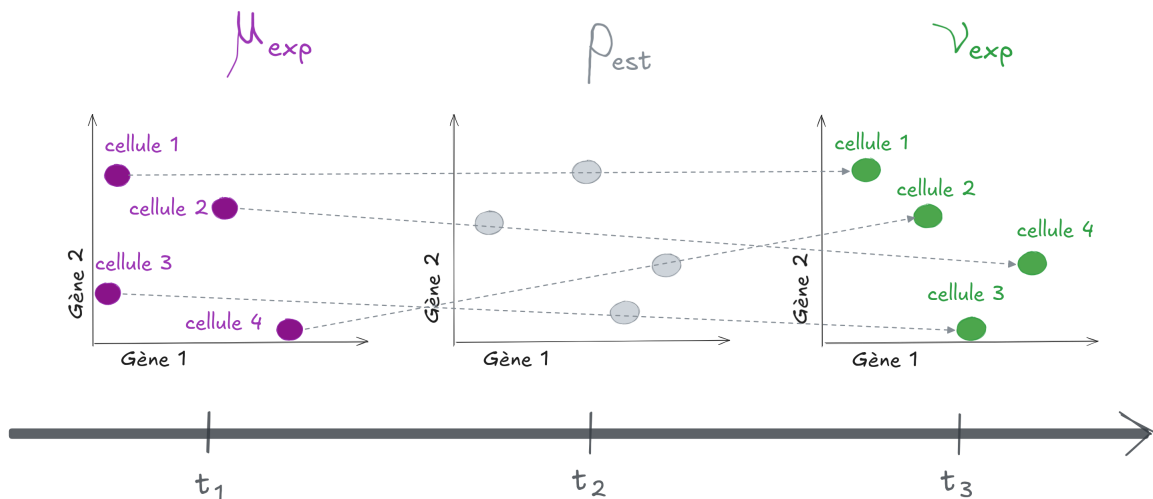


Figure 1.13. Représentation schématique de l'interpolation dans l'espace d'expression des gènes en 2 dimensions. À partir des deux distributions expérimentales μ_{exp} et ν_{exp} et leur couplage, j'en déduis à un temps intermédiaire (t_2), une distribution estimée ρ_{est} .

Le Chapitre suivant présente les principaux résultats de ma thèse (Fournié et al., 2026). Il s'agit du manuscrit de l'article intitulé "*Cell trajectory inference based on schrödinger problem and a mechanistic model of stochastic gene expression*" de Fournié C., Ventre E., Herbach U., Baradat A., Gandrillon O. et Crauste F. publié dans **npj Systems Biology and Applications** en 2026 (doi : doi.org/10.1038/s41540-026-00743-x).

2

Cell Trajectory Inference based on Schrödinger Problem and a Mechanistic Model of Stochastic Gene Expression

2.1	Introduction	38
2.2	Results	41
2.2.1	Schrödinger problem for a single gene	42
2.2.2	Case study: the toggle switch	44
2.2.3	Analysis for other GRNs	49
2.2.4	Application to experimental data from Semrau et al. (2017)	50
2.3	Discussion	54
2.4	Methods	56
2.4.1	Gene dynamics	56
2.4.2	Reduced model with analytical solution	58
2.4.3	Coupling between time points	60
2.4.4	Schrödinger problem	63
2.4.5	Performances of the models.	64
2.4.6	Application to Semrau et al. (2017) dataset	65
2.5	Data Availability	67
2.6	Code Availability	67
2.7	Acknowledgments	67
2.8	Author Contributions	68
2.9	Competing Interests	68
2.10	Supplementary information	69
2.10.1	Analytical solution for the reduced Bursty model	69
2.10.2	Sensitivity results	71
2.10.3	The stationary case	73
2.10.4	Interpolation parameter α and empirical measures	74
2.10.5	Comparison of the Bursty model-based inference method and the Diffusion-based inference method for the experimental data	75

Abstract

Cellular differentiation is the biological process that leads a cell to opt for a particular cellular identity. Recently, single-cell RNA-sequencing has enabled the simultaneous measurement of gene expression levels at specific times for a large number of individual cells and a large number of genes. Repeating such measurements at different time points gives then access to the temporal variation, or transport, of a distribution on a gene expression space. The whole temporal trajectory of distributions thus characterizes the differentiation process at population level, but trajectories of individual cells are still out of reach since most measurement techniques are destructive.

The optimal transport theory that has been used so far to infer cellular differentiation trajectories from time-stamped single-cell RNA-seq data involves solving the so-called Schrödinger problem in its most common version. This implies assuming that cells move, in the gene expression space, by diffusion. Yet, real gene dynamics are much more complex.

In the present work, we assume that mRNA dynamics are characterized by brief and important production of RNA, with long periods of inactivity in between, and consider the so-called Bursty model of gene dynamics. We use this model to define a reference process for the Schrödinger problem. By comparing the solutions of the Schrödinger problems with a Diffusive and a Bursty reference process, under different conditions, we show that the Bursty model provides a better approximation of the underlying gene dynamics than the standard Diffusive process when inferring cell trajectories.

2.1 Introduction

While two cells in a multicellular organism may possess the same genome, they can express their genes differently, leading to distinct cellular outcomes. Cellular differentiation is the biological process that leads a cell to opt for a particular cellular identity (Clevers et al., 2017). In contemporary biology, the cellular identity is defined by the expression of cell markers, mostly surface proteins.

Recently, single-cell RNA sequencing (scRNA-seq) has enabled the simultaneous measurement of gene expression levels for up to millions of individual cells and tens of thousand of genes (Klein et al., 2015; Macosko et al., 2015; Tanay & Regev, 2017). Thanks to this technique, a cell can be represented in a high-dimensional space, where each dimension corresponds to the expression level of a given gene, so the space dimension equals the number of genes. Re-

peating such measurements at different time points gives access to the temporal variation, or transport, of a distribution on a gene expression space. The entire temporal trajectory of the distribution characterizes the differentiation process at population level, but trajectories of individual cells are still out of reach since measurement process is destructive (Schiebinger et al., 2019). Hence the goal to develop tools to infer trajectories between measurements to reconstruct the differentiation process has become crucial. Since cells are destroyed in the data acquisition process, inference of cell trajectories allows to determine potential cell fate under the assumption that the cell sample used at each measurement time is representative of the whole cell population.

In practice, gene expression is measured either as the number of mRNA molecules or the level of protein expression. Hence, the above-mentioned gene space is either a mRNA counts space or a protein level space. While the differentiation process is fundamentally a continuous process, most of the experimental measurements provide access only to discrete temporal distributions. In order to study a differentiation process and more specifically to infer differentiation pathways, many works focused on the reconstruction of the continuous transcriptomics evolution based on such distributions.

Schiebinger et al. (2019) were pioneers in using the optimal transport (OT) theory to infer such trajectories from time-stamped transcriptomics data, introduced in the Waddington Optimal Transport (WOT) algorithm. WOT enables the reconstruction of a “timeline” between two cell distributions measured at different times. WOT solves the OT problem in a standard way: by minimizing a cost function defined as the squared Euclidean distance between cells in a PCA-reduced space.

Since then, several works proposed methods for cell trajectory inference inspired by OT theory (Bunne et al., 2024; Jiang et al., 2022; Lange et al., 2024; Lavenant et al., 2024; Liu et al., 2020; Lotfollahi et al., 2023; Prasad et al., 2022; Sha et al., 2023; Tong et al., 2020; Tronstad et al., 2025; Yachimura et al., 2024; Zhu et al., 2025). Authors of OT-based studies chose to work with quadratic OT, which consists in using the squared Euclidean distance as a cost in the OT problem (see Section 2.4.3). More specifically, they used the so-called entropic approximation of OT—which is equivalent to solving the Schrödinger problem—in order to use the efficient Sinkhorn algorithm (Peyré & Cuturi, 2019a; Sinkhorn, 1967). The Schrödinger problem consists in finding the stochastic process that is the closest to a given reference process while fitting the observations. Originally, the aim of the Schrödinger problem was to interpolate between cell distributions observed at discrete timepoints in order to reconstruct the full time-varying distribution. The mathematical theory supporting this problem ensures that this optimal interpolation is associated to an optimal coupling corresponding to the transition probability kernel between each pair of distributions of the underlying stochastic process. Thus, the destructive scRNA-seq measurements are perfectly aligned with the Schrödinger problem: the inferred cou-

pling characterizes the optimal process describing the observed data under the hypothesis of the Schrödinger problem.

The reference process used in the Schrödinger problem is in general a Brownian motion, hence it is implicitly assumed that cell trajectories are close to diffusion. In some cases, distances different from the square Euclidean distance have been considered, mostly for taking into account additional information like lineage information (Lange et al., 2024; Majima et al., 2024) or coexpression between genes (Du et al., 2019).

Some methods nevertheless used a different cost function when solving the OT problem. For instance, Liu et al. (2020) criticized the use of Euclidean distance when applying OT theory and proposed an implementation of a learned cost function in the OT problem based on the Sinkhorn algorithm (Sinkhorn, 1967). However, they do not use OT theory to infer cell trajectories, instead they focus on alignment of scRNA-seq datasets. Additionally, the learned cost is non-interpretable, a property that is shared by computational methods based on machine learning. In order to reduce the computational cost and gain interpretability, scEGOT (Yachimura et al., 2024) combines OT with a cell state graph. By adding Gaussian mixture distribution on each cell's dots, scEGOT allows to consider cell clusters instead of individual cells. The Gaussian mixture makes the results interpretable, however their objectives are clearly different from ours because the authors use Euclidean distance between Gaussian distributions to infer trajectories, relying once again on a standard approach.

The main idea of the present paper is to assert that in order to reconstruct cell differentiation trajectories, a more realistic, quantitative model of stochastic gene expression should be considered as reference process. We focus on solving the Schrödinger problem while using an alternative reference process that would be closer to realistic gene dynamics. Indeed, the Diffusive process generates unrealistic dynamics for proteins and mRNA (Magaña-López et al., 2024). The Schrödinger problem is not seen here as an approximation of a transport problem, but as a way to identify a process close to an a priori quantitative model that would fit the observations.

When examined at the single-cell level, gene expression is undoubtedly a stochastic process (Chalancon et al., 2012; Eldar & Elowitz, 2010; Huang, 2010; McCullagh et al., 2009; Niepel et al., 2009; Raj & van Oudenaarden, 2008). The simplest model that can correctly reproduce experimental data distribution is the well-known two-state model of gene expression (Peccoud & Ycart, 1995; Raj et al., 2006; Suter et al., 2011), a refinement of the model introduced in 1991 by Ko (1991). In this model, a gene is described by its promoter which can be either active or inactive—possibly representing a transcription complex being “bound” or “unbound” although it may be more complicated (Coulon et al., 2013)—with mRNA being transcribed only during the active periods.

Genes may mechanistically influence the activation rate of other genes, thus creating a Gene

Regulatory Network (GRN). We consider an extension of the two-state model for several genes, that takes into account the GRN structure. For certain parameter values, mRNA dynamics are characterized by brief and important production of RNA, with long periods of inactivity in between: this regime corresponds to the widely observed phenomenon called transcriptional bursting (Suter et al., 2011). Herbach et al. (2017) introduced a mathematical framework for modeling GRN driven by transcriptional bursting, which we call the Bursty model here. This model has been shown to reproduce realistic gene dynamics at the single-cell level (Albayrak et al., 2016; Ventre et al., 2023) and will provide our reference process.

In this work we compare the performances of a Diffusive process, associated to standard OT theory, and a reference process based on the Bursty model to solve the Schrödinger problem, when initial and final distributions are based on realistic, simulated data. For the sake of clarity, in the following we refer to our approach as the Bursty model-based inference method. We first explore the case of a single gene, then of a specific two-gene GRN, the toggle switch, and finally more complex three-gene GRNs. We focus on the influence of some indicators, namely the time difference between the measurements of the two distributions and the number of cells in the distributions. We show that using the Bursty model as a reference process provides a better solution to the Schrödinger problem for all tested networks. Additionally, we show that the Bursty model-based inference method performs better than a Diffusion-based inference approach on a dataset made of mRNA counts for 41 genes.

2.2 Results

We consider two distributions of protein counts obtained from a fixed number of cells, simulated from the Bursty model (see Section 2.4.1) and measured respectively at times t_1 and t_2 ($t_2 > t_1$). Noticeably, we focus only on protein counts since protein dynamics, contrary to mRNA dynamics, allow to preserve the markovian property in the Bursty model therefore providing a rigorous mathematical framework. We solve the associated Schrödinger problem for several reference processes, including the Bursty process and the standard Diffusive process. Solutions of the Schrödinger problem are compared using an entropy function. If the two processes are similarly able to correctly infer trajectories, the entropy equals or nears zero.

When the final distribution is measured a “long” time after the initial distribution, the Bursty model has reached a so-called stationary state, and both initial and final distributions can be considered independent. In this case, trajectory inference should depend less on the selected reference coupling. Consequently, we focus our study on distributions in transient states, meaning the time difference $t_2 - t_1$ is small enough and t_1 is close to the initialisation of the simulations, as represented in Figure 2.12. Details are provided in Section 2.4.

Noticeably, since we are working with data that we generated, so-called *simulated datasets*,

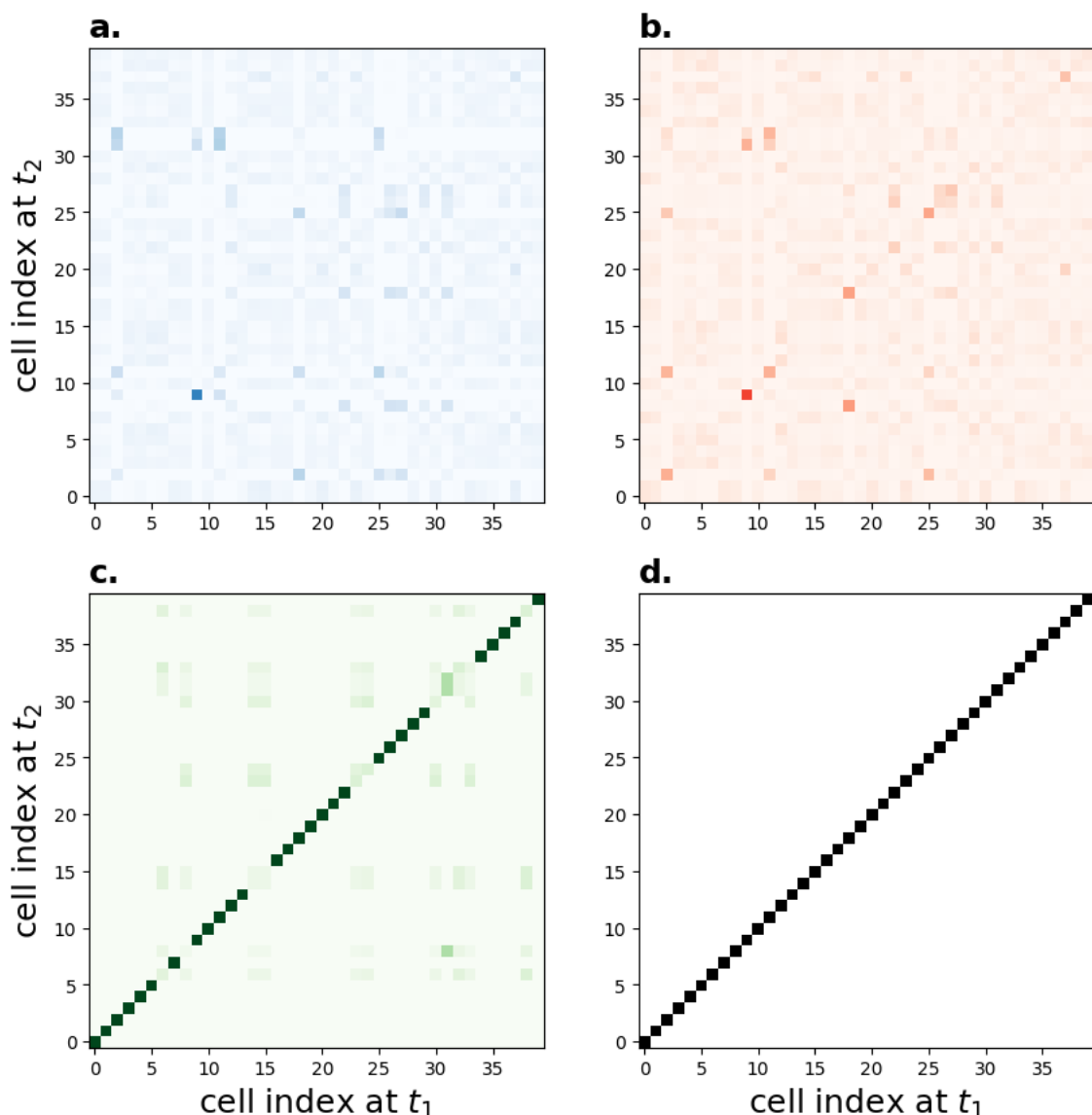


Figure 2.1. Solution of the Schrödinger problem for a single gene. a. D_{sch} . b. B_{sch} . c. RB_{sch} . d. Ground truth. Each colored square represents the probability of the cell index at t_2 starting from the cell index at t_1 for $N = 40$ cells: the darker the box, the greater the probability. All matrices sum to 1. Parameters are given in Table 2.1.

it is possible and efficient to define a ground truth. More precisely, it is convenient to use an identity-based ground truth, defined such that cell i at t_1 corresponds to cell i at t_2 . However, when using experimental single cell datasets, no ground truth is available and only marginal observations can be used to infer the cell differentiation process. In this case, no reference exists to validate the result of any method.

2.2.1 Schrödinger problem for a single gene

We first considered the case of a single gene with no feedback, as a toy model to illustrate the influence of the underlying model of gene dynamics on trajectory inference.

Figure 2.1 shows the solutions of the Schrödinger problem computed with different approximated processes: a Diffusive model, D_{sch} (Figure 2.1a), the non-reduced Bursty model, B_{sch}

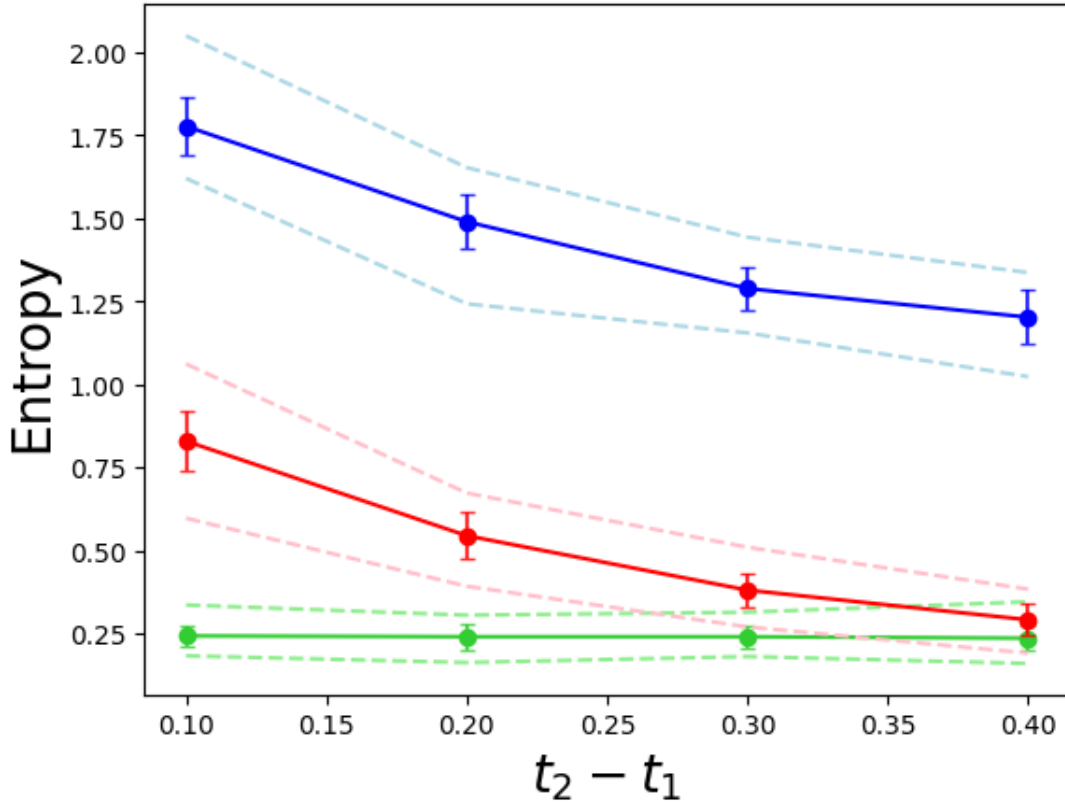


Figure 2.2. Entropies between reference couplings as a function of the time interval for a one-gene GRN. Mean and standard deviation (straight lines) of entropies $H_{D,R}$ (blue), $H_{B,R}$ (red) and H_R (green) are displayed for various values of the time difference $t_2 - t_1$, when the gene dynamics model is in transient state. Additionally, minimum and maximum values over 50 iterations (dashed lines) are also displayed. Other parameters are detailed in Table 2.1.

(Figure 2.1b), and the reduced Bursty model, RB_{sch} (Figure 2.1c). The ground truth of the simulated dataset (Figure 2.1d) is a coupling of cells with the same index (cell i in the initial distribution is cell i in the final distribution). We observe that RB_{sch} yields the closest result to the ground truth of the simulated dataset (Figure 2.1c), in the form of a very accurate transition matrix, while D_{sch} yields the furthest (Figure 2.1a). A darker diagonal is observed for B_{sch} (Figure 2.1b), meaning that this process better approximates the ground truth than D_{sch} . Noticeably, a limitation of the representation in Figure 2.1 is that the process could be not expressed (light color) and yet be relevant. For example, if two cells are close in protein space at both t_1 and t_2 , then the process between those cells could be mixed up but relevant.

Figure 2.2 presents the entropy between D_{sch} and RB_{sch} , denoted by $H_{D,R}$, and the entropy between B_{sch} and RB_{sch} , denoted by $H_{B,R}$ (see Section 2.4.5). The curve H_R also appears in Figure 2.2, it serves as a control by computing the entropy between the reference coupling and its approximation. The initial distribution is measured at time $t_1 = 0.1$ h, and the final distribution is computed at various times t_2 in the range $[0.2; 0.5]$ h.

We observe that quantities $H_{D,R}$ and $H_{B,R}$ decrease as $t_2 - t_1$ increases and H_R remains constant. We also observe that $H_{D,R}$ is always higher than the entropy between B_{sch} and

RB_{sch} , and remains much higher than H_R even for large values of $t_2 - t_1$ (here 0.4 h). Entropy $H_{B,R}$ is always larger than H_R , yet it is much smaller than $H_{D,R}$, and for $t_2 - t_1 = 0.4$ h there is no difference between $H_{B,R}$ and H_R . Both the exact temporal distribution of the process and its approximation provide comparable results.

The study of this toy model highlights that using the exact temporal distribution of the reference process to solve the Schrödinger problem provides much better results than using the approximation of a gene dynamics model. However, it is a complex, most of the time impossible, task to find the analytical expression of this distribution for a GRN comprising at least two connected genes. Consequently, for a more relevant GRN, we will focus on the estimations provided by the reference process of the non-reduced Bursty model and the Diffusive process.

2.2.2 Case study: the toggle switch

We focused on a two-gene GRN, called ‘toggle switch’, introduced in Section 2.4.1 and illustrated in Figure 2.6b (blue GRN). This specific GRN consists of two self-activating genes that inhibit each other.

2.2.2.1 Influence of the time interval and the number of cells

We first investigated the influence of the time between measurements of the initial and the final distributions (i.e. $t_2 - t_1$) on the ability of two models, the Bursty and the Diffusive models, to correctly transport cell distribution.

Figure 2.3a shows values of entropies $H_{D,B}$ and H_B as functions of the time difference $t_2 - t_1$. The $H_{D,B}$ curve is always above the H_B curve and remains approximately constant and independent of the time difference. The entropy between the two processes is the same for all $t_2 - t_1$ values: whatever the time difference, the Diffusive process infers trajectories with the same uncertainty. For the time difference $t_2 - t_1 = 0.2$ h, the entropy $H_{D,B}$ has a larger standard deviation, due to one extreme value.

Although it is more relevant to focus on the transient state (Figure 2.3a), it may be informative to question what happens in the stationary state. In this case, cells may either remain in their basin of attraction or switch to the other basin of attraction. The basin of attraction is a notion related to deterministic dynamics, consequently in the current framework where we consider stochastic models of gene dynamics it refers to the basin of attraction of the coarse approximation in the form of a system of ordinary differential equations of a piecewise-deterministic Markov process (Ventre et al., 2021). Results in the stationary state are illustrated in Supplementary File (Section 2.10.3), they highlight conclusions similar to the ones obtained for the transient state. Although the Diffusive model-based inference method is less accurate than the Bursty model-based inference method for inferring trajectories, its ability increases when the

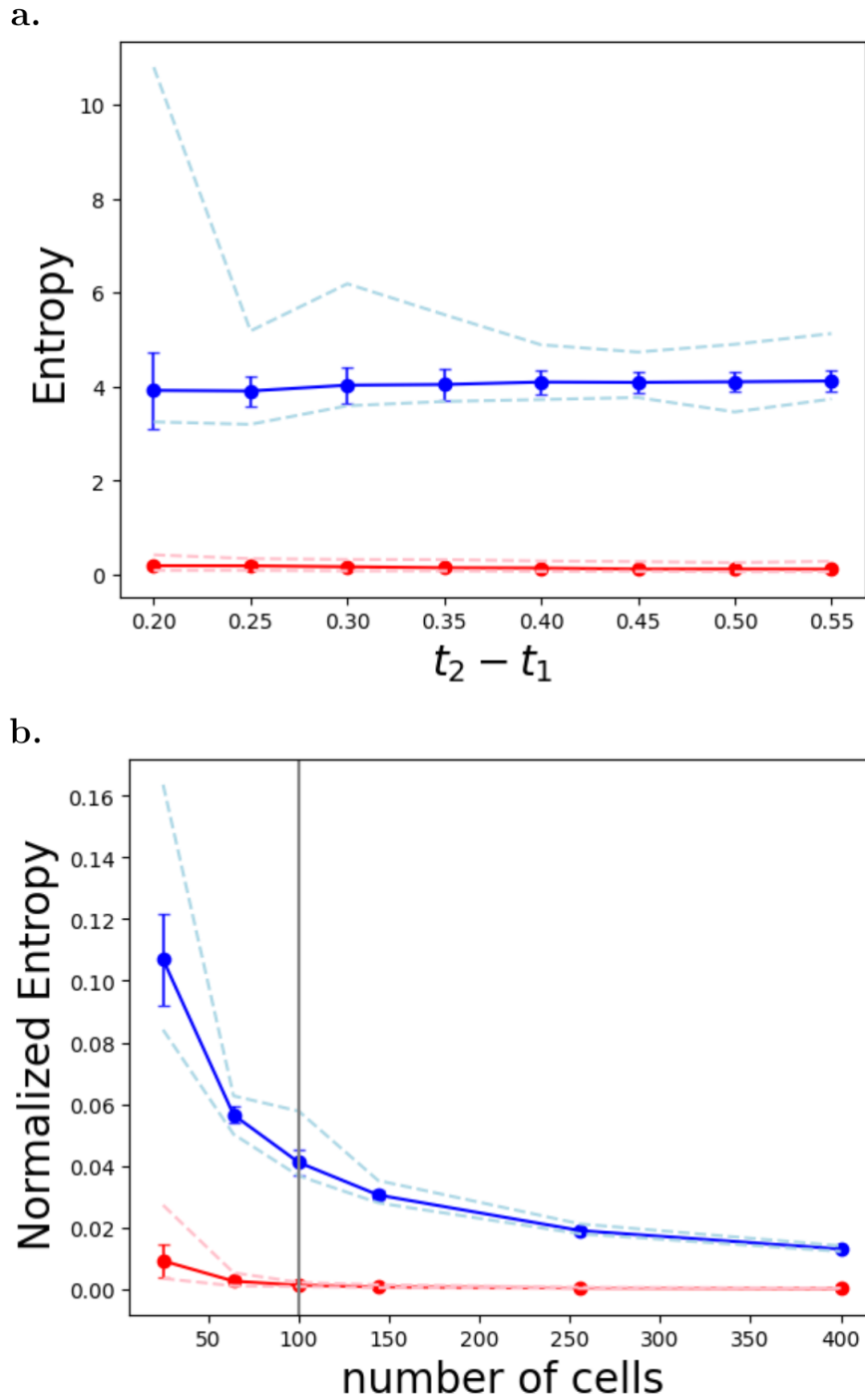


Figure 2.3. Entropies between reference processes for the toggle switch GRN, in the transient state. Mean and standard deviation (straight lines) of (a.) $H_{D,B}$ (blue) and H_B (red) as functions of the time interval $t_2 - t_1$, and (b.) $H_{D,B}/N$ (blue) and H_B/N (red) as functions of the number of cells N . Additionally, minimum and maximum values over 100 iterations are represented (dashed lines). The thick vertical grey line in Panel B is the default value $N = 100$ cells used in other figures. Other parameters are given in Table 2.1.

final distribution becomes independent of the initial one (later time measurements).

Figure 2.3b illustrates the influence of the number of cells N considered when measuring both initial and final distributions in transient state. To that aim, the entropy has been normalized by the number of cells to avoid a bias due to the size of the collections in the computation of the entropy: the computation of the entropy (2.8) explicitly depends on the number of cells N . First, we observed that H_B/N does not depend on the number of cells. Second, $H_{D,B}/N$ decreases as the number of cells increases. We observed (not shown) that the Diffusive process tends to couple a small sample of cells from the initial collection to all other cells in the final collection.

Both numerical investigations on the dependence of $H_{D,B}$ and H_B on the length of the time interval $t_2 - t_1$ and the number of cells N highlighted a poor ability of the Diffusive model to reconstruct cell trajectories when the underlying model is the Bursty model of gene dynamics. These results are based on the use of the entropy to measure the differences between solutions of the Schrödinger problem. We then questioned the relevance of the entropy criterion and the parameter values of the Bursty model in reaching this conclusion.

2.2.2.2 Unspecificity of the entropy indicator and the distribution parameters

In Figure 2.4, we compared the performance of different indicators when evaluating the difference between the Diffusive and the Bursty process. Figure 2.4 shows mean and standard deviation of normalized indicators, computed as $(Indic - ControlIndic)/ControlIndic$, where $Indic$ is either $H_{D,B}$, $D_{TV}(D_{sch}, B_{sch})$, $JSD(D_{sch}, B_{sch})$ or the Frobenius norm $\|D_{sch} - B_{sch}\|_F$ (see Section 2.4.5) and $ControlIndic$ its associated control measure (respectively, H_B , $D_{TV}(B_{ref}, B_{sch})$, $JSD(B_{ref}, B_{sch})$ and $\|B_{ref} - B_{sch}\|_F$). We observe that for all indicators, the curves are all strictly positive and increasing. So, despite quantitative differences, all indicators yielded the same qualitative conclusion: the Bursty process provides a better trajectory inference than the Diffusive process.

To challenge the performance of the B_{ref} process, we generated the two distributions with a given GRN and used a different GRN for the reference process (mostly the same structure, but different interaction parameter values), that is a different Bursty model. The use of simulated data provides full knowledge of the structure of the GRN and of its parameter values, however when dealing with experimental data the GRN is not known and can be at best inferred from the data, without any way to validate per se either the structure of the GRN or the parameter values. Hence, using a GRN that does not correspond to the data used to generate the initial and final distributions challenges the method and allows to test the robustness of the approach. In practice, we computed the initial distribution μ and the final distribution ν using the default parameter values of the toggle switch GRN, but we solved the Schrödinger problem with a different B_{ref} process (different gene interaction matrix).

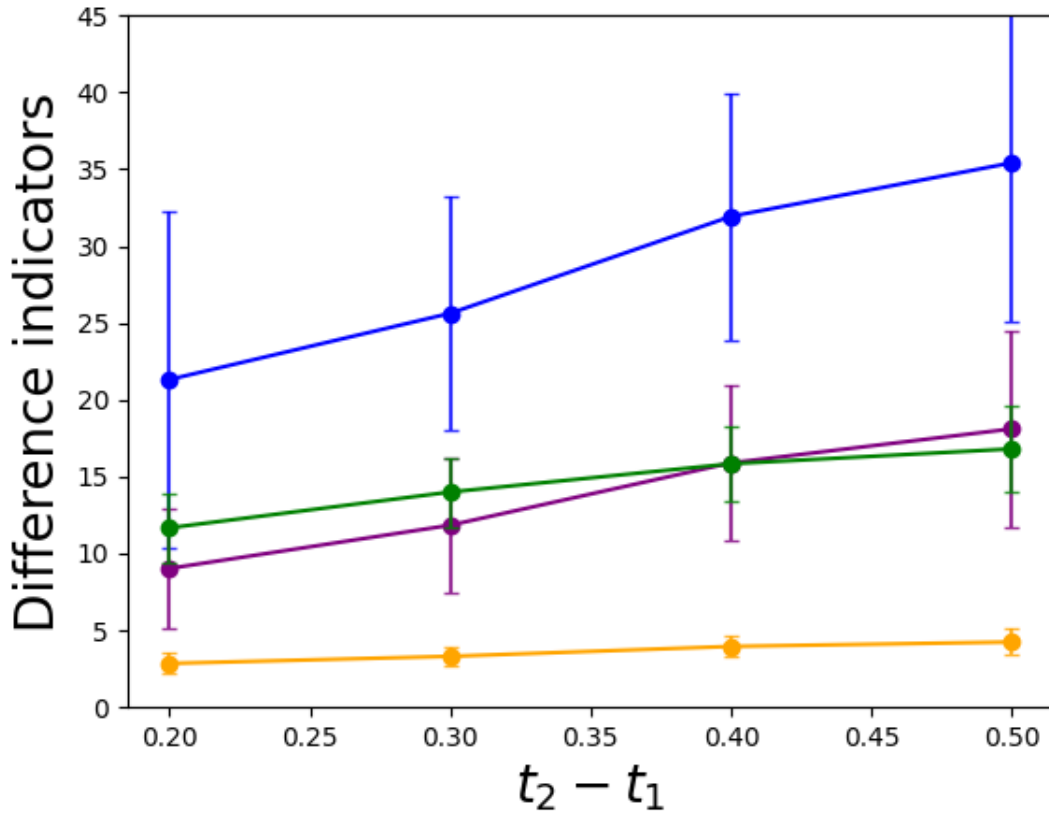


Figure 2.4. Difference indicators between reference processes as a function of the time interval for the toggle switch GRN. Mean and standard deviation of difference indicators comparing D_{sch} and B_{sch} normalized by the difference indicators comparing B_{ref} and B_{sch} respectively: Entropy (blue), total variation distance (yellow), Jensen–Shannon Divergence (purple), and Frobenius norm (green). For standard deviations < 0.1 , error bars are not visible. Parameters are detailed in Table 2.1.

Figure 2.5a highlights that for all the two-gene-based Bursty models (listed in Figure 2.5c), the curves are very close. The $H_{D,B}$ curves are all higher than the H_B curves, indicating that any Bursty process, even one with parameter values different from the ones corresponding to the data, provides a better approximation than the Diffusive process. In Figure 2.5b, one observes that even though H_B curves are nearly indistinguishable at early times they become separated as $t_2 - t_1$ increases. We observed differences when parameter values of the GRN associated with self-activation were reduced (purple and yellow curves and GRN).

Overall, the Bursty model provides a framework to cell trajectory inference that performs better than standard diffusion.

We studied numerically the sensitivity of our method to hyperparameters (number of simulations n , standard deviation ε in (2.6), precision and truncation thresholds used to prevent zero values in the Sinkhorn algorithm) and to parameter values of the GRN, specifically focusing on degradation rates and basal activity levels (Supplementary File). No significant impact was observed on the entropy $H_{D,B}$, pointing towards a weak sensitivity of the results on parameterization.

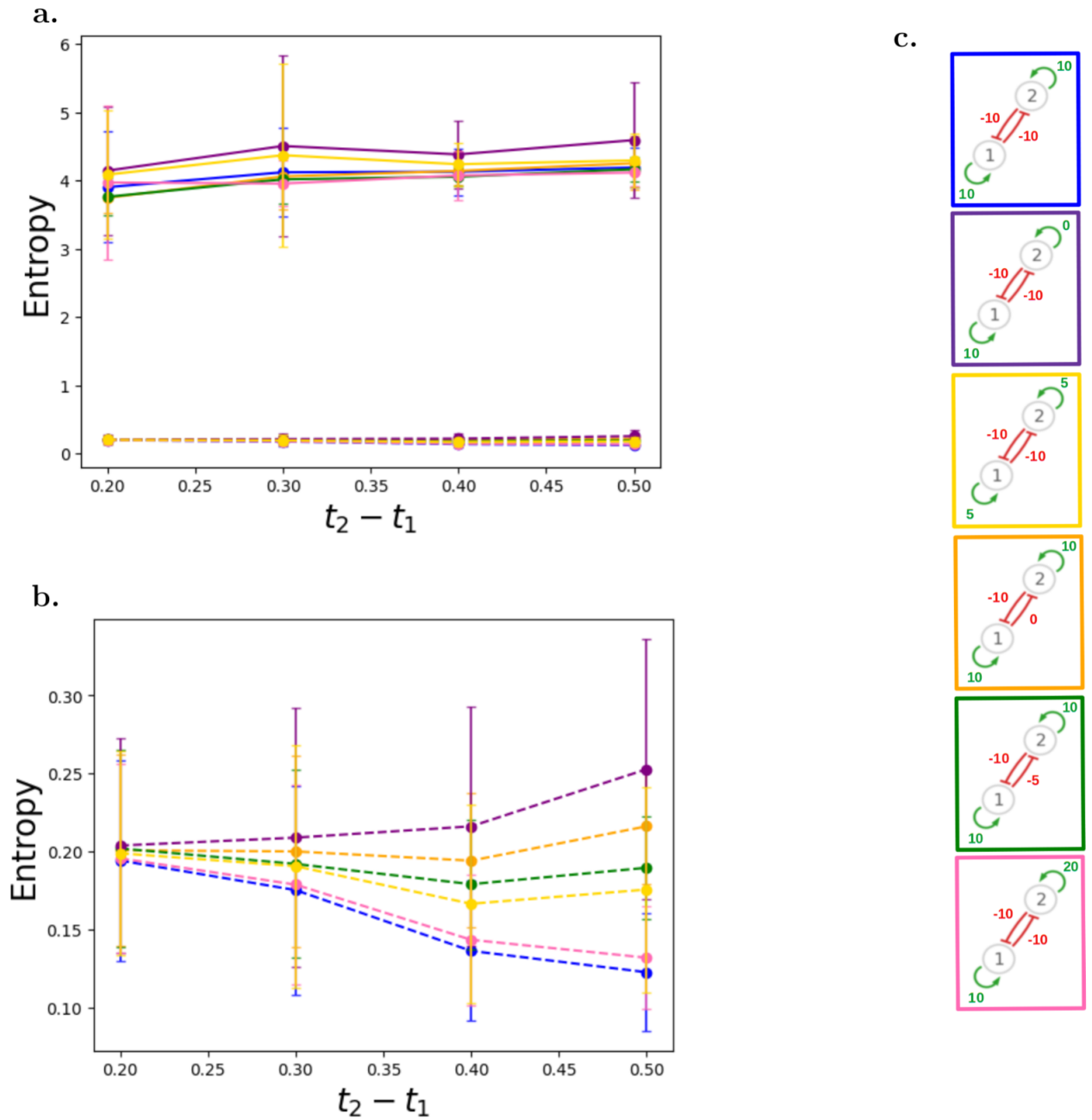


Figure 2.5. Entropy $H_{D,B}$ as a function of the time difference $t_2 - t_1$ for different Bursty models, in transient state. **a.** Entropies H_B (dashed line) and $H_{D,B}$ (straight line) means and standard deviations. Each color represents a set of interaction parameters of B_{ref} used to compute $H_{D,B}$ (illustrated in Panel **c**). **b.** Zoom in on H_B values in Panel **a** for $t_2 - t_1 \in [0.2, 0.5]$. **c.** Two-gene GRN used to compute B_{ref} . Each scheme depicts interactions between gene 1 and gene 2, with green arrows representing activation ($\theta > 0$) and red arrows representing inhibition ($\theta < 0$). The blue-squared GRN is the toggle switch used in previous figures. Parameters values are detailed in Table 2.1.

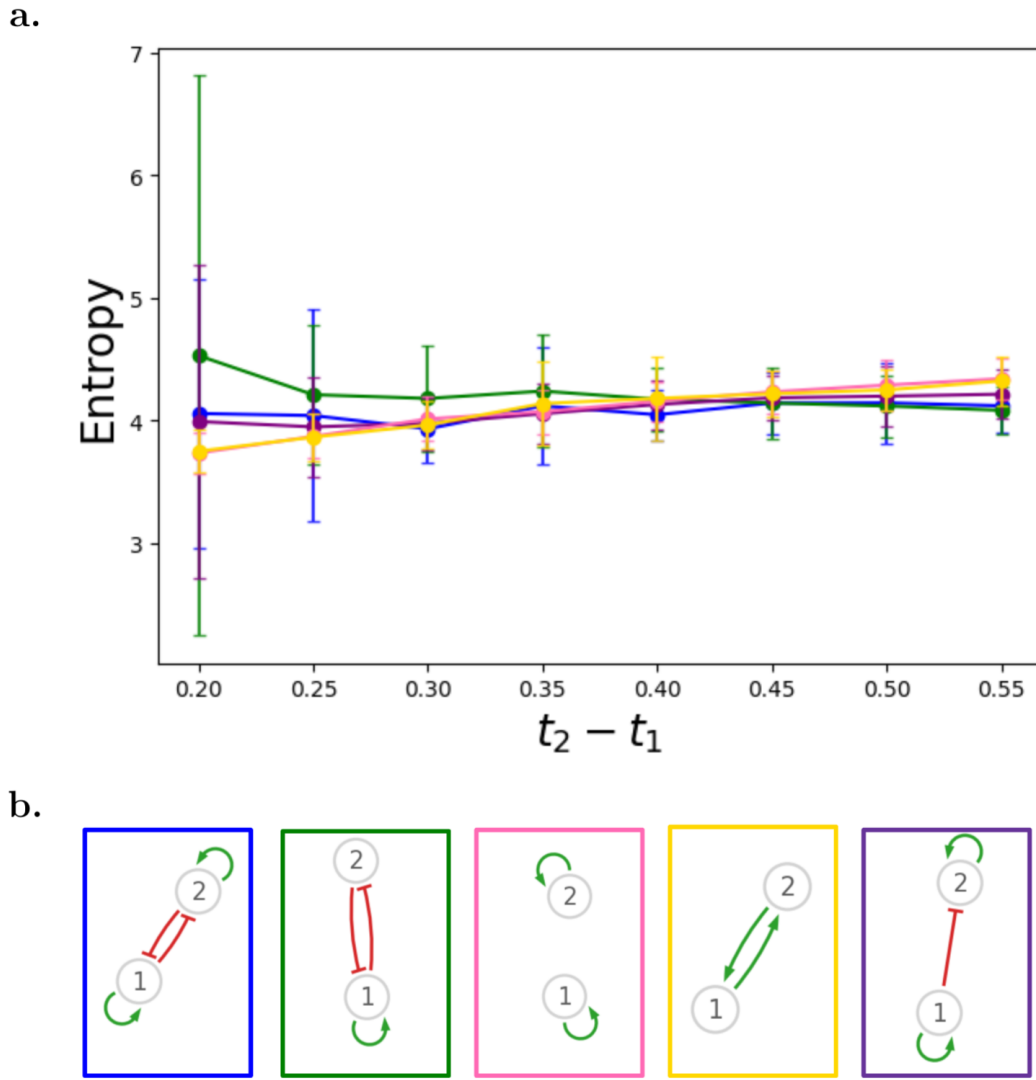


Figure 2.6. Entropy $H_{D,B}$ as a function of the time difference $t_2 - t_1$ for various two-gene GRNs. **a.** Entropy $H_{D,B}$ in transient state, mean and standard deviation. Each color represents a GRN introduced in Panel **b.**. Parameters are detailed in Table 2.1. **b.** Schematic representations of two-gene GRNs used in Panel **a.** Green arrows represent activations ($\theta > 0$) and red arrows represent inhibitions ($\theta < 0$).

2.2.3 Analysis for other GRNs

As illustrated in Figure 2.11, the toggle switch model of gene dynamics exhibits what can be considered as specific temporal dynamics. We subsequently challenged the Bursty process on other GRNs, first with various two-gene GRNs and then on three-gene GRNs. Results discussed hereafter and illustrated in Figures 2.6 and 2.7 have been obtained using various GRNs to compute both the distributions (μ and ν) and the reference coupling. This contrasts with the results shown in Figure 2.5, where only the reference coupling was modified by the GRNs, not the data.

Figure 2.6 shows the entropy $H_{D,B}$ for different two-gene GRNs. For each GRN, the distributions and the Bursty process were computed from the corresponding Bursty model. The blue curve represents $H_{D,B}$ values computed for the toggle switch. We observed that, whatever the two-gene GRN, dynamics of $H_{D,B}$ were similar, quantitatively and qualitatively: it was mostly

constant as the time difference $t_2 - t_1$ increased, and the variability was higher for small time differences.

Figure 2.7 shows the entropy $H_{D,B}$ for two 3-gene GRNs. In the first one, gene 1 activates gene 2 which activates gene 3 (Figure 2.7a); in the second one, the so-called repressilator model, gene 1 inhibits gene 2, gene 2 inhibits gene 3, and gene 3 inhibits gene 1 (Figure 2.7b). Parameter values are the same than for the toggle switch (Table 2.1).

Figure 2.7a shows a similar range of variations for $H_{D,B}$ and H_B as observed for the two-gene GRNs. Variations about the mean were, however, more important than for the two-gene GRNs, as highlighted by error bars and minimum and maximum values (dashed lines). The $H_{D,B}$ curve remains well above the H_B curve, similarly to what has been obtained for two-gene GRNs. The same results held for the Repressilator model (Figure 2.7b), a standard three-gene GRN. Whatever the time difference, the mean of $H_{D,B}$ was always higher than H_B . For both GRNs, when $t_2 - t_1$ was small, the standard deviation of $H_{D,B}$ was larger and the minimum and maximum values were widely separated. The two three-gene GRNs considered generated more differences between the Diffusive and the Bursty processes than the toggle switch model or any 2-gene GRN, as well as more variability.

In conclusion, for any two-gene or three-gene GRN, whatever the gene-gene interactions considered, the Bursty process always performed better than the standard Diffusive process in correctly transporting the distribution at time t_1 into the distribution at time t_2 .

2.2.4 Application to experimental data from Semrau et al. (2017)

While previous sections have been devoted to validating the Bursty-based inference method on small GRN and simulated data, here we apply the method on an experimental dataset. We use a time-stamped in vitro dataset from Semrau et al. (2017) made of mRNA counts obtained from scRNA-seq on mouse embryonic stem cells. A GRN in the form of a bursty model has been inferred from these data by Ventre et al. (2023), therefore we use this GRN and the parameter values obtained during the inference procedure to compute our reference process B_{ref} .

Since all previous investigations were performed on protein data, we first check whether or not the Bursty model-based inference method would also perform well on simulated mRNA data. Similarly to what we did in Figure 2.3a., considering the toggle-switch GRN, we investigated the impact of time difference $t_2 - t_1$ on the entropy $H_{D,B}$ but this time using only simulated mRNA data. Figure 2.8 shows that the $H_{D,B}$ curve is higher than the H_B curve. Contrary to what is observed with protein level distributions, the $H_{D,B}$ curve slightly decreases as $t_2 - t_1$ increases, yet remaining far above the control curve. mRNAs are produced and degraded more rapidly than proteins, accelerating the approach to the stationary state. The same conclusions then hold, whether mRNA or protein data are used to infer cell trajectories.

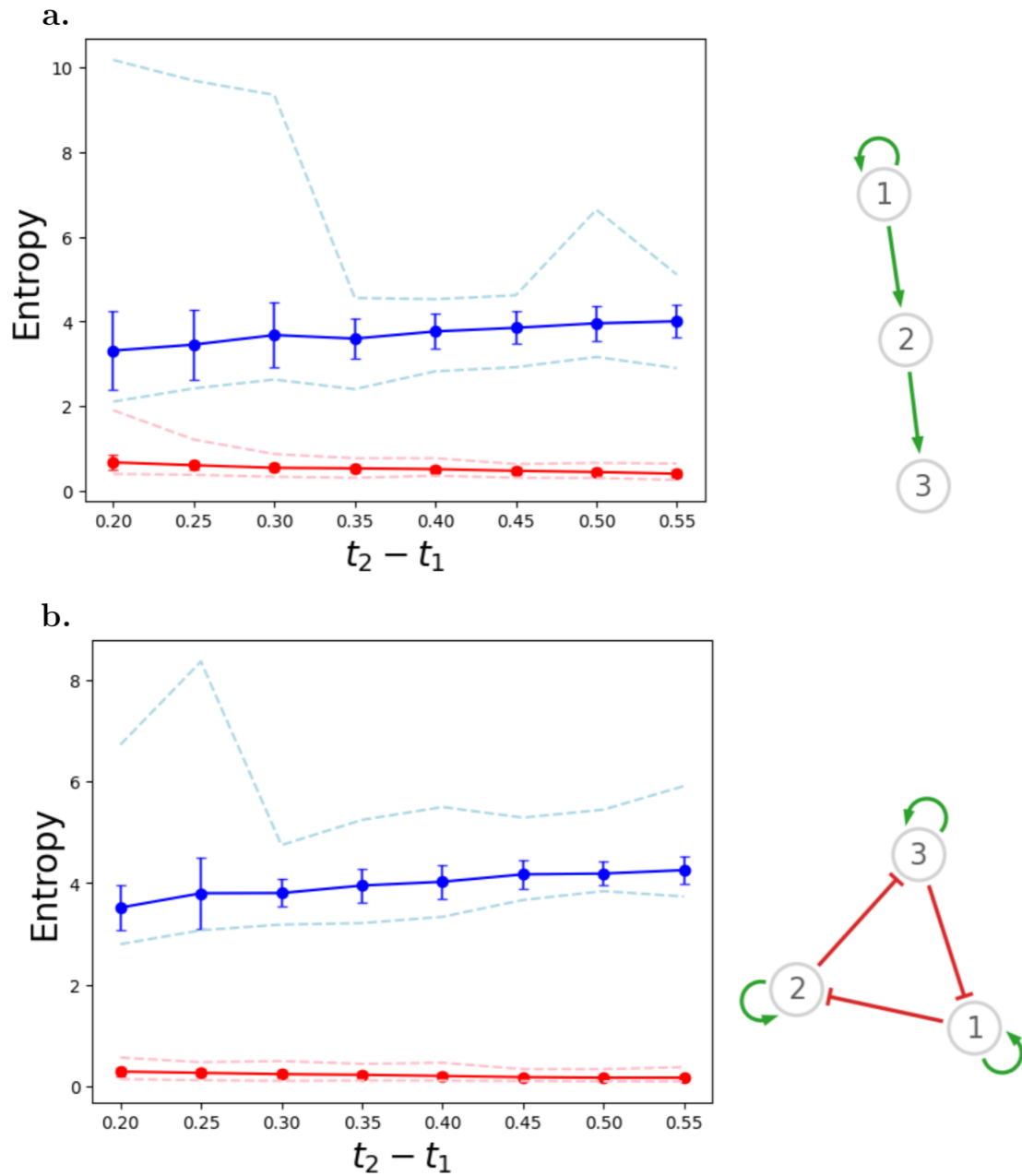


Figure 2.7. Entropy evolution as a function of the time difference $t_2 - t_1$ in the transient state for three-gene GRNs. **a.** $H_{D,B}$ mean and standard deviations (blue line with error bars) as well as min and max values (dash blue lines) and H_B mean and standard deviation (red line with error bars) and min and max values (dash red lines), and the associated three-gene GRN. The schematic representation of the GRN highlights positive interactions (activation) between genes. **b.** $H_{D,B}$ mean and standard deviations (blue line with error bars) as well as min and max values (dash blue lines) and H_B mean and standard deviation (red line with error bars) and min and max values (dash red lines), and the associated three-gene GRN (repressilator). The schematic representation of the GRN shows interactions between genes, with activations in green and inhibitions in red. Parameters are detailed in Table 2.1.

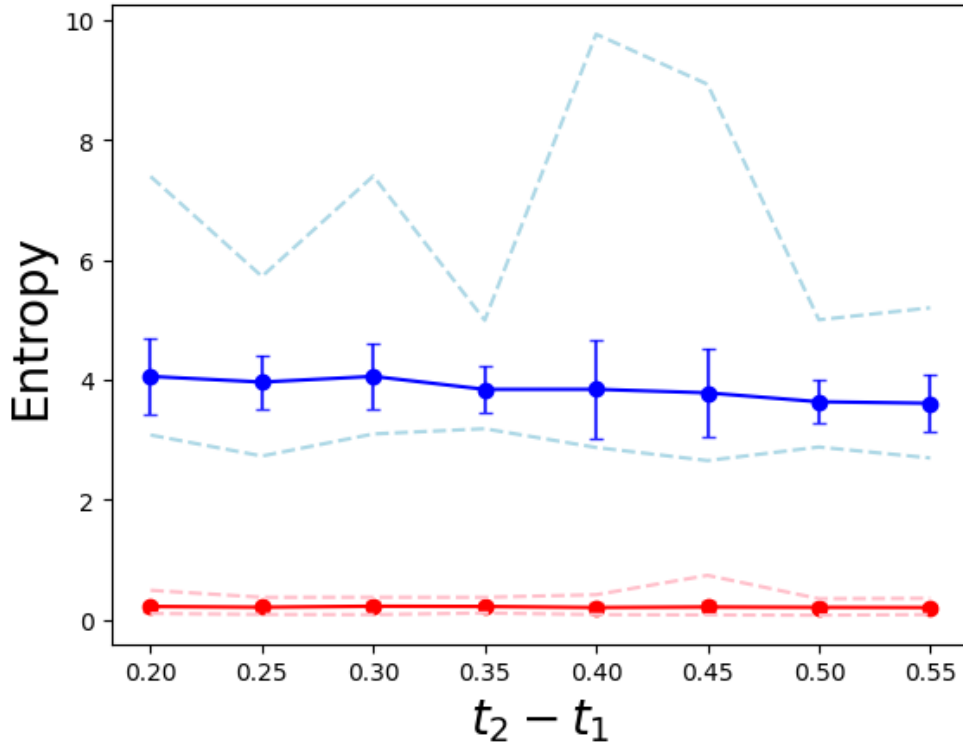


Figure 2.8. Entropies between reference processes for the toggle switch GRN for mRNA data. Mean and standard deviation (straight lines) of $H_{D,B}$ (blue) and H_B (red) are displayed as functions of the time interval $t_2 - t_1$. Additionally, minimum and maximum values over 100 iterations are represented (dashed lines). Parameters are given in Table 2.1.

We focus now on experimental mRNA data from Semrau et al. (2017). When applying the Bursty model-based inference method to experimental data, no ground truth is available therefore, in a general context, standard or recently proposed methods—such as cluster-to-cluster fate probabilities (Lange et al., 2022), ancestral/source compositions (Schiebinger et al., 2019)—may allow a posteriori validation by focusing on checkable quantities obtained from the identified coupling, provided that the size of the sample is large enough. Here, in order to apply our Bursty model-based inference method, we chose to focus on its predictive power. We first computed an mRNA-based version of the process B_{sch} between time points $t_1 = 0$ and $t_3 = 72\text{h}$, and estimated the distribution $\rho_{\text{est},36}$ at $t_2 = 36\text{h}$ (see Section 2.4.6 and Supplementary File, Section 2.10.5). Next, we computed the Earth Mover’s Distance (EMD) between the experimental distribution at 36h and, on one side, $\rho_{\text{est},36}$, denoted by EMD_B , and on the other side, the distribution obtained using regularized OT, $\rho_{D,36}$, denoted by EMD_D .

Figure 2.9 shows $\Delta_{EMD} = EMD_D - EMD_B$ for each of the 41 genes in the dataset. It allows a measure of the performance of the Bursty model-based inference method compared to the Diffusion-based inference method. Green bar-plots correspond to a positive Δ_{EMD} , i.e. EMD_B is smaller than EMD_D , and red bar-plots correspond to negative Δ_{EMD} , i.e. when EMD_B is larger than EMD_D . Heights of bar-plots quantify the contributions of Bursty model-based versus Diffusion-based inference methods.

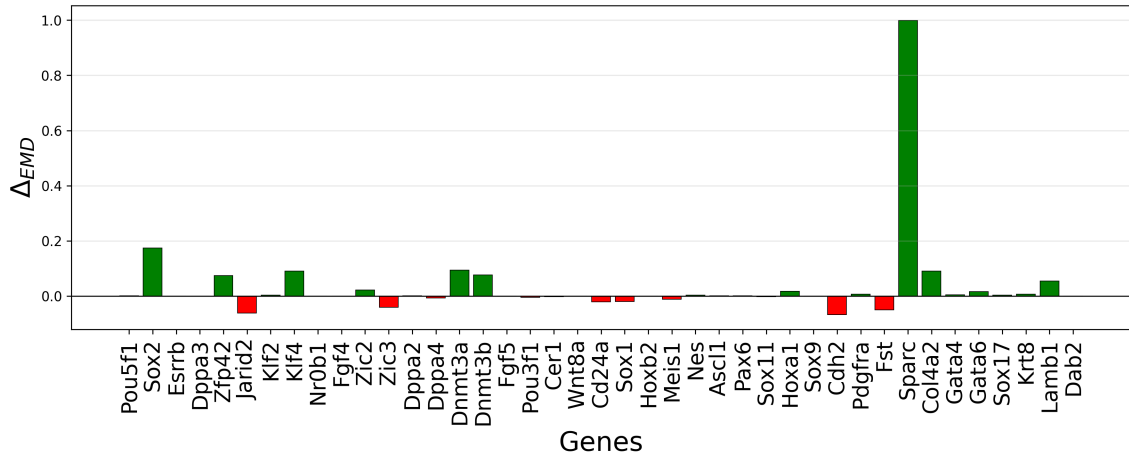


Figure 2.9. Gene-by-gene differences between EMD_B and EMD_D at $t_2 = 36h$, using data at $t_1 = 0h$ and $t_3 = 72h$. Bar plots are green and positive when the Bursty model-based inference method performs better, they are red and negative when the Diffusion-based inference method performs better.

For 11 genes, $\Delta_{EMD} = 0$. This means either that both EMD_B and EMD_D equal 0 (genes *Fgf5*, *Cer1*, *Ascl1* and *Sox9*), or that they are equal (*Esrrb*, *Dppa3*, *Nr0b1*, *Fgf4*, *Wnt8a*, *Pax6* and *Gata4*). In the first case, both methods are able to correctly predict the gene expression level, which is close to zero for the considered genes (no expression). In the second case, both methods similarly predict the gene expression level. Noticeably, gene distributions that do not evolve much between $t_1 = 0h$ and $t_3 = 72h$ display a small EMD.

Out of the remaining 30 genes, for 22 genes the Bursty model-based inference method performs better. Values of EMD_B and EMD_D for each gene are displayed in Supplementary File (Supp. Section 2.10.5). Noticeably, the large bar-plot of gene *Sparc* indicates that the distance between the two EMD is negative and rather large but, in that case, both EMD_B and EMD_D are also large (both methods have difficulties correctly predicting this gene dynamics). For genes *Col4a2* and *Lamb1*, differences between EMD_D and EMD_B are also negative, yet smaller than for *Sparc*, but in that case both EMD are small (both methods perform well in predicting these gene dynamics, the Bursty model-based inference method performing better).

One may note that we only focus here on an illustrative example where we predict the distribution at 36h by using distributions at 0h and 72h. However, the results are representative of other predictions that could be tested (not shown). Sometimes both Bursty model-based and Diffusion-based inference methods compare well (in particular in the early times, when only few genes drive the dynamics and their levels of expression do change slowly over time), nevertheless the above results show that the Bursty model-based inference method can be applied on experimental data and performs better than the Diffusion-based method.

2.3 Discussion

We explored the possibility that the Bursty model of gene dynamics would represent an improved alternative to the standard diffusion model used to infer trajectories in OT theory. Based on distributions of protein counts, the Bursty model proved indeed quite efficient in capturing cells trajectories more accurately than the Diffusive model (Figure 2.1). This would be expected if we considered cells moving around in a Waddington's landscape (Waddington, 1957) whose shape is constrained by the GRN structure (Huang, 2010; Ventre, 2021). In this case, diffusion could not easily represent the variables acceleration and deceleration due to valleys and hills. Noticeably, we assumed for the sake of simplicity the same number of cells in both initial and final distributions, yet our entire analysis could be extended to the case with different cell numbers by normalizing the distributions.

When considering a toy model made of only one gene, we concluded that using the exact time-dependent distribution of the process was the best alternative to diffusion. Although expected, this result indicates that it is possible to very efficiently infer cell trajectories in a transcriptomic space. However, the exact distribution is only explicit for one gene. Extending the reference coupling RB_{ref} to GRN with multiple genes seems a very challenging task, probably impossible to do for complex gene interactions. An approximation of this distribution could be considered, yet the approximation might rapidly become coarse.

For a one-gene GRN, the computation of RB_{ref} and D_{ref} couplings is almost instantaneous on a personal computer (64 GB RAM and 13th Gen Intel(R) Core(TM) i7-13700H), whereas the computation of B_{ref} takes approximately 7s (value averaged over 100 simulations). Increasing the number of genes does not impact the computational time for D_{ref} , but it increases for B_{ref} . The computational time is proportional to the number of genes in the GRN with almost no time-variability from one simulation to another, for two-gene up to six-gene GRNs where genes self-activate and inhibit the other genes (similar to the toggle switch model). For two genes, the computational time is 12s, for 3 genes, it is 13s, for 4 genes, 14s, for 5 genes 15.5s and for 6 genes 17s per simulation of the B_{ref} coupling, on a computational server (High-Throughput Computing platform with 1 CPU and 2 GiB memories). For most applications, such computational time would be acceptable. If needed, computations of B_{ref} coupling can be parallelized, for instance when the number of simulations, the number of cells, or the number of genes is large. For the application to experimental data in Section 2.4.6, the computation of the B_{ref} coupling took 23min using parallelization and 168min without parallelization.

The full two-state model provides both mRNA and protein counts. Since experimentally available scRNA-seq datasets provide mRNA counts, it would have been appropriate to focus only on mRNA counts. However, it is not possible to reduce the Bursty model to an mRNA-only model and to preserve mathematical results, due to the fact that mRNA are typically much less sta-

ble than proteins and thereby would not preserve the Markovian property of the GRN. We can however reduce the Bursty model to a protein-only model that is well defined, along with quasi-steady-state distributions of mRNA counts conditionally on proteins (Gaillard & Herbach, 2025; Herbach et al., 2017). Moreover, Baradat and Ventre (2024) showed that for such a model it is still possible to define and solve a Schrödinger problem. We therefore decided to focus on distributions of protein counts when bench-marking on simulated datasets.

In Section 2.2.4, we applied our Busty model-based inference method on experimental data consisting of mRNA counts. We showed first that despite the above arguments in favor of preferentially using protein data, the Bursty model can also be applied to simulated mRNA data and provides in this case similar performances. To use our approach on an experimental dataset, it is first required to get a GRN, associated with the data, that is used to compute the reference coupling. This motivated the use of the dataset from (Semrau et al., 2017), since Ventre et al. (2023) had already inferred a GRN from this dataset using CARDAMOM. The inference and calibration of the mechanistic model with CARDAMOM are based on a two-step procedure, with bursting parameters being identified in the first step (Ventre, 2021; Ventre et al., 2023). Noticeably, other algorithms have been proposed to infer the bursting kinetics from single-cell RNA sequencing data (Huang et al., 2026; Kim & Marioni, 2013; Luo et al., 2023a; 2023b), the more recent ones benefiting from AI-based methods to infer kinetics and the method proposed in Huang et al. (2026) relying on the use of the Telegraph model, as does our method. Importantly, even if the underlying GRN structure is not perfectly known, which is expected from using inference methods to obtain the GRN, we established in Section 2.2.3 (see for instance Figure 2.5 and Supplementary File) that the Bursty model-based inference method is not very sensitive to parameter values of the GRN and correctly catches the trajectories even though the GRN used to compute the reference coupling is not correct.

When trying to predict an intermediate observation with Semrau et al. (2017), the interpolation made with the Bursty model-based inference method performed better than the one made with Diffusion-based inference method. It is however worth noticing that the Diffusion-based inference method works well when observations are made on a short time period and when gene activity is limited. When observations are made on longer periods of time and genes exhibit important activity, the Bursty model-based inference method performs systematically better. Altogether the results presented in this article thus pave the way to the extension of GRN inference algorithms to methods combining GRN and trajectory inference in an integrated framework.

2.4 Methods

2.4.1 Gene dynamics

We model gene expression using the framework introduced by Herbach et al. (2017), and later simplified in the bursty regime (Gaillard & Herbach, 2025; Herbach, 2021; Ventre, 2021). In this model, a gene is characterized by its amount of mRNA and its level of proteins: mRNA is synthesized through stochastic transcriptional bursts, and degraded at rate d_0 ; proteins are synthesized from mRNA at rate s_1 and degraded at rate d_1 . The Bursty model is a piecewise-deterministic Markov process, with deterministic part given by

$$\begin{cases} \dot{M}(t) = -d_0 M(t), \\ \dot{P}(t) = s_1 M(t) - d_1 P(t), \end{cases} \quad (2.1)$$

where $M(t)$ and $P(t)$ are continuous approximations of the quantity of mRNA molecules and the amount of protein, respectively, at time t , and a stochastic part consisting of mRNA bursts occurring at random times with rate k_{on} and random sizes that follow an exponential distribution (Herbach, 2021; Ventre, 2021).

Within a cell, genes interact with each other via a GRN. Hence, a gene can be activated or inhibited by itself but also by other genes in the GRN. A burst is assumed to occur for gene i at a rate $k_{on,i}(P_1, \dots, P_G)$, where G is the number of genes in the GRN and $(P_1, \dots, P_G)$ is the vector of protein levels of all genes in the GRN. The burst rate of gene i accounts for a basal activity parameter β_i and the influence of other genes in the GRN, expressed via a gene-gene interaction matrix $\theta = \{\theta_{ji}\}_{j,i \in \{1, \dots, G\}}$ (Herbach, 2021; 2023), so that

$$k_{on,i}(P_1, \dots, P_G) = \frac{k_{0,i} + k_{1,i} \exp(\beta_i + \sum_{j=1}^G \theta_{ji} P_j)}{1 + \exp(\beta_i + \sum_{j=1}^G \theta_{ji} P_j)}, \quad (2.2)$$

where $k_{0,i}$ and $k_{1,i}$ correspond respectively to the minimum and maximum burst frequencies of gene i , and θ_{ji} is the intensity of gene j action on gene i (Herbach et al., 2017). Parameters θ_{ji} can be positive or negative, depending on whether gene j activates or inhibits gene i (Figure 2.10).

Example of GRN: the Toggle Switch. We consider a standard GRN made of two genes, the so-called “toggle switch” model. The two genes self-activate and inhibit each other, i.e.

$$\theta_{11} = \theta_{22} \geq 0 \quad \text{and} \quad \theta_{12} = \theta_{21} < 0.$$

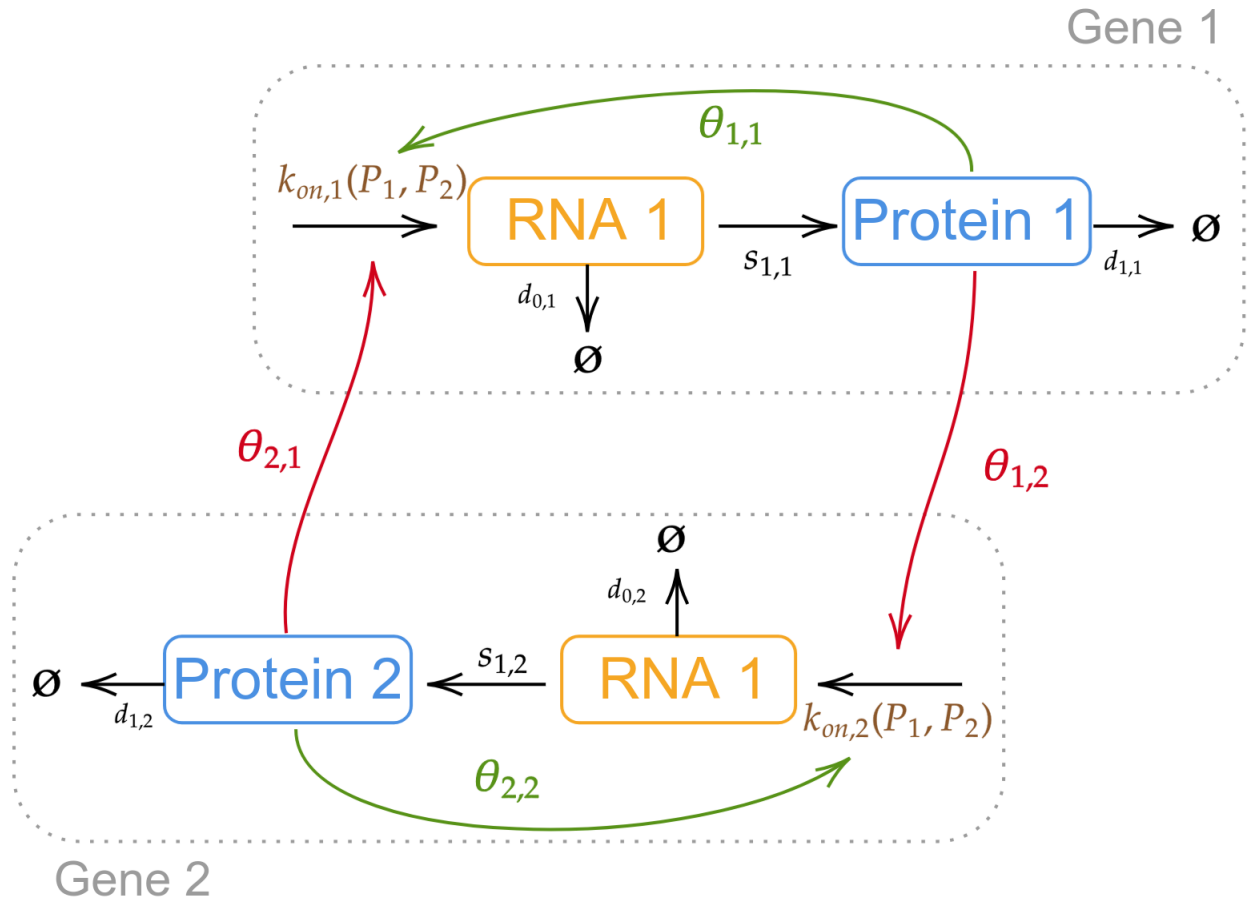


Figure 2.10. Schematic representation of a two-gene GRN based on the Bursty model of gene dynamics. For each gene $i = 1, 2$, mRNA synthesis is activated with a rate $k_{on,i}(P_1, P_2)$, defined in (2.2), and mRNA and protein dynamics follow system (2.1). Parameters of the GRN are described in the text. In the case of the toggle switch model, the GRN consists of 2 genes that self-activate ($\theta_{1,1} > 0$ and $\theta_{2,2} > 0$, green arrows) and inhibit each other ($\theta_{2,1} < 0$ and $\theta_{1,2} < 0$, red arrows).

This GRN is illustrated in Figure 2.10, while Figure 2.11 illustrates the temporal evolutions of mRNA and proteins for a toggle switch GRN where gene dynamics obey the Bursty model, computed with parameter values given in Table 2.1. A deterministic toggle switch model would favor one gene over the other, depending on initial conditions of the system. In this stochastic version of the toggle switch, one observes that one gene activity always prevails, yet the dominant gene switches stochastically.

More generally, by modifying the values of the interaction parameters θ_{ji} in model (2.1)-(2.2), any two-gene GRN can be implemented (see, for example, Figure 2.6). Extensions to three-gene GRNs can be straightforwardly implemented (see, for example, Figure 2.7).

Simulations of protein dynamics. Computations of the temporal evolution of protein and mRNA counts were performed using the Python package Harissa (Herbach, 2023), which implements an exact simulation procedure for the Bursty model (Gaillard & Herbach, 2025) (Figure 2.11). This allows to simulate one GRN, hence simulating N times the evolution of a GRN results in N

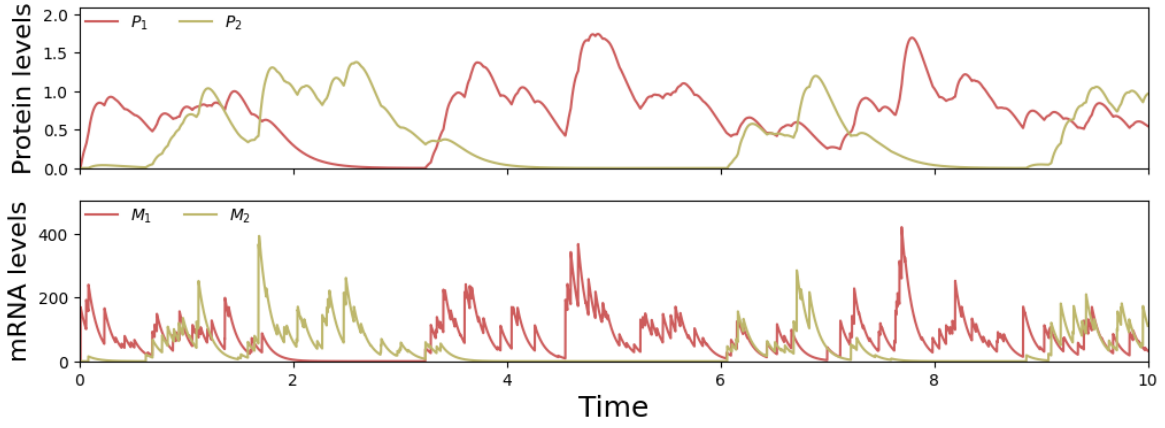


Figure 2.11. Temporal evolution of protein and mRNA levels for the toggle switch model. Dynamics of one gene are illustrated in yellow, the other in red. Protein levels are displayed on the upper panel, while mRNA dynamics are displayed in the lower panel.

independent cell trajectories (Figure 2.12).

In order to mimic biological experiments aimed at measuring a GRN activity at given observation times, we simulate N times the GRN with initial conditions IC_M and IC_P (see Table 2.1), select a simulation time t , and extract all mRNA and proteins counts for all genes.

Throughout this work, we focus on distributions of protein counts only, yet the case of mRNA distributions is discussed in the last section. We denote by μ the distribution of G protein counts in a population of N cells at time $t = t_1$, and ν the distribution of G protein counts in the same population at time $t = t_2$, with $t_2 > t_1$. They are given by:

$$\mu = \frac{1}{N} \sum_{i=1}^N \delta_{x_i} \quad \text{and} \quad \nu = \frac{1}{N} \sum_{j=1}^N \delta_{y_j},$$

where δ is the Dirac distribution, $x_i = (P_g^i(t_1))_{g=1,\dots,G}$ is the level of gene g 's protein in cell x_i at t_1 and $y_j = (P_g^j(t_2))_{g=1,\dots,G}$ is the level of gene g 's protein in cell y_j at t_2 . For the sake of simplicity, when using simulated data cells have the same index in distributions μ and ν (i.e., the i -th component of μ is associated to the same cell than the i -th component in ν). Additionally, the two distributions are jointly normalized (between 0 and 1). Default parameter values are given in Table 2.1.

2.4.2 Reduced model with analytical solution

In addition to the model introduced in the previous section, for which couplings will be estimated from simulations, we consider a simplified model (*reduced Bursty model*) that will provide closed-form couplings. This simplification is achieved in two steps. First, a protein-only model is derived from a time scale separation of mRNA and proteins corresponding to $d_0 \gg d_1$ (i.e., proteins assumed to be more stable than mRNAs). This model is defined by (Friedman

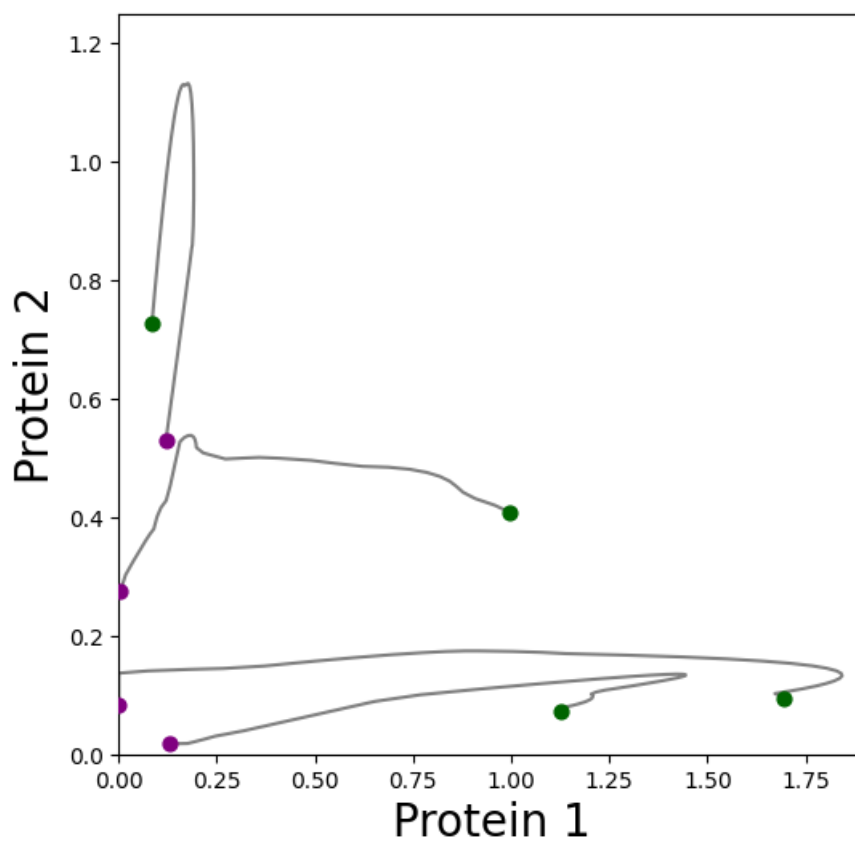


Figure 2.12. Cell trajectories of four cells driven by a toggle switch. Initial (time $t_1 = 0.1h$) location of cells is displayed as purple dots, while their final (time $t_2 = 0.5h$) location is represented by green dots. Each grey line is the temporal trajectory of each cell. Parameters are given in Table 2.1.

et al., 2006; Herbach, 2023):

$$\begin{cases} P(t) \xrightarrow{k_{on}} P(t) + \mathcal{E}(1/\eta) \\ \dot{P}(t) = -d_1 P(t) \end{cases} \quad (2.3)$$

where $\mathcal{E}$ is the exponential distribution characterizing the random size of each burst (with mean value η), k_{on} is the burst rate and d_1 the degradation rate of proteins, as before. Second, we consider a particular 1-dimensional case of model (2.3) with a constant k_{on} : this setting can be interpreted either as a single gene with no feedback, or part of a GRN within a time range sufficiently short so that $k_{on}(P_1, \dots, P_G)$ does not vary significantly during this period.

This model turns out to be analytically tractable (see Supplementary File, Section 2.10.1). The time-dependent distribution of $P(t)$ knowing $P(0) = P_0$ is given by

$$\mathcal{L}(P(t)|P(0) = P_0) = e^{-k_{on}t} \delta_{e^{-d_1 t} P_0} + (1 - e^{-k_{on}t}) \nu_t \quad (2.4)$$

where $\delta_{e^{-d_1 t} P_0}$ is the Dirac measure localized at $e^{-d_1 t} P_0$, and ν_t has a density function $g_t(x|P_0)$ given by

$$g_t(x|P_0) = f_t(x - e^{-d_1 t} P_0), \quad (2.5)$$

where

$$f_t(x) = \frac{1}{\eta} \frac{(e^{d_1 t} - 1) k_{on}}{(e^{k_{on} t} - 1) d_1} \exp\left(-\frac{x}{\eta} e^{d_1 t}\right) \Phi\left(\frac{k_{on}}{d_1} + 1; 2; (e^{d_1 t} - 1) \frac{x}{\eta}\right) \mathbb{1}_{\{x \geq 0\}},$$

and Φ denotes the confluent hypergeometric (Kummer's) function of the first kind.

For numerical applications, we regularize the singular part in (2.4) by replacing it with a normal distribution with small standard deviation $\varepsilon \ll 1$,

$$\mathcal{L}(P(t)|P(0) = P_0) \approx e^{-k_{on}t} \mathcal{N}(e^{-d_1 t} P_0, \varepsilon^2) + (1 - e^{-k_{on}t}) \nu_t, \quad (2.6)$$

so that the equality holds when $\varepsilon \rightarrow 0$.

2.4.3 Coupling between time points

Let us recall that experimental data would be collections of cell positions in the protein space at distinct time points, and that the objective of cell trajectory inference is to reconstruct cell trajectories in this space. Hence, we restrict the problem to two consecutive time points t_1 and t_2 . Given two collections $X = (x_1, \dots, x_N) \in (\mathbb{R}_+^G)^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{R}_+^G)^N$, we search to construct a matrix $Q = (Q_{ij}) \in \mathbb{R}_+^{N \times N}$ of probabilities Q_{ij} that a cell at position x_i at time t_1 reaches the state y_j at time t_2 , in the most realistic way. In the following, the matrix Q will be called a coupling between the collections X and Y .

By solving the Schrödinger problem (see Section 2.4.4), we look for the coupling that is the closest to a certain reference coupling $R = (R_{ij}) \in \mathbb{R}_+^{N \times N}$, obtained using the Bursty model (2.1)-(2.2), while being compatible with the data X and Y . For either the model (2.1)-(2.2) or the Diffusive model, the reference coupling R is built by associating to the model a function being an exact or an approximated version of the transition kernel of the underlying theoretical process.

Coupling for the reduced Bursty model. In the case of a single gene following model (2.3), we use the regularized analytical solution (2.6) to compute the probability density of reaching state y starting from x , that is $p_t(y|x)$ defined by

$$p_t(y|x) = e^{-k_{on}t} h_t(y|x) + (1 - e^{-k_{on}t}) g_t(y|x)$$

where g_t is defined in (2.5) and

$$h_t(y|x) = \frac{1}{\sqrt{2\pi\varepsilon^2}} \exp \left\{ -\frac{(y - e^{-d_1 t} x)^2}{2\varepsilon^2} \right\}.$$

Considering this transition kernel $p_t(y|x)$ and the uniform distribution $p(x) \equiv 1/N$ over initial states $(x_1, \dots, x_N)$, the related reference coupling $RB_{\text{ref}}(x, y)$, for “reduced Bursty reference coupling”, is defined as:

$$RB_{\text{ref}}(x, y) = \frac{1}{N} \frac{p_{t_2-t_1}(y|x)}{c(x)}$$

where the normalizing term $c(x) = \sum_{1 \leq j \leq N} p_{t_2-t_1}(y_j|x)$ is defined so that $p_{t_2-t_1}(y|x)/c(x)$ is a discrete probability distribution over $(y_1, \dots, y_N)$.

Coupling for the non-reduced Bursty model. Given a GRN whose dynamics are given by a Bursty model (2.1)-(2.2), we construct a corresponding approximated coupling $B_{\text{ref}}(x, y)$, denoted most of the time B_{ref} , an approximation of its coupling between two collections at two different time points. It is the reference coupling defined by the probability of reaching y when starting from x , computed from the Bursty model (2.1)-(2.2).

We estimate B_{ref} using a Gaussian method (illustrated in Figure 2.13). For each cell i in the initial distribution μ at $t = t_1$, we simulate the Bursty model (2.1)-(2.2) n times with the cell i 's coordinates as an initial condition with the same parameters than those used to compute the simulated data detailed in Table 1 (with one exception, when we test the influence of GRN parameter values in Section 2.2.2.2: in this case, the parameter values used to compute the couplings are the ones of the tested GRN, not the ones used to generate simulated data). We obtain a distribution denoted $\tilde{\nu}$ at time $t = t_2$, representing the potential fates of cell i . An estimated density of $\tilde{\nu}$ is computed using Gaussian approximation, i.e. we create Gaussian

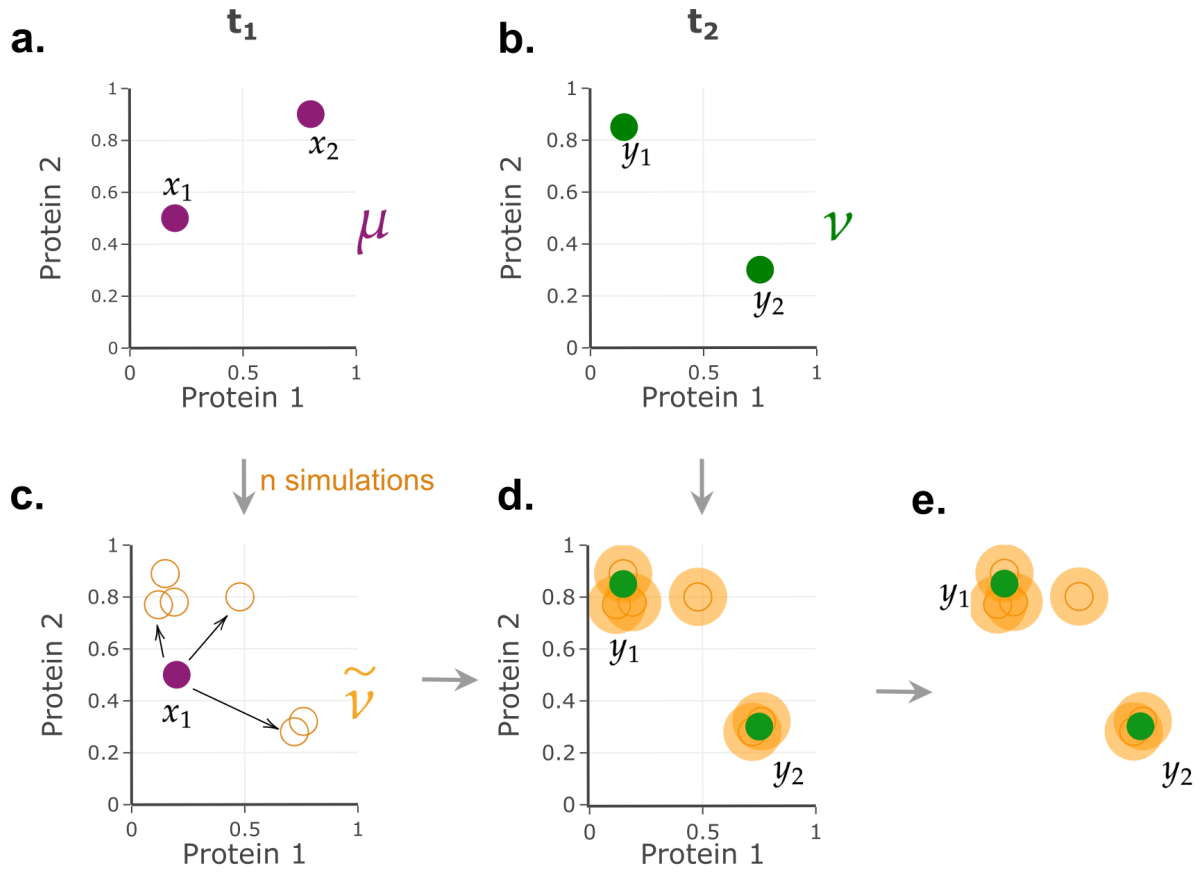


Figure 2.13. Schematic of the procedure used for constructing the B_{ref} estimation with a Gaussian method. **a.** In this example, $N = 2$ and cells x_1 and x_2 (purple dots) are the two cells of the distribution μ at time t_1 in protein space. **b.** Cells y_1 and y_2 (green dots) are the two cells belonging to the distribution ν at time t_2 in the protein space. **c.** The approximated distribution $\tilde{\nu}$ is computed for $n = 6$ and cells are displayed as orange circles. **d.** Gaussian distributions centered on $\tilde{\nu}$ cells generate a smooth distribution (orange areas). **e.** Probabilities $B_{\text{ref}}(y_1|x_1)$ and $B_{\text{ref}}(y_2|x_1)$ are deduced from the construction process.

distributions centered on cells in the distribution $\tilde{\nu}$, and we compute the Gaussian kernel of this distribution using the default parameters of the `gaussian_kde()` python function. The result is then divided by the total sum to ensure the resulting distribution is a probability distribution.

Coupling for a Diffusive process. When the underlying model is a Diffusive process, we consider the natural discrete counterpart used in (Lavenant et al., 2024). Specifically, the reference coupling denoted by D_{ref} is defined by

$$D_{\text{ref}}(x, y) = \frac{1}{\bar{D}_{\text{ref}}} \exp \left(-\frac{\|x - y\|^2}{2\sigma^2(t_2 - t_1)} \right)$$

where σ^2 is the diffusion coefficient, $\|x - y\|$ is the Euclidean distance (here, in protein space)

between x and y , and $\bar{D}_{\text{ref}}$ is the normalizing constant given by

$$\bar{D}_{\text{ref}} = \sum_{1 \leq i, j \leq N} \exp \left(-\frac{\|x_i - y_j\|^2}{2\sigma^2(t_2 - t_1)} \right).$$

The value of the diffusion coefficient was chosen to be both relevant and close to the value used in Schiebinger et al. (2019). The challenge was to consider a σ large enough to ensure a fast convergence of the Sinkhorn algorithm, while still capturing sufficiently accurate couplings ($D_{\text{ref}}(x, y) > 0$).

2.4.4 Schrödinger problem

In its modern formulation, the Schrödinger problem, first introduced in 1932 (Schrödinger, 1932), aims at minimizing the relative entropy of a coupling with respect to a reference coupling, with marginal constraints. This problem is deeply linked with OT: when the reference coupling is the joint law of a reversible Brownian motion with a given diffusion coefficient between times t_1 and t_2 , then the Schrödinger problem converges towards quadratic OT as the diffusion coefficient goes to 0 (Christian, 2013; Mariani, 2018).

Let Π and Ψ be two discrete and finite spaces, e.g. $\Pi = \{x_1, \dots, x_N\}$ and $\Psi = \{y_1, \dots, y_N\}$ with $N \in \mathbb{N}$, the number of cells. We denote by $R \in \mathcal{P}(\Pi \times \Psi)$, the reference process, where $\mathcal{P}(\Pi \times \Psi)$ is the space of probability measures on $\Pi \times \Psi$.

Since Π and Ψ are discrete spaces of size N , then R is a $N \times N$ matrix. We denote by $u \in \mathcal{P}(\Pi)$ and $v \in \mathcal{P}(\Psi)$ two uniform vectors of size N (each cell has the same weight), and we search for the process Q solution of

$$Sch(R; u, v) = \underset{Q \in \mathcal{P}(\Pi \times \Psi)}{\text{argmin}} \left\{ H(Q|R); \forall i, u_i = \sum_j Q_{ij}, \forall j, v_j = \sum_i Q_{ij} \right\}, \quad (2.7)$$

where H is the *relative entropy* of Q with respect to R (also called Kullback–Leibler divergence of Q from R) defined by

$$H(Q|R) = \sum_{ij} Q_{ij} \log \left(\frac{Q_{ij}}{R_{ij}} \right). \quad (2.8)$$

This quantity will be called *entropy* throughout this paper. The Schrödinger problem (2.7) can be solved with the Sinkhorn algorithm (Peyré & Cuturi, 2019a; Sinkhorn, 1967).

We solve the Schrödinger problem for the reference couplings introduced in Section 2.4.3. When $R = RB_{\text{ref}}$, we denote by RB_{sch} its solution, $RB_{\text{sch}} := Sch(RB_{\text{ref}}; u, v)$; for the reference coupling $R = B_{\text{ref}}$, we denote by B_{sch} its solution, $B_{\text{sch}} := Sch(B_{\text{ref}}; u, v)$; and when $R = D_{\text{ref}}$, then we denote D_{sch} its solution, $D_{\text{sch}} = Sch(D_{\text{ref}}; u, v)$.

2.4.5 Performances of the models.

In order to compare the performances of the Bursty model-based inference method with the Diffusion-based inference method, we rely on the computation of a standard criterion, the entropy (2.8) between solutions of the Schrödinger problem (2.7). However, to ensure that results do not depend on the criterion, we also test other criteria measuring the difference between two stochastic processes.

Entropy between Schrödinger problem solutions. For one-gene GRN, we define

$$H_{B,R} := H(B_{\text{sch}}|RB_{\text{sch}}) \quad \text{and} \quad H_{D,R} := H(D_{\text{sch}}|RB_{\text{sch}}),$$

i.e. the entropy between B_{sch} or D_{sch} and RB_{sch} . These values characterize the ability of a Diffusive process or a Bursty model versus RB_{sch} to predict cell trajectories correctly. The process RB_{sch} is considered as the reference process because it provides an approximation of the exact probability distribution. In parallel, we compute as a control measure

$$H_R := H(RB_{\text{ref}}|RB_{\text{sch}})$$

in order to estimate the error made when performing stochastic simulations.

For two-or-more-gene GRN, we define

$$H_{D,B} := H(D_{\text{sch}}|B_{\text{sch}}),$$

the entropy between D_{sch} and B_{sch} , in order to evaluate the difference between the ability of a Diffusive process versus a Bursty model to correctly predict cell trajectories. Similarly to the one-gene GRN case above, we introduce

$$H_B := H(B_{\text{ref}}|B_{\text{sch}})$$

to estimate the error made when performing stochastic simulations.

Values of H_B should be smaller than $H_{D,B}$ values if B_{ref} and B_{sch} are closer than D_{sch} and B_{sch} . If the Bursty model-based reference process better predicts individual cell trajectories, the difference between H_B and $H_{D,B}$ will be important.

Due to stochasticity, we perform 100 iterations to compute these quantities and focus on the mean and standard deviation.

Other difference indicators. To certify that our results do not depend on the choice of the entropy, we also compute the following criteria, for two processes Q and R :

- the Total Variation Distance:

$$D_{\text{TV}}(Q, R) = \frac{1}{2} \sum_i \sum_j |Q_{ij} - R_{ij}|;$$

- the Jensen–Shannon Divergence:

$$\text{JSD}(Q, R) = (H(Q | W) + H(R | W))/2,$$

where $W = (Q + R)/2$;

- the Frobenius norm:

$$\|Q - R\|_F = \sqrt{\sum_{i,j} (Q_{ij} - R_{ij})^2}.$$

2.4.6 Application to Semrau et al. (2017) dataset

We use a time-stamped in vitro dataset from Semrau et al. (Semrau et al., 2017) made of mRNA counts obtained from scRNA-seq experiments on mouse embryonic stem cells. The dataset comprises 41 genes whose levels of expression were measured at 9 observation times (0h, 6h, 12h, 24h, 36h, 48h, 60h, 72h and 96h) with approximately hundreds of cells obtained at each observation time. Genes are listed in Figure 2.9. Ventre et al. (2023) inferred a GRN based on the Bursty model driving the dynamics observed in the dataset. We use this GRN and the associated parameter values to compute the reference coupling B_{ref} , similarly to what has been done with protein inputs but using mRNA counts. Compared to the work done on simulated protein data, we use a larger number of simulations when constructing the approximated reference process, $n = 10,000$, while other parameter values are given in Table 2.1.

Given two experimental distributions μ at time t_1 and ν at time t_3 , we hereafter describe how we predict the distribution at time $t_2 \in (t_1, t_3)$.

By simulating the Bursty model with the GRN from Ventre et al. (2023), we first compute $B_{\text{ref}}(t_1, t_3)$, the reference coupling between times t_1 and t_3 , and a reference distribution ρ_{ref} at time t_2 . Second, we solve the Schrödinger problem and determine the coupling $B_{\text{sch}}(t_1, t_3)$. We sample cell pairs (x, y) based on the coupling $B_{\text{sch}}(t_1, t_3)$: the goal is to generate a certain number of pairs for each cell x in μ , where the probability of selecting a cell y in ν is determined by $B_{\text{sch}}(t_1, t_3)$. Third, we build an interpolated distribution at t_2 , rooted in distributions μ and ν , as described below.

We define the estimated, interpolated distribution at time t_2 as

$$\rho_{\text{est}} := T_{t_2}^{\alpha*} \# B_{\text{sch}}(t_1, t_3), \quad (2.9)$$

with $T_{t_2}^{\alpha^*} : (x, y) \mapsto \alpha^* x + (1 - \alpha^*) y$ for each pair (x, y) , $\alpha^* \in \mathbb{R}^G$ is an optimized vector and $\#$ is the push-forward function. Details of the optimization procedure leading to the estimation of α^* are provided in Supplementary File. We compare the estimated distribution ρ_{est} with the experimental distribution, using the Earth mover's distance (EMD), denoted by EMD_B . We also compute the coupling between the distributions μ at t_1 and ν at t_3 using regularised OT (similarly to WOT (Schiebinger et al., 2019)), denoted by $D_{\text{sch}}(t_1, t_3)$. In this case, we use a log-transform of mRNA counts, we do not account for growth and death since this is not relevant in this case, and we interpolate a distribution at t_2 as in Schiebinger et al. (2019) by using $\alpha^* = (t_2 - t_1)/(t_3 - t_1) = 0.5$ in (2.9) for all genes. We then compare the interpolated distribution and the experimental distribution by computing the EMD, denoted by EMD_D . Finally, we compare the two distances, EMD_B and EMD_D , by computing

$$\Delta_{EMD} = EMD_D - EMD_B.$$

Table 2.1: **Default parameter values.**

Description	Parameter	Value	Unit
Number of genes	G	2	genes
Number of cells	N	100	cells
Number of simulations	n	100	simulations
Initial conditions for mRNA counts	IC_M	(0, 0)	mRNA
Initial conditions for protein counts	IC_P	(0, 0)	protein
Smoothing standard deviation	ε	10^{-4}	protein
Creation rate for proteins of gene i	s_i	10	protein h^{-1} mRNA $^{-1}$
Minimum burst frequency of gene i	$k_{0,i}$	0	h^{-1}
Maximum burst frequency of gene i	$k_{1,i}$	2	h^{-1}
Basal activity of gene i	β_i	5	h^{-1}
Degradation rate of mRNA of gene i	$d_{0,i}$	10	h^{-1}
Degradation rate of protein of gene i	$d_{1,i}$	5	h^{-1}
Intensity of self-activation of gene i	θ_{ii}	10	
Intensity of inhibition of gene i	θ_{ij}	-10	
Initial time point in transient state	t_1	0.1	h
Final time point in transient state	t_2	0.5	h
Diffusion coefficient	σ^2	0.001/2	protein ^{2} h^{-1}
Precision of Sinkhorn algorithm		0.1	

2.5 Data Availability

All results included in this manuscript can be reproduced using the codes mentioned below (2.6). Experimental data used in this work have been provided by Semrau et al. (2017) and are accessible at <https://www.ncbi.nlm.nih.gov/geo/> under accession number GSE79578.

2.6 Code Availability

All python codes are available at

https://github.com/fourniecl/inference_trajectories_PDMP.

2.7 Acknowledgments

We thank the computational center of IN2P3 (Villeurbanne/France), where part of the computations were performed.

This work has received financial support from the CNRS through the MITI interdisciplinary programs through its exploratory research program (TROPIC project under Action 80PRIME 2023). This work was supported by the PEPR Santé Numérique under Project AI4scMED MultiScale AI for Single Cell-Based Precision Medicine (ANR-22-PESN-0002).

2.8 Author Contributions

CF performed the analyses and numerical simulations, and wrote the initial version of the manuscript. OG and FC supervised the work. All authors designed the methodology, and read and approved the final manuscript.

2.9 Competing Interests

The authors declare no competing financial or non-financial interests.

2.10 Supplementary information

2.10.1 Analytical solution for the reduced Bursty model

Let us consider the reduced Bursty model

$$\begin{cases} P(t) \xrightarrow{k_{on}} P(t) + \mathcal{E}(1/\eta), \\ \dot{P}(t) = -d_1 P(t), \end{cases}$$

where $P(t)$ is the protein level at time t , $\mathcal{E}$ is the exponential distribution characterizing the random size of each burst (with mean value η), k_{on} is the burst rate, assumed to be constant, and d_1 is the degradation rate of proteins.

We briefly explain how to derive the formula

$$\mathcal{L}(P(t)|P(0) = P_0) = e^{-k_{on}t} \delta_{e^{-d_1 t} P_0} + (1 - e^{-k_{on}t}) \nu_t$$

where $\delta_{e^{-d_1 t} P_0}$ is the Dirac measure localized at $e^{-d_1 t} P_0$, and $\nu_t = \delta_{e^{-d_1 t} P_0} * \mu_t$ with μ_t the probability measure defined by the following density function:

$$f_t(x) = \frac{1}{\eta} \frac{(e^{d_1 t} - 1) k_{on}}{(e^{k_{on} t} - 1) d_1} \exp\left(-\frac{x}{\eta} e^{d_1 t}\right) \Phi\left(\frac{k_{on}}{d_1} + 1; 2; (e^{d_1 t} - 1) \frac{x}{\eta}\right) \mathbb{1}_{\{x \geq 0\}}, \quad (2.10)$$

where Φ denotes the confluent hypergeometric (Kummer's) function of the first kind. Hence the probability measure ν_t has a density function $g_t(x|P_0)$ given by

$$g_t(x|P_0) = f_t(x - e^{-d_1 t} P_0). \quad (2.11)$$

The corresponding Markov process $(X_t)_{t \geq 0}$, with values in $\mathbb{R}_+$, is characterized by its infinitesimal generator L , defined in Malrieu (2015) *Some simple but challenging Markov processes*, Annales de la Faculté de Sciences de Toulouse, 24, 857–883,

$$Lf(x) = -d_1 x f'(x) + k_{on} \int_0^\infty (f(x+h) - f(x)) \frac{1}{\eta} e^{-\frac{1}{\eta} h} dh,$$

from which an evolution equation for the Laplace transform $\phi(s, t) = \mathbb{E}[e^{-sX_t}]$ can be derived, leading to the general solution

$$\phi(s, t) = \phi(se^{-d_1 t}, 0) \left(\frac{1 + \eta e^{-d_1 t} s}{1 + \eta s} \right)^{\frac{k_{on}}{d_1}}.$$

The first term in this product can be identified as the Laplace transform of $e^{-d_1 t} X_0$, while the second term is the Laplace transform of X_t conditional on $X_0 = 0$ (i.e., $\mathcal{L}(X_t|X_0 = 0)$). It

turns out that the latter can be inverted explicitly using a classical series known as the confluent hypergeometric function of the first kind (denoted by $\Phi(a; b; z)$), which is of practical interest since this special function is implemented in standard computing libraries such as SciPy. Manipulation of the series leads to the probability measure

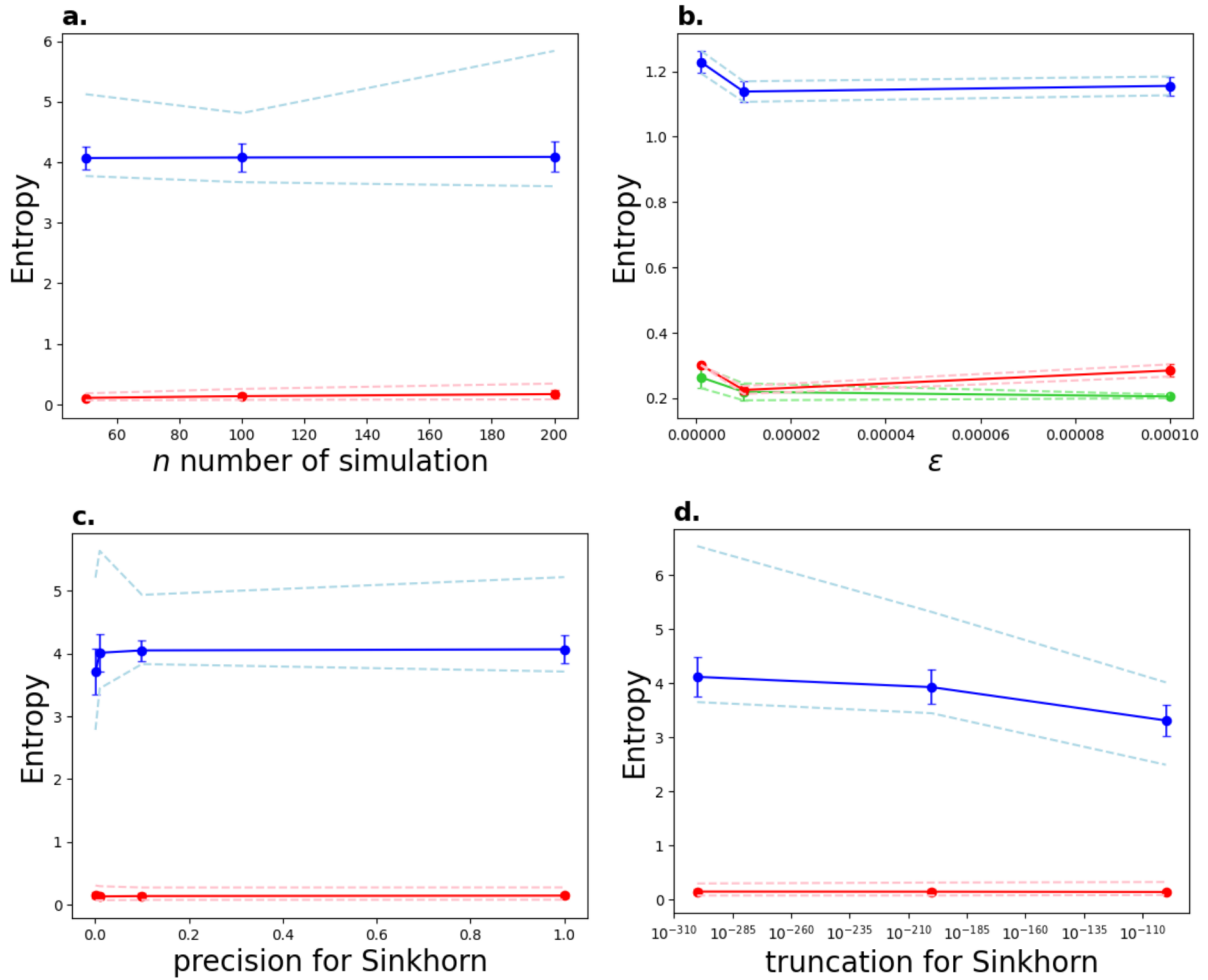
$$\mathcal{L}(X_t|X_0 = 0) = e^{-k_{on}t}\delta_0 + (1 - e^{-k_{on}t})\mu_t,$$

where μ_t has density function $f_t(x)$ given in (2.10), and the final expression (2.10)-(2.11) follows.

Interestingly, the weight of the singular part (Dirac measure at $e^{-d_1t}P_0$) decreases exponentially with rate k_{on} , in line with the property that in this model, a trajectory ceases to be deterministic precisely when the first burst occurs.

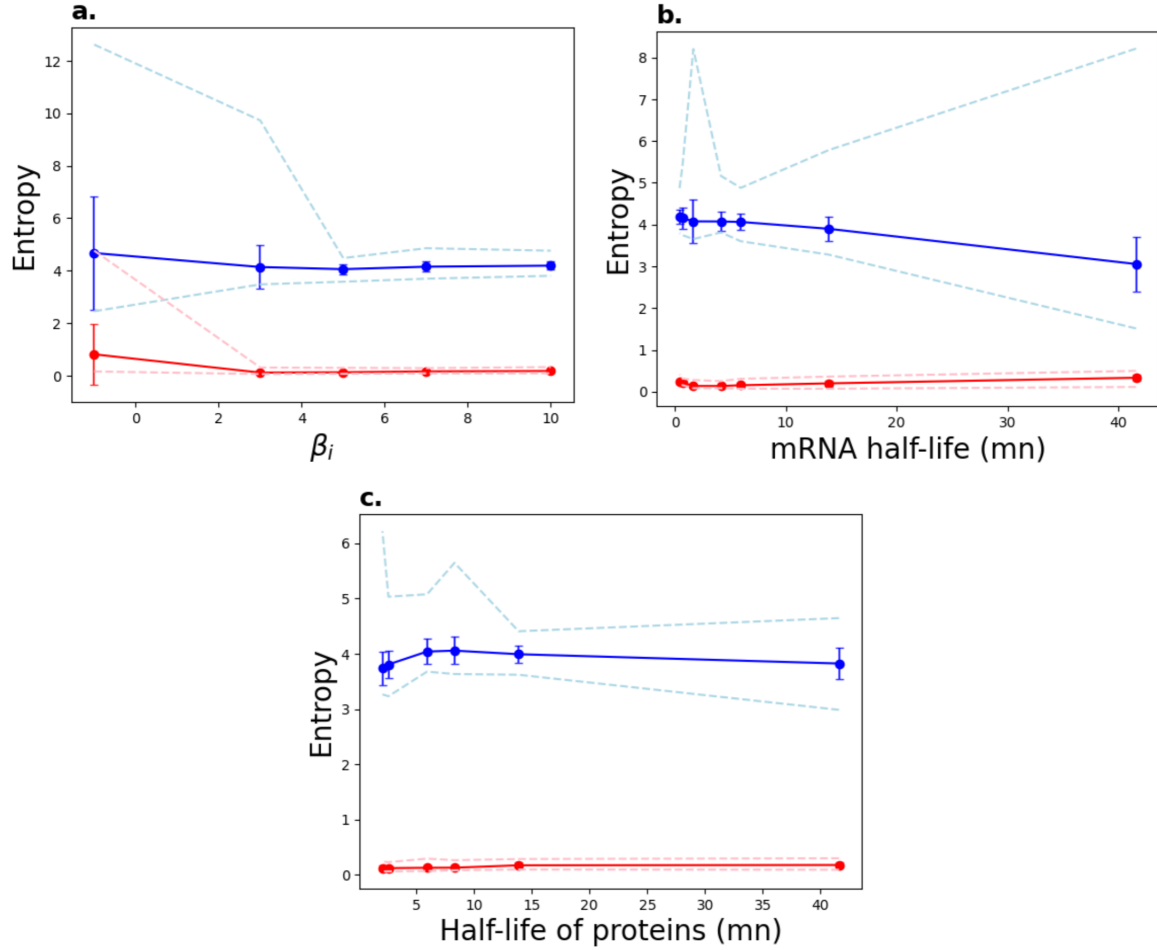
2.10.2 Sensitivity results

Supp. Figure 1 shows sensitivity results for some parameters of the Bursty model-based inference method. We focused on the impact on the Entropy $H_{D,B}$ of the number of simulations used to compute B_{ref} (Supp. Figure 1a), the standard deviation ε of the Gaussian law in (2.6) (Supp. Figure 1b), the precision of Sinkhorn algorithm (Supp. Figure 1c) and the truncation used to avoid zeros in Sinkhorn algorithm (Supp. Figure 1d). Results highlight a weak sensitivity to variations in these parameters, as long as their values remain reasonable.



Supp. Figure 1. Entropy as a function of hyperparameters. Mean (dots), standard deviation, min and max values (dashed lines) of $H_{D,B}$ (blue), H_B (red) and H_R (green, only in b.) are represented as functions of **a.** the number of simulations used to compute B_{ref} ; **b.** ε , the standard deviation of the Gaussian law in (6); **c.** the precision of Sinkhorn algorithm; **d.** the truncation used to avoid zeros in Sinkhorn algorithm. All other parameters are fixed and given in Table 2.1 (main text).

Supp. Figure 2 shows sensitivity results for GRN parameters β_i the basal activity rate, the half-life of mRNA $\log(2)/d_{0,i}$, and the half-life of proteins $\log(2)/d_{1,i}$, of gene i (note that all genes in the Toggle Switch are assumed to have the same parameter values). Their impact on the entropies $H_{D,B}$ and H_B is limited.

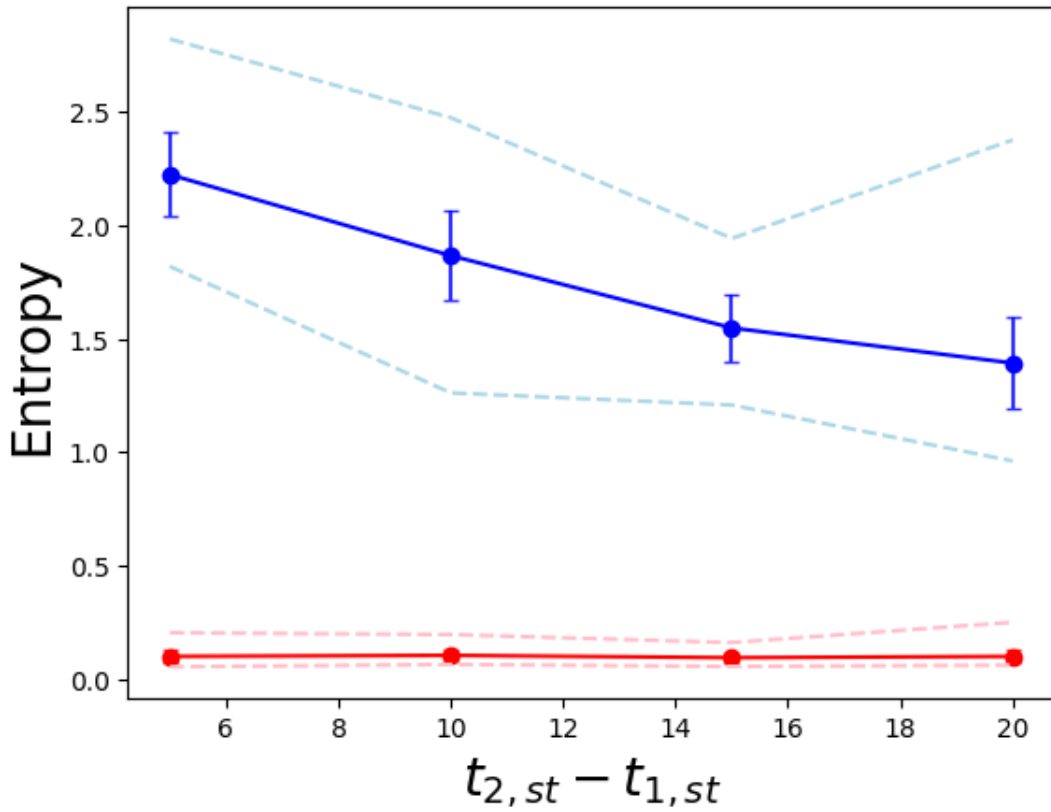


Supp. Figure 2. Entropy as a function of GRN parameters. Mean (dots), standard deviation, minimum and maximum values (dashed lines) of $H_{D,B}$ (blue) and H_B (red) as functions of **a.** the rate of basal activity, β_i ; **b.** the half-life of mRNA, $\log(2)/d_{0,i}$; **c.** the half-life of proteins, $\log(2)/d_{1,i}$, of gene i (all genes have the same parameter values by default). All other parameters are fixed and given in Table 2.1 (main text).

2.10.3 The stationary case

We studied the influence of the time difference $t_2 - t_1$ on the solutions of the Schrödinger problem when the system is in a stationary state. To do so, we defined $t_{1,st} = 40$ h and $t_{2,st} \in [42; 64]$ h. Results are displayed in Supp. Figure 3.

Results are similar to the transient state case, except that the $H_{D,B}$ curve decreases as $t_{2,st}$ increases. This shows that despite the Diffusive model being less accurate than the Bursty model to infer trajectories, their performances become comparable when the final distribution becomes independent of the initial one, that is for large $t_{2,st}$ values.



Supp. Figure 3. Entropies between reference couplings for the toggle switch GRN, in the stationary state. Mean and standard deviation (straight lines) of $H_{D,B}$ (blue) and H_B (red) are displayed as functions of the time interval $t_{2,st} - t_{1,st}$. Additionally, minimum and maximum values over 100 iterations are represented (dashed lines). Other parameters are given in Table 2.1 (main text).

2.10.4 Interpolation parameter α and empirical measures

The interpolation application at time t_2 is defined as

$$T_{t_2}^\alpha : (x, y) \mapsto \alpha x + (1 - \alpha)y, \quad \alpha \in \mathbb{R}^G. \quad (2.12)$$

The optimized value of α , denoted by α^* , is obtained as the solution of a minimisation problem:

$$\alpha^* = \arg \min_{\alpha} \mathcal{L}(\alpha),$$

where the cost function $\mathcal{L}$ is the quadratic Wasserstein distance between the empirical distribution ρ_{ref} at t_2 and the empirical interpolated distribution $T_{t_2}^\alpha \# B_{\text{ref}}(t_1, t_3)$, given by

$$\mathcal{L}(\alpha) := W_2^2(T_{t_2}^\alpha \# B_{\text{ref}}(t_1, t_3), \rho_{\text{ref}}).$$

The optimization is performed with the Python package geomloss, using the SamplesLoss function (<https://www.kernel-operations.io/geomloss/index.html>), and minimization is done with PyTorch.

Once α^* is found, the estimated distribution at time t_2 is defined as

$$\rho_{\text{est}} := T_{t_2}^{\alpha^*} \# B_{\text{sch}}(t_1, t_3).$$

In practice, we consider empirical measures. The push-forward function of the reference coupling by $T_{t_2}^\alpha$ is given explicitly by

$$T_{t_2}^\alpha \# B_{\text{ref}}(t_1, t_3) = \frac{1}{N} \sum_{i=1}^N \delta_{\alpha x_i + (1-\alpha)y_i},$$

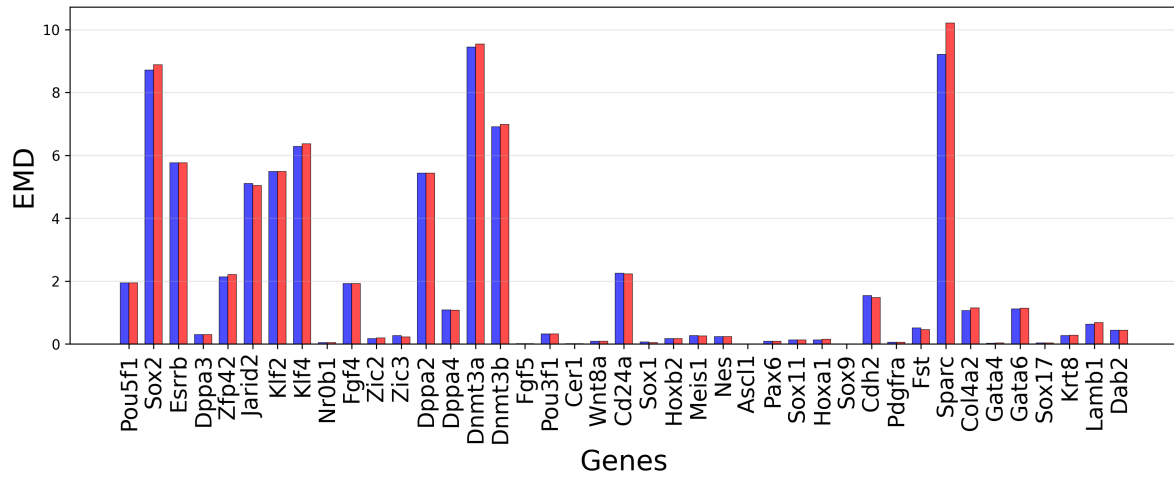
where N is the number of sampled pairs (x_i, y_i) from the coupling $B_{\text{ref}}(t_1, t_3)$, that is N equals the number of cells x times the number of samples for each cell. δ denotes the Dirac measure.

Similarly, the reference distribution at time t_2 writes

$$\rho_{\text{ref}} = \frac{1}{m} \sum_{j=1}^m \delta_{z_j}, \quad z_j \in \mathbb{R}^G,$$

with m depending on the number of simulated vectors from the initial distribution μ .

2.10.5 Comparison of the Bursty model-based inference method and the Diffusion-based inference method for the experimental data



Supp. Figure 4. EMD between Bursty model-based and Diffusion-based inference methods and experimental data for each gene in the dataset. The EMDs are computed as described in Section 2.6, at the interpolated time $t_2 = 36h$, based on the coupling between times $t_1 = 0h$ and $t_3 = 72h$. In blue are shown EMD between the Bursty model-based prediction and the experimental data, and in red the EMD between the Diffusion-based prediction and the experimental data.

Supp. Figure 4 shows the values of the Earth Mover's Distances (EMD) between the experimental data and the inference methods (Bursty model-based inference method in blue, Diffusion-based inference method in red). They are the EMD used to compute Δ_{EMD} in Figure 2.9.

3

Discussion

3.1 Résumé des contributions

Dans mon travail de thèse, j'ai étudié l'hypothèse selon laquelle le modèle de gènes dit *Bursty*, qui permet de décrire une dynamique aléatoire markovienne d'un réseau de régulation génique générant des séries temporelles réalistes, constituerait une alternative au modèle de diffusion standard, utilisé classiquement dans la théorie du transport optimal, pour inférer des trajectoires de différenciation. La *méthode d'inférence basée sur le modèle Bursty*, que j'ai introduite, s'est avérée efficace pour capturer les trajectoires cellulaires, avec une plus grande précision que l'inférence réalisée avec un processus diffusif.

Pour valider cette méthode, j'ai inféré des trajectoires de différenciation sur des données simulées, à la fois dans l'espace des niveaux de protéines et dans l'espace d'expression des comptes d'ARNm, et des données expérimentales, dans l'espace d'expression des ARNm uniquement.

En utilisant des données simulées dans les deux espaces, l'inférence est pertinente même si les résultats sont meilleurs lorsqu'on utilise les données protéiques. Cela n'est pas surprenant : les protéines sont des molécules plus stables, dont les dynamiques sont plus lentes et plus régulières, ce qui les rend plus compatibles avec le cadre des processus de Markov sur lequel repose le modèle *Bursty*. En revanche, les ARNm présentent une variabilité plus élevée et une demi-vie plus courte. Ce contraste motive les réflexions sur la réduction du modèle discutées plus bas.

Pour les données expérimentales de Semrau et al. (2017), l'interpolation réalisée avec la *méthode d'inférence basée sur le modèle Bursty* a donné de meilleurs résultats que celle réalisée avec la méthode d'inférence basée sur la diffusion. Toutefois, on note que la méthode

d'inférence basée sur la diffusion fonctionne bien lorsque les observations sont effectuées sur une courte période et lorsque l'activité des gènes est faible. En revanche, lorsque les observations sont effectuées sur des périodes plus longues et que les gènes présentent une activité importante, la *méthode d'inférence basée sur le modèle Bursty* est systématiquement plus performante.

Ce travail illustre plus généralement l'intérêt d'incorporer une connaissance mécaniste du système biologique dans le processus d'inférence. L'utilisation d'un processus de référence prenant en compte les dynamiques biologiques permet d'orienter les couplages vers des trajectoires cohérentes avec la dynamique sous-jacente. Dans ce chapitre, je discute certains aspects de mon travail et présente quelques perspectives.

3.2 Prétraitement des données

Dans cette section, je discute les choix que j'ai faits quant au prétraitement des données. Les données simulées dans cette thèse sont des données générées à partir du modèle Bursty, et les données expérimentales sont issues d'une expérience de scRNA-seq réalisée par Semrau et al. (2017). Dans les deux cas, une étape de prétraitement sur les valeurs d'expression des gènes a été réalisée avant de réaliser l'inférence.

Pour les données simulées comme pour les données expérimentales, j'ai normalisé les données par rapport à chaque expérience. J'ai fait une mise à l'échelle en divisant les valeurs d'une distribution par son maximum (par exemple, la distribution μ est normalisée par $\max(\mu)$). J'ai choisi de normaliser les données principalement afin de m'appuyer sur l'état de l'art ; en effet, WOT (Schiebinger et al., 2019) et les travaux qui s'en sont inspirés (TrajectoryNet (Tong et al., 2020), MIOFlow (Huguet et al., 2022), CellOT (Bunne et al., 2023)...) normalisent leurs données. Ensuite, j'ai normalisé pour faciliter la comparaison des analyses : toutes les trajectoires sont ainsi à la même échelle transcriptomique, quel que soit le processus de référence utilisé ou les intervalles de temps étudiés. Enfin, pour des temps longs (plus d'une heure), la mesure de l'entropie sans normalisation présentait une variance bien plus élevée que celle avec normalisation. Il est toutefois important de souligner que cette étape implique que les résultats d'inférence sont obtenus sur des trajectoires dans un espace normalisé, et non dans l'espace d'expression des gènes par défaut.

Pour les données expérimentales, j'ai en plus appliqué une transformation logarithmique, telle que la valeur x_g , l'expression d'un gène g d'une cellule x , soit associée à la quantité $\log(x_g - 1)$. Cette transformation est classique en analyse de données scRNA-seq (Anscombe, 1948; Harrison, 2017), elle permet de ramener à une même échelle des différences d'expression dans les

gènes. En effet, il ne faudrait pas qu'un gène sur-exprimé prédomine et gomme les variations significatives des gènes moins exprimés. Cette transformation n'a pas été appliquée aux données simulées car celles-ci sont générées directement par le modèle Bursty dans un espace restreint et normalisé. Il ne me semblait pas nécessaire d'appliquer cette transformation en plus pour des données simulées.

Pour conclure cette section, la normalisation est plutôt un choix de praticité pour l'analyse tandis que la transformation logarithmique est un choix plus important pour la cohérence dans le traitement des données.

3.3 Modélisation et paramétrisation du processus diffusif

Dans le Chapitre 2, on a comparé les résultats obtenus par l'inférence avec le processus de référence à base du modèle Bursty aux résultats obtenus avec le processus diffusif. Aucun travail n'utilise directement le problème de Schrödinger pour l'inférence de trajectoire, et on a cherché à comparer notre méthode à l'état de l'art (section 1.5). Dans cette section, je discute des choix de paramétrisation que nous avons dû faire pour paramétrer le processus de diffusion. Le but n'était pas de trouver le processus de référence de diffusion qui infère le mieux les trajectoires cellulaires, mais de paramétrer un processus qui reflète au mieux l'état de l'art, pour le comparer à la *méthode d'inférence basée sur le modèle Bursty*.

Les paramètres du processus diffusif utilisés ont été choisis pour être cohérents avec la distance euclidienne utilisée dans l'état de l'art, en particulier dans le travail de Schiebinger et al (2019), et compatibles avec notre formalisme. La valeur du coefficient de diffusion, σ^2 dans l'expression (1.5), doit être la plus petite possible pour la convergence de l'algorithme de Sinkhorn, tout en maintenant une réalité biologique, c'est à dire capturer des couplages suffisamment précis. Si on avait choisi un σ^2 trop grand, le processus de diffusion aurait permis aux cellules de diffuser dans tout l'espace et les couplages trouvés n'auraient pas du tout été spécifiques, toutes les cellules auraient été couplées entre elles. À l'inverse, un σ^2 trop petit permet d'atteindre un espace trop restreint ; $D_{\text{ref}}(x, y)$ est proche de 0, pour tout x et y , les cellules ne se coupleraient pas. La valeur du coefficient de diffusion a été choisie pour être à la fois pertinente et proche de la valeur utilisée dans Schiebinger et al. (2019) : $2\sigma^2 = 0.001$.

Il existe d'autres manières de modéliser le processus diffusif, il peut aussi être associé au noyau de la chaleur. Nous n'avons pas choisi ce formalisme, car il n'est pas restreint à $\mathbb{R}_+^G$, où G est le nombre de gènes (et de dimensions).

La paramétrisation du processus diffusif a un impact sur l'évaluation de la méthode. On aurait pu choisir un cadre plus avantageux pour la paramétrisation du processus diffusif comme un exemple de GRN sans inhibition, optimiser le coefficient de diffusion, regarder des temps plus

courts ou appliquer la méthode sur un espace réduit par analyse en composantes principales (ACP). Mais le processus diffusif ne sert que de comparaison avec les méthodes de l'état de l'art. Ainsi, les résultats obtenus ne sont pas les plus optimaux pour le processus diffusif mais les plus cohérents avec la littérature.

3.4 Perspectives

Dans cette section, je rappelle certaines des hypothèses faites pour la modélisation et propose des pistes de réflexions et d'améliorations pour alléger les hypothèses. D'abord, je vais plus loin en abordant le cas où les distributions sont en état stationnaire. Ensuite, je discute le fait que mon travail soit réalisé principalement sur des données protéiques alors que les données de scRNA-seq sont des données ARNm. Après, je propose des pistes pour inclure la prolifération et la mort. De plus, je présente une nouvelle technologie qui pourrait fournir des données alternatives aux données simulées pour tester la méthode, de façon plus réaliste. Et enfin, je propose une extension du modèle Bursty pour inclure l'aspect spatial de la différenciation.

3.4.1 Étude des clusters dans le cas de la distribution stationnaire

Dans cette section, je reviens sur l'analyse des résultats en état stationnaire de la section 2.10.3, pour l'approfondir. Cette analyse permet de proposer une autre manière de tester la méthode en profitant de la structure particulière de l'état stationnaire du toggle-switch.

Dans le cas stationnaire (comme détaillé dans la section 1.8.3), c'est à dire après un certains temps de simulation, on observe deux clusters de cellules : la distribution stationnaire est bimodale. Les cellules évoluent soit au sein de leur cluster (la quantité d'une seule protéine augmente ou diminue, tandis que l'autre est très peu exprimée), soit entre les deux clusters (une protéine prend l'avantage sur l'autre). J'étudie les regroupements en cluster de cellules caractéristiques du toggle-switch en état stationnaire.

Étudier la capacité de la *méthode d'inférence basée sur le modèle Bursty* à retrouver ces regroupements en cluster est une autre alternative pour déterminer sa performance. On s'attend à ce que la méthode trouve des couplages qui ne sont pas individuellement exacts mais soient au sein du même cluster. La reconstruction des couplages individuels n'est pas le but recherché. En effet, dans les données expérimentales, il n'y a pas de couplage réel deux à deux.

Les résultats sur des données simulées présentés dans la section 2.2 reposent principalement sur des temps courts, proches du temps initial. Le choix de t_1 et t_2 proches de 0 implique que la dynamique n'a pas atteint son équilibre, c'est ce que j'ai appelé l'état transitoire. Quand t_1 et t_2 sont suffisamment grands (dans le cas des protéines du toggle-switch, c'est au-delà de 5h), on observe deux clusters de cellules. La dynamique du réseau fait que, pour la plupart

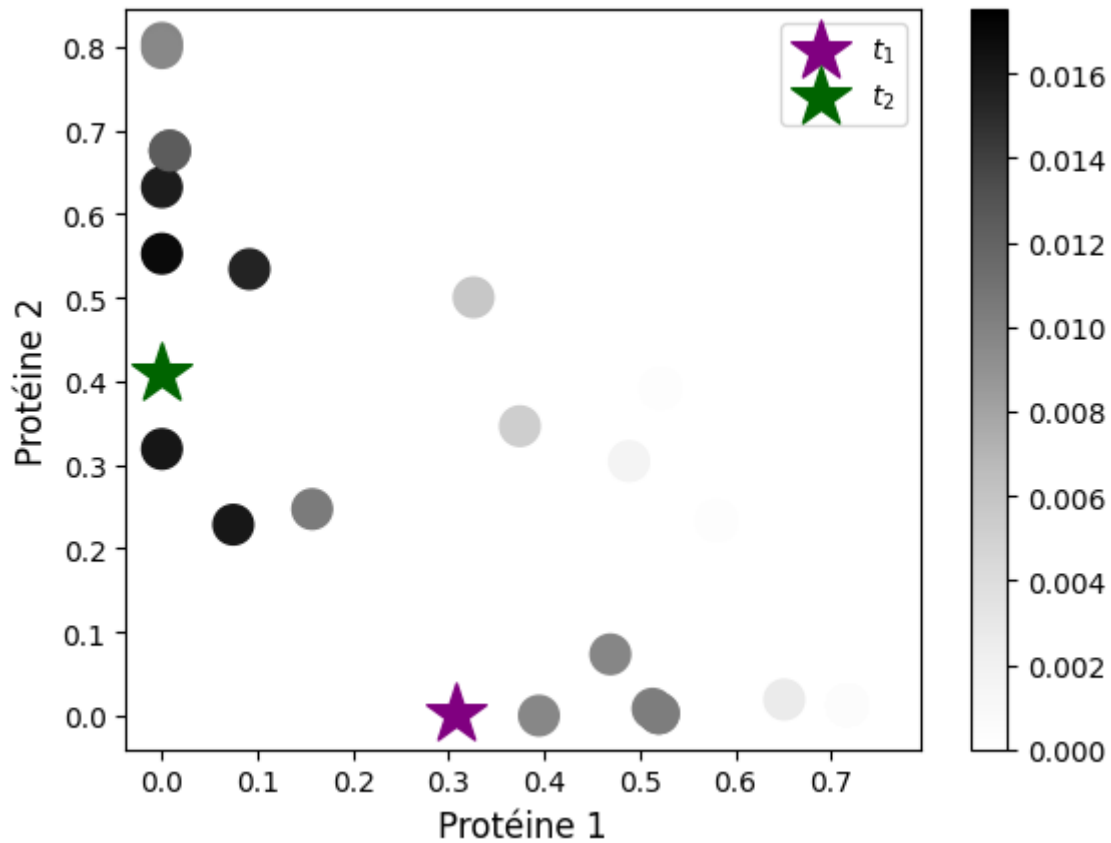


Figure 3.1. Dans l'espace d'expression des protéines, une cellule commence à t_1 (l'étoile violette) et termine à t_2 (l'étoile verte). Les points représentés sont les meilleurs couplages trouvés avec B_{sch} partant de la cellule initiale à t_1 . Plus la couleur est foncée plus la probabilité du couplage est élevée.

des cellules, une seule des deux protéines est exprimée ; ainsi, les deux clusters correspondent aux deux cas où la protéine 1 ou la protéine 2 est exprimée. Les deux clusters ne sont ni fixés ni délimités au cours du temps ; les cellules évoluent soit au sein de leur cluster, la quantité d'une seule protéine augmente ou diminue tandis que l'autre est très peu exprimée, soit entre les deux clusters, une protéine prend l'avantage sur l'autre. Ces deux regroupements sont caractéristiques de la dynamique dite stationnaire du toggle-switch.

Après avoir évalué les performances de la méthode d'inférence basée sur le modèle *Bursty* en régime stationnaire, dans la section 2.10.3, j'ai étudié plus en détail le comportement du toggle-switch dans ce régime stationnaire. Je propose une discussion plus approfondie de la capacité du modèle *Bursty* à capturer les dynamiques entre et au sein des clusters.

La Figure 3.1 illustre le comportement d'une cellule de la distribution μ , en position x_0 à t_1 et y_0 à t_2 , dans la distribution ν . Cette cellule change de clusters entre t_1 et t_2 . La Figure 3.1 représente les 25 meilleurs couplages identifiés par la méthode d'inférence basée sur le modèle *Bursty*, nuancés en fonction de leur probabilité. Même si le couplage de la méthode d'inférence faite avec le modèle *Bursty* ne retrouve pas le vrai couplage individuel (x_0, y_0) , on voit qu'il attribue une probabilité plus élevée aux cellules du même cluster que y_0 . Les cellules avec la plus

forte probabilité (points foncés) sont proches de la cellule y_0 (étoile verte). La Figure 3.1 montre que la méthode d'inférence capture bien la structure des clusters et retrouve correctement le cluster correspondant dans ce cas, même si elle ne peut pas identifier précisément le devenir des cellules individuellement. Cette illustration est un cas pour une cellule, je cherche à voir si cet exemple est représentatif et que la conclusion est générale. Le cas de x_0 présenté dans la Figure 3.1 est un cas idéal, où en faisant le test de Pearson¹, la valeur de r avoisine les -0.78 (avec une valeur p de $p = 3 \times 10^{-21}$).

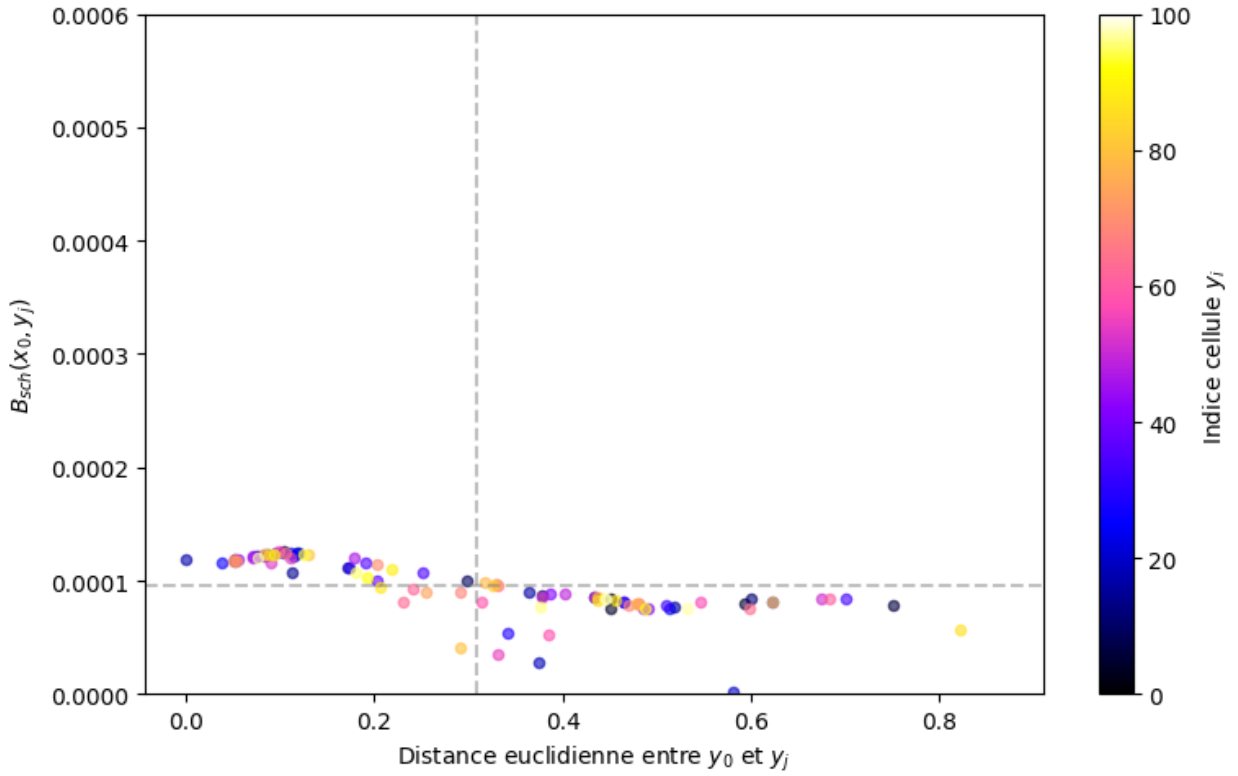
Au-delà de l'exemple d'une seule cellule, la Figure 3.2A. représente l'ensemble des cellules y_i de ν en fonction de la distance euclidienne avec y_0 en abscisses et la valeur du couplage $B_{sch}(x_0, y_i)$ en ordonnées. Le couplage réel est entre x_0 et y_0 , je cherche à évaluer si les y_i avec une valeur élevée de B_{sch} sont proches de y_0 . Idéalement, les couplages de forte probabilité devraient correspondre à de faibles distances euclidiennes : un point en haut à gauche du graphe indique que le couplage le plus probable est également le plus proche du couplage réel. On aimerait qu'il y ait une corrélation négative entre la distance euclidienne avec y_0 et la valeur du couplage $B_{sch}(x_0, y_i)$. En faisant le test de Pearson pour l'ensemble des cellules, on obtient $r = -0.0718$, avec une valeur p de 6.5×10^{-13} . La corrélation est faiblement négative. Comme le montre le test de Pearson, une forte corrélation négative n'est pas systématique, ce qui indique que majoritairement, la méthode d'inférence basée sur le modèle Bursty capture bien la structure globale du cluster, mais pas systématiquement et demeure imprécise au niveau des trajectoires individuelles.

J'ai représenté les mêmes quantités pour la méthode basée sur le processus diffusif dans la Figure 3.2B. Les valeurs de D_{sch} sont concentrées autour de zéro pour la grande majorité des cellules, avec quelques valeurs extrêmes à de grandes distances. Le test de Pearson pour l'ensemble des cellules, aboutit à $r = -0.0446$, avec une valeur p de 8.1×10^{-6} . La corrélation est légèrement moins négative avec D_{sch} qu'avec B_{sch} . Le processus diffusif répartit les probabilités uniformément sur l'ensemble des cellules, sans distinction entre les deux clusters. Il est donc moins capable de capturer les transitions entre les clusters alors que le modèle Bursty le fait.

Ces résultats s'ajoutent aux résultats de la section 2.10 et donnent des informations plus explicites sur les couplages que l'entropie. Ils soutiennent également que la méthode d'inférence basée sur le modèle Bursty constitue une meilleure approximation pour l'inférence des trajectoires cellulaires en régime stationnaire que le processus diffusif (Figure 3.2B.), notamment pour les transitions entre les clusters.

¹Le test de corrélation de Pearson examine la relation entre deux variables (Pearson, 1900). Le coefficient de corrélation est la variable r qui en fonction de son signe indique si la corrélation est positive ou négative.

A.



B.

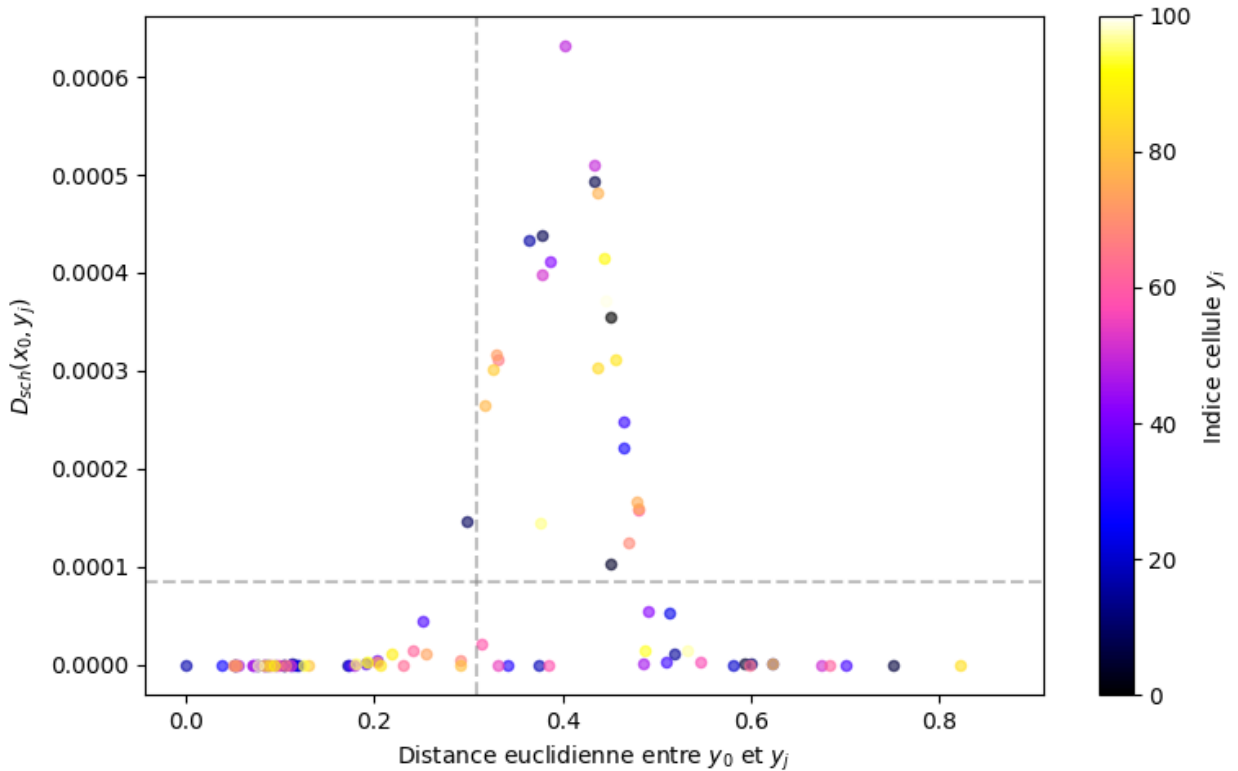


Figure 3.2. Distribution des cellules de ν en fonction de la valeur du couplage B_{sch} (A.) ou de la valeur du couplage D_{sch} (B.), et de leur distance euclidienne avec y_0 (le vrai couplage). Chaque couleur correspond à l'indice de la cellule y_i couplée à x_0 . Les lignes pointillées représentent la moyenne de l'axe correspondant. Les moyennes pour la valeur du couplage B_{sch} est 9.5×10^{-5} et pour la valeur du couplage D_{sch} est 8.5×10^{-5} et la moyenne de la distance euclidienne avec y_0 est 0.31.

3.4.2 Choix des conditions initiales lors de la construction du modèle Bursty et réduction du modèle

Le modèle Bursty décrit conjointement les quantités d'ARNm et de protéines. On a démontré, grâce aux simulations, qu'inférer des trajectoires d'ARN ou de protéines par diffusion était moins efficace qu'avec l'inférence basée sur le modèle Bursty. Or, la technologie du scRNA-seq donne des informations sur les quantités d'ARNm. Je discute de l'impact du choix des conditions initiales lors de la construction du modèle Bursty et propose une perspective pour l'améliorer en réduisant le modèle Bursty.

Si on ne se concentre que sur les données protéiques dans le modèle Bursty complet, dans la construction de B_{ref} , j'initialise les protéines par les valeurs de la distribution initiale et l'ARNm à 0. L'initialisation de l'ARNm à 0 n'a pas réellement d'impact sur la dynamique. En effet, quatre situations peuvent se présenter : soit les protéines activatrices sont déjà exprimées, auquel cas la probabilité du burst est élevée et perpétue la dynamique ; soit elles sont peu exprimées, et la probabilité de burst reste faible, mais la dynamique reste cohérente. De même, soit les protéines inhibitrices sont déjà exprimées et réduisent la probabilité de burst, soit elles ne le sont pas et la dynamique est préservée. Dans tous les cas, la dynamique est principalement pilotée par l'état protéique, et l'ARNm à 0 est une perturbation rapidement absorbée.

Cependant, la simulation des ARNm en l'absence de données de protéines soulève des questions, notamment quant au choix des conditions initiales du modèle Bursty lors de sa construction. En effet, lors du calcul du processus de référence B_{ref} pour les données simulées, on considère les coordonnées des cellules dans l'espace d'expression des gènes comme conditions initiales pour l'ARNm et on associe 0 aux protéines, valeur par défaut d'HARISSA (Herbach, 2021). Le choix de 0 pour les protéines a un effet négatif direct sur la dynamique de l'ARNm, car les taux de protéines interviennent directement dans le calcul du taux k_{on} . Une concentration de protéine nulle impact la régulation. Or, cette initialisation n'est pas la seule envisageable : on pourrait, par exemple, initialiser les protéines à une valeur conditionnelle de l'ARNm observé, ou encore les tirer aléatoirement. C'est en effet ce qu'on a fait pour les données expérimentales, on a associé des valeurs aléatoires à la condition initiale des protéines (selon le logarithme de la loi Gamma-Poisson en fonction des taux de burst minimal et maximal du gène (k_0 et k_1)) ; cette solution permet d'atténuer l'effet répressif artificiel des protéines. Cette solution reste toutefois une approximation. Le choix des conditions initiales lors de la construction du processus est un axe pour améliorer les résultats sur les données ARNm. La perspective envisagée est de réduire le modèle Bursty.

Dans la section 2.4 une version du modèle fondée uniquement sur les protéines a été proposée, mais la technologie du scRNA-seq donne des informations sur les quantités d'ARNm, il paraît naturel de se concentrer uniquement sur ces données. Cependant, il n'est pas possible

de réduire le modèle Bursty à un modèle ne prenant en compte que l'ARNm tout en préservant les propriétés mathématiques, car les ARNm sont généralement beaucoup moins stables que les protéines et ne permettraient donc pas de préserver la propriété markovienne du réseau de régulation génique. Leur demi-vie est typiquement beaucoup plus faible que celle des protéines, de l'ordre de 5 fois moins importante.

Face à cette contrainte, une piste consiste à apprendre conjointement la distribution des protéines et l'inférence des trajectoires des données expérimentales, en traitant les protéines comme des variables latentes à inférer à partir des données d'ARNm. C'est précisément l'approche proposée dans CardamomOT (Maugé & Ventre, 2026), une extension récente de CARDAMOM. Les auteurs ont développé un modèle qui reconstruit les trajectoires protéiques non observées en les considérant comme des variables latentes apprises simultanément à l'inférence faite avec le transport optimal. Cette formulation présente l'avantage de ne plus nécessiter de spécifier manuellement les conditions initiales des protéines : celles-ci sont déduites de manière cohérente à partir des données d'ARNm et de la structure du modèle dynamique. On ne peut pas adapter directement cette méthode avec la résolution du Problème de Schrödinger et le modèle Bursty. Mais, une autre possibilité, qui s'inspire de la logique de CardamomOT en conservant le travail fait sur le modèle Bursty, consiste à transporter une variable latente déduite de l'ARNm, plutôt que les protéines directement. L'idée serait de construire une variable latente à partir des données scRNA-seq, puis d'inférer les trajectoires de protéines à partir de cette variable latente, en utilisant le processus de référence présenté dans cette thèse. Cette approche peut soulever des questions sur l'interprétabilité des trajectoires des variables protéiques latentes ; mais elle permet de retirer la contrainte des conditions initiales des protéines et d'utiliser le modèle Bursty dans son intégralité, tout en conservant sa propriété markovienne. Cette approche pourrait constituer une direction d'extension de la méthode présentée dans ce travail.

3.4.3 Prise en compte de plusieurs intervalles de temps

Dans le Chapitre 2, on a résolu le problème de Schrödinger entre deux temps. La *méthode d'inférence basée sur le modèle Bursty* a été conçue pour ce cadre ; si on veut utiliser la méthode entre plusieurs temps, il faut réaliser l'inférence plusieurs fois, entre deux temps consécutifs. Serait-il envisageable de généraliser cette approche ? En effet les couplages trouvés deux à deux sont indépendants.

D'un point de vue théorique, le problème de Schrödinger est bien posé pour plusieurs marginales (≥ 2) et pour trouver sa solution, il existe un analogue multi-marginal de l'algorithme de Sinkhorn, qui fonctionne en balayant cycliquement les marginales (Chen et al., 2019) (Figure 1.6 et Algorithme 1.1). Ce cadre permet de chercher un seul processus de référence qui pourrait

s'appliquer à l'ensemble des intervalles de temps et résoudre le problème de manière globale. La difficulté réside dans la construction d'un processus de référence multi-marginal.

Dans le cadre à deux marginales, le processus de référence est construit avec le modèle Bursty à partir de la distribution μ en condition initiale. Il faudrait pouvoir implémenter, en plus, une condition "intermédiaire" : la condition à passer par des distributions intermédiaires, comme des "points d'étapes" obligatoires. Aujourd'hui, HARISSA, qui est utilisé pour résoudre numériquement le modèle Bursty, ne le permet pas. On pourrait utiliser un processus de référence entre le temps initial et le temps final, et s'en servir pour approcher par interpolation les distributions intermédiaires.

Soient μ , ρ et ν , trois distributions marginales aux temps respectifs $t_1 < t_2 < t_3$. Dans la même logique que l'interpolation réalisée entre deux distributions, en simulant le modèle Bursty avec la distribution μ en condition initiale, je peux extraire plusieurs nouvelles distributions : $\tilde{\rho}$, à t_2 , et $\tilde{\nu}$, à t_3 , et les comparer aux distributions ρ et ν . L'avantage de cette méthode est que je réduis le temps computationnel car je simule une fois le processus pour trois temps au lieu de deux. Mais la limitation de cette méthode est peut être que je dois augmenter le nombre de simulations pour trouver des trajectoires qui respectent les trois conditions : initiale, intermédiaire et finale.

3.4.4 Prise en compte de la mort et la prolifération

La méthode introduite dans ce travail omet de prendre en compte explicitement la mort et la prolifération cellulaires. Pourtant, elle ne nécessite pas d'avoir le même nombre de cellules dans les distributions initiale et finale, ce qui permet déjà de traiter partiellement ce cas. En effet, quel que soit le nombre de cellules impliquées, la méthode d'inférence peut coupler une cellule de la distribution de départ à plusieurs cellules de la distribution d'arrivée, ce qui peut être interprété comme une unique cellule à t_1 se divisant et générant des cellules clonales à t_2 . On pourrait associer ce phénomène à la prolifération. De plus, certaines cellules peuvent ne pas être couplées à des cellules à l'arrivée, ce qui peut être interprété comme une forme de mort cellulaire entre les deux temps d'observation. Ce phénomène est toutefois rare, car les couplages sont obtenus en utilisant l'algorithme de Sinkhorn qui a pour propriété d'empêcher ce type de situation (Algorithme 1.1).

Cela étant dit, il convient de souligner une limitation venant du type de données utilisées pour réaliser l'inférence : la technologie ScRNA-seq détruit les cellules utilisées pour réaliser les comptages. Il ne s'agit donc pas de reconstituer avec précision des lignées cellulaires individuelles, mais plutôt d'inférer des trajectoires populationnelles à partir de snapshots.

D'autres expériences moins agressives pour les cellules, comme la cytométrie en flux (section 1.3), permettent de suivre les mêmes cellules au cours du temps sans les détruire, et donc

d'accéder directement aux lignées cellulaires. Une nouvelle technologie (présentée dans la section 3.4.5) ou des jeux de données plus riches (données ARN et données sur les lignées cellulaires comme moslin (Lange et al., 2024)) peuvent permettre d'inclure de l'information sur la dynamique cellulaire.

Concrètement une extension simple pourrait être de modifier le poids des cellules dans l'algorithme de Sinkhorn 1.1, i.e. les valeurs des vecteurs u et v . Si on augmente le poids d'une cellule de la distribution de départ elle a plus de chance de se coupler avec plusieurs cellules à l'arrivée, comme si elle se divisait. Et si on attribue une très faible valeur à une cellule de la distribution de départ, elle a peu de chances de se coupler, comme si elle mourait. Cette extension reste néanmoins limitée : elle ne permet pas d'incorporer directement une information sur les lignées cellulaires individuelles.

3.4.5 Raman2RNA

Une dernière piste pour améliorer la validation de la *méthode d'inférence basée sur le modèle Bursty* consiste à comparer les couplages obtenus avec des données d'ARNm temporelles issues de la technologie Raman2RNA (Kobayashi-Kirschvink et al., 2024). Raman2RNA est une nouvelle méthode permettant de mesurer l'ARNm en continu, sans lyse des cellules. Raman2RNA mesure, par microscopie Raman hyperspectrale, une cellule unique afin d'obtenir un spectre. La microscopie Raman mesure les niveaux d'énergie vibrationnels des molécules (protéines, acides nucléiques et métabolites). Les auteurs entraînent un réseau de neurones afin d'établir un lien entre le spectre RAMAN et les données ScRNA-seq. Le principe consiste à déduire les informations transcriptomiques à partir du spectre RAMAN. Pour entraîner le réseau, les auteurs mesurent d'abord le spectre Raman, puis lysent la cellule *in vitro* afin d'obtenir les données scRNA-seq correspondantes. Ainsi, on peut suivre les cellules par spectroscopie RAMAN en fonction du temps, sans destruction des cellules ni contraintes sur le nombre d'observations dans le temps, et en déduire l'évolution des transcriptomes au cours du temps. Kobayashi-Kirschvink et al. (2024) appliquent leur méthode à des données de fibroblastes de souris ; la technologie étant récente, on peut espérer que les applications s'étendent à d'autres données. La limitation est que les données proviennent uniquement de cellules observées en 2D et *in vitro*. Cette technologie supprime les contraintes des pas de temps et offre une évolution temporelle continue. Ces données constituent une validation plus réaliste que les données simulées présentées dans la sous-section 1.8.3.

3.4.6 Intégration de l'aspect spatial

La méthode présentée dans le Chapitre 2 omet l'aspect spatial de la différenciation. Or, la différenciation cellulaire est induite par des facteurs extrinsèques, comme la communication en-

tre les cellules, les signaux chimiques et mécaniques de l'environnement (Engler et al., 2006; Morrison & Spradling, 2008). Une perspective d'amélioration est de s'inspirer des travaux de Merkt et al. (Merkt et al., 2025) : Waddington spatial, pour modéliser les facteurs extrinsèques de la différenciation. Waddington spatial modélise le processus de différenciation en intégrant explicitement la communication intercellulaire en couplant des EDO et EDP pour simuler les interactions entre la densité cellulaire et la concentration de ligands. Pour chaque cellule, les auteurs couplent des équations : une équation indique l'état de différenciation dans le paysage de Waddington, et d'autres équations modélisent la dynamique des ligands, considérés comme les outils de communication cellulaire. Deux types de données de cellules dendritiques sont utilisés, issus de Shalek et al. (2014) : "in-tube" où les cellules communiquent et "on-chip" où les cellules sont isolées physiquement et ne communiquent pas. Pour chaque jeu de données, des cellules sont étudiées au fur et à mesure avec le ScRNA-seq et les signaux sont étudiés avec les données "in-tube". Les auteurs confirment que les interactions façonnent les trajectoires de différenciation. Pour adapter notre méthode, une possibilité serait de coupler le modèle Bursty avec la concentration de ligands. J'intégrerais l'aspect spatial en modélisant la concentration de ligands, $l(t, s)$ qui dépend du temps t et de l'espace s en s'inspirant des travaux de Merkt et al. (2025).

Et j'incorpore la concentration de ligands dans le taux k_{on} du modèle Bursty (1.4) de la façon suivante :

$$k_{on,i}(P_1(t), \dots, P_G(t), l(t, s)) = \frac{k_{0,i} + k_{1,i} \exp(\beta_i + \sum_{j=1}^G \theta_{ji} P_j(t) + \phi_i l(t, s))}{1 + \exp(\beta_i + \sum_{j=1}^G \theta_{ji} P_j(t) + \phi_i l(t, s))},$$

ϕ_i est une constante qui représente l'impact du ligand sur le gène i . Si le gène est inducteur de différenciation, ϕ est positif, si le gène est répresseur de différenciation, ϕ est négatif ou si le gène n'a pas ou peu d'impact, ϕ est proche de 0. L'incorporation de la concentration de ligands dans le coefficient k_{on} permettrait que le modèle Bursty soit dépendant du temps et de l'espace. Si je choisis d'adapter la méthode avec cette extension, il faudrait étudier si le problème est bien posé, pour cela, je pourrais m'inspirer des travaux de Austin (2008) et Buckwar et Riedler (2011). Cette extension suppose que la concentration des ligands ne dépende pas du niveau de différenciation des cellules, il faudrait alors dans un second temps l'intégrer. Pour cela, on pourrait considérer que la sécrétion des ligands serait déclenchée par un noeud du réseau de gènes.

Bibliographie

- ADIMY, M., CHEKROUN, A., & TOUAOULA, T.-M. (2015). Age-structured and delay differential-difference model of hematopoietic stem cell dynamics. *Discrete and Continuous Dynamical Systems - B*, 20(9), 2765-2791 (cité p. 14).
- ALBAYRAK, C., JORDI, C. A., ZECHNER, C., LIN, J., BICHSEL, C. A., KHAMMASH, M., & TAY, S. (2016). Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Molecular Cell*, 61(6), 914-924 (cités pp. 24-26, 41).
- ALWINE, J. C., KEMP, D. J., & STARK, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences*, 74(12), 5350-5354 (cité p. 11).
- ANSCOMBE, F. J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika*, 35(3/4), 246-254 (cité p. 78).
- ARMOND, J. W., SAHA, K., RANA, A. A., OATES, C. J., JAENISCH, R., NICODEMI, M., & MUKHERJEE, S. (2014). A stochastic model dissects cell states in biological transition processes. *Sci. Rep.*, 4(1), 3692 (cité p. 14).
- ARNOL'D, V. I., AFRAIMOVICH, V., IL'YASHENKO, Y., & SHILNIKOV, L. (1989). *Dynamical systems V : bifurcation theory and catastrophe theory*. (Cité p. 17).
- AUSTIN, T. D. (2008). The emergence of the deterministic Hodgkin–Huxley equations as a limit from the underlying stochastic ion-channel mechanism. *The Annals of Applied Probability*, 18(4), 1279-1325 (cité p. 88).
- BALÁZSI, G., van OUDENAARDEN, G., & COLLINS, J. J. (2011). Cellular Decision Making and Biological Noise : From Microbes to Mammals. *Cell*, 144(6), 910-925 (cité p. 10).
- BARADAT, A., & VENTRE, E. (2024). Convergence of the Sinkhorn algorithm when the Schrödinger problem has no solution. *Annales de la Faculté des sciences de Toulouse : Mathématiques, Ser. 6*, 33(5), 1297-1371 (cités pp. 23, 55).
- BARRESI, M. J. F., & GILBERT, S. F. (2019). *Developmental Biology* (12^e éd.). Sinauer Associates, imprint of Oxford University Press. (Cité p. 6).
- BAWA, G., LIU, Z., YU, X., QIN, A., & SUN, X. (2022). Single-cell RNA sequencing for plant research : Insights and possible benefits. *Int. J. Mol. Sci.*, 23(9), 4497 (cité p. 13).
- BECSKEI, A., KAUFMANN, B. B., & van OUDENAARDEN, A. (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.*, 37(9), 937-944 (cité p. 25).
- BECSKEI, A., & SERRANO, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, 405(6786), 590-593 (cité p. 10).

- BÉLAIR, J., MACKEY, M. C., & MAHAFFY, J. M. (1995). Age-structured and two-delay models for erythropoiesis. *Mathematical Biosciences*, 128(1), 317-346 (cité p. 14).
- BENAMOU, J., & BRENIER, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math*, 84 (cité p. 19).
- BENDALL, S. C., DAVIS, K. L., DAVID AMIR, E., TADMOR, M. D., SIMONDS, E. F., CHEN, T. J., SHENFELD, D. K., NOLAN, G. P., & PE'ER, D. (2014). Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell*, 157(3), 714-725 (cité p. 18).
- BERGEN, V., LANGE, M., PEIDL, S., WOLF, F. A., & THEIS, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.*, 38(12), 1408-1414 (cité p. 19).
- BERNARD, S., CRAUSTE, F., GANDRILLON, O., KNIBBE, C., & PARSONS, D. (2024). *Simuscale : A Modular Framework for Multiscale Single-Cell Modelling* (rapp. tech. N° RT-0520). Inria Lyon. (Cité p. 15).
- BESSONNARD, S., DE MOT, L., GONZE, D., BARRIOL, M., DENNIS, C., GOLDBETER, A., DUPONT, G., & CHAZAUD, C. (2014). Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development*, 141(19), 3637-3648 (cité p. 16).
- BESSONOV, N., PUJO-MENJOUET, L., & VOLPERT, V. (2006). Cell Modelling of Hematopoiesis. *Math. Model. Nat. Phenom.*, 1(2), 81-103 (cité p. 15).
- BINTU, L., YONG, J., ANTEBI, K., Y. E. and. McCue, KAZUKI, Y., UNO, N., OSHIMURA, M., & ELOWITZ, M. B. (2016). Dynamics of epigenetic regulation at the single-cell level. *Science*, 351(6274), 720-724 (cité p. 9).
- BOETTIGER, A. N. (2013). Analytic Approaches to Stochastic Gene Expression in Multicellular Systems. *Biophysical Journal*, 105(12), 2629-2640 (cité p. 24).
- BOLAND, M. J., NAZOR, K. L., & LORING, J. F. (2014). Epigenetic regulation of pluripotency and differentiation. *Circ. Res.*, 115(2), 311-324 (cité p. 6).
- BONNER, W., HULETT, H., SWEET, R., & HERZENBERG, L. (1972). Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3), 404-409 (cités pp. 7, 11).
- BRUNS, I., LUCAS, D., PINHO, S., AHMED, J., LAMBERT, M. P., KUNISAKI, Y., SCHEIERMANN, C., SCHIFF, L., PONCZ, M., BERGMAN, A., & FRENETTE, P. S. (2014). Megakaryocytes regulate hematopoietic stem cell quiescence through CXCL4 secretion. *Nature medicine*, 20(11), 1315-1320 (cité p. 8).
- BUCKWAR, E., & RIEDLER, M. (2011). An exact stochastic hybrid model of excitable membranes including spatio-temporal evolution. *Journal of mathematical biology*, 63, 1051-93 (cité p. 88).
- BUNNE, C., SCHIEBINGER, G., KRAUSE, A., REGEV, A., & CUTURI, M. (2024). Optimal transport for single-cell and spatial omics. *Nature Reviews Methods Primers*, 4(1) (cités pp. 20, 39).
- BUNNE, C., STARK, S. G., GUT, G., DEL CASTILLO, M., J. S. and Levesque, LEHMANN, K.-V., PELKMANS, L., KRAUSE, A., & RÄTSCH, G. (2023). Learning single-cell perturbation responses using neural optimal transport. *Nat. Methods*, 20(11), 1759-1768 (cités pp. 21, 78).

- ÇAĞATAY, T., TURCOTTE, M., ELOWITZ, M. B., GARCIA-OJALVO, J., & SÜEL, G. (2009). Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139(3), 512-522 (cité p. 10).
- CATLIN, S. N., ABKOWITZ, J. L., & GUTTORP, P. (2001). Statistical inference in a two-compartment model for hematopoiesis. *Biometrics*, 57(2), 546-553 (cité p. 15).
- CHALANCON, G., RAVARANI, C. N., BALAJI, S., MARTINEZ-ARIAS, A., ARAVIND, L., JOTHI, R., & BABU, M. M. (2012). Interplay between gene expression noise and regulatory network architecture. *Trends Genet*, 28, 221-232 (cités pp. 10, 40).
- CHANG, H. H., HEMBERG, M., BARAHONA, M., INGBER, D. E., & HUANG, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194), 544-547 (cité p. 16).
- CHEN, S., & MAR, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1), 232 (cité p. 26).
- CHEN, Y., CONFORTI, G., GEORGIOU, T. T., & RIPANI, L. (2019). Multi-marginal Schrödinger Bridges, 725-732 (cité p. 85).
- CHIZAT, L., PEYRÉ, G., SCHMITZER, B., & VIALARD, F.-X. (2018). Unbalanced optimal transport : Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11), 3090-3123 (cité p. 20).
- CHONG, S., CHEN, C., GE, H., & XIE, X. S. (2014). Mechanism of Transcriptional Bursting in Bacteria. *Cell*, 158(2), 314-326 (cité p. 25).
- CHRISTIAN, L. (2013). A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint* (cités pp. 24, 63).
- CIVIN, C. I., STRAUSS, L. C., BROVALL, C., FACKLER, M., SCHWARTZ, J., & SHAPER, J. (1984). Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells. *The Journal of Immunology*, 133(1), 157-165 (cité p. 7).
- CLEVERS, H., RAFELSKI, S., ELOWITZ, M. B., KLEIN, A., SHENDURE, A., TRAPNELL, C., LEIN, E., LUNDBERG, E., UHLEN, M., MARTINEZ ARIAS, A., SANES, J. R., BLAINEY, P. C., EBERWINE, J., KIM, J., & LOVEN, J. C. (2017). What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism? *Cell Systems*, 4, 255-259 (cité p. 38).
- COULON, A., CHOW, C. C., SINGER, R. H., & LARSON, D. R. (2013). Eukaryotic transcriptional dynamics : from single molecules to cell populations. *Nat Rev Genet*, 14, 572-584 (cités pp. 25, 40).
- COVER, T. M., & THOMAS, J. A. (2005). Elements of Information Theory. Wiley. (Cité p. 33).
- CRAUSTE, F. (2009). Delay Model of Hematopoietic Stem Cell Dynamics : Asymptotic Stability and Stability Switch. *Math. Model. Nat. Phenom.*, 4(2), 28-47 (cité p. 14).
- CREMER, T., CREMER, M., HÜBNER, B., STRICKFADEN, H., SMEETS, D., POPKEN, J., STERR, M., MARKAKI, Y., RIPPE, K., & CREMER, C. (2015). The 4D nucleome : Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. *FEBS Lett.*, 589, 2931-2943 (cité p. 9).

- CRUDU, A., DEBUSSCHE, A., MULLER, A., & RADULESCU, O. (2012). Convergence of stochastic gene networks to hybrid piecewise deterministic processes. *Ann. Appl. Probab.*, 22(5), 1822-1859 (cit   p. 25).
- CUTURI, M. (2013). Sinkhorn Distances : Lightspeed Computation of Optimal Transportation Distances (cit   p. 19).
- DATTANI, J., & BARAHONA, M. (2017). Stochastic models of gene transcription with upstream drives : exact solution and sample path characterization. *Journal of the Royal Society, Interface*, 14(126) (cit   p. 24).
- DAVIDSON, E. H., RAST, J. P., OLIVERI, P., RANSICK, A., CALESTANI, C., YUH, C.-H., MINOKAWA, T., AMORE, G., HINMAN, V., ARENAS-MENA, C., OTIM, O., BROWN, C. T., LIVI, C. B., LEE, P. Y., REVILLA, R., RUST, A. G., PAN, Z., SCHILSTRA, M. J., CLARKE, P. J. C., ... BOLOURI, H. (2002). A Genomic Regulatory Network for Development. *Science*, 295(5560), 1669-1678 (cit   p. 26).
- DAVIS, M. H. A. (1984). Piecewise-Deterministic Markov Processes : A General Class of Non-Diffusion Stochastic Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3), 353-388 (cit   p. 25).
- DE ROSA, S. C., HERZENBERG, L. A., HERZENBERG, L. A., & ROEDERER, M. (2001). 11-color, 13-parameter flow cytometry : identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nature medicine*, 7(2), 245-248 (cit     pp. 7, 12).
- DIBAEINIA, P., & SINH, S. (2020). SERGIO : A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *Cell Systems*, 11(3), 252-271.e11 (cit   p. 26).
- D'INVERNO, M., & SAUNDERS, R. (2004). Agent-based modelling of stem cell self-organisation in a niche. *International Workshop on Engineering Self-Organising Applications*, 52-68 (cit   p. 15).
- DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M., & GINGERAS, T. R. (2013). STAR : ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21 (cit   p. 12).
- DOUMIC, M., MARCINIAK-CZOCRA, A., PERTHAME, B., & ZUBELLI, J. (2011). A Structured Population Model of Cell Differentiation. *SIAM Journal on Applied Mathematics*, 71(6), 1918-1940 (cit   p. 14).
- DRIESCH, H. (1891). Entwicklungsmechanische Studien. I. Der Wert der beiden ersten Furchungszellen in der Echinodermenentwicklung. *Zeitschrift f  r wissenschaftliche Zoologie*, 53, 160-178 (cit   p. 6).
- DU, J., JIA, P., DAI, Y., TAO, C., ZHAO, Z., & ZHI, D. (2019). Gene2vec : distributed representation of genes based on co-expression. *BMC Genomics*, 20(S1) (cit     pp. 20, 40).
- DUNN, S.-J., MARTELLO, G., YORDANOV, B., EMMOTT, S., & SMITH, A. (2014). Defining an essential transcription factor program for naive pluripotency. *Science*, 344(6188), 1156-1160 (cit   p. 16).
- DUPUIS, X. (2014). Optimal Control of Leukemic Cell Population Dynamics. *Mathematical Modelling of Natural Phenomena*, 9(1), 4-26 (cit   p. 14).
- EHRLICH, P. (1879). *Beitr  ge zur Kenntniss der granulirten Bindegewebszellen und der eosinophilen Leukocyten.* (Cit   p. 7).

- ELDAR, A., & ELOWITZ, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312), 167-73 (cités pp. 10, 40).
- ENGLER, A. J., SEN, S., SWEENEY, H. L., & DISCHER, D. E. (2006). Matrix Elasticity Directs Stem Cell Lineage Specification. *Cell*, 126(4), 677-689 (cités pp. 4, 88).
- EVANS, M. J., & KAUFMAN, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292, 154-156 (cités pp. 4, 6).
- EYMARD, N., BESSONOV, N., GANDRILLON, O., KOURY, M. J., & VOLPERT, V. (2014). The role of spatial organization of cells in erythropoiesis. *Journal of Mathematical Biology*, 70, 71-97 (cité p. 14).
- FARD, A. T., SRIHARI, S., MAR, J. C., & RAGAN, M. A. (2016). Not just a colourful metaphor : modelling the landscape of cellular development using Hopfield networks. *NPJ Syst. Biol. Appl.*, 2(1), 16001 (cité p. 17).
- FISCHER, S., KURBATOVA, P., BESSONOV, N., GANDRILLON, V., O. and. Volpert, & CRAUSTE, F. (2012). Modeling erythroblastic islands : Using a hybrid model to assess the function of central macrophage. *Journal of Theoretical Biology*, 298, 92-106 (cité p. 16).
- FOUREL, G., MAGDINIER, F., & GILSON, E. (2004). Insulator dynamics and the setting of chromatin domains. *Bioessays*, 26(5), 523-532 (cité p. 10).
- FOURNIÉ, C., VENTRE, E., HERBACH, U., BARADAT, A., GANDRILLON, O., & CRAUSTE, F. (2026, mai). *Cell Trajectory Inference based on Schrödinger Problem and a Mechanistic Model of Stochastic Gene Expression*. (Cités pp. 30, 36).
- FRIEDMAN, N., CAI, L., & XIE, X. S. (2006). Linking Stochastic Dynamics to Population Distribution : An Analytical Framework of Gene Expression. *Phys. Rev. Lett.*, 97, 168302 (cités pp. 10, 26, 58).
- FURUSAWA, C., & KANEKO, K. (2012). A dynamical-systems view of stem cell biology. *Science*, 338(6104), 215-217 (cité p. 16).
- GAILLARD, M., & HERBACH, U. (2025). Efficient Stochastic Simulation of Gene Regulatory Networks Using Hybrid Models of Transcriptional Bursting. *Computational Methods in Systems Biology*, 109-125 (cités pp. 27, 32, 55-57).
- GERIKE, T. G., PAULUS, U., POTTEN, C. S., & LOEFFLER, M. (1998). A dynamic model of proliferation and differentiation in the intestinal crypt based on a hypothetical intraepithelial growth factor. *Cell Prolif.*, 31(2), 93-110 (cité p. 5).
- GERMANO, D. P., & OSBORNE, J. M. (2021). A mathematical model of cell fate selection on a dynamic tissue. *Journal of Theoretical Biology*, 514, 110535 (cité p. 5).
- GIRARD, O., MOINARD, E., DAVID, L., & FRÉOUR, T. (2026). The contribution of single-cell omics technologies to improve knowledge on human early embryo development. *Human Reproduction*, 41(1), 6-15 (cités pp. 6, 8, 16).

- GLAUCHE, I., CROSS, M., LOEFFLER, M., & ROEDER, I. (2007). Lineage specification of hematopoietic stem cells : mathematical modeling and biological implications. *Stem cells*, 25(7), 1791-1799 (cité p. 15).
- GLEN, C. M., KEMP, M. L., & VOIT, E. O. (2019). Agent-based modeling of morphogenetic systems : Advantages and challenges. *PLOS Computational Biology*, 15(3), 1-31 (cité p. 15).
- GÖKBUĞET, D., & BLELLOCH, R. (2019). Epigenetic control of transcriptional regulation in pluripotency and early differentiation. *Development*, 146(19), dev164772 (cité p. 6).
- GOODWIN, J., LASLETT, A. L., & RUGG-GUNN, P. J. (2020). The application of cell surface markers to demarcate distinct human pluripotent states. *Experimental Cell Research*, 387(1), 111749 (cité p. 12).
- GRAF, T., & ENVER, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273), 587-594 (cité p. 4).
- GURDON, J. B. (1962). The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Journal of Embryology and Experimental Morphology*, 10(3), 622-640 (cité p. 6).
- HADDAD, N., JOST, D., & VAILLANT, C. (2017). Perspectives : using polymer modeling to understand the formation and function of nuclear compartments. *Chromosome Res.*, 25(1), 35-50 (cité p. 10).
- HAGHVERDI, L., BÜTTNER, M., WOLF, F. A., BUETTNER, F., & THEIS, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods*, 13(10), 845-848 (cités pp. 16, 19).
- HAMMING, R. W. (1950). Error detecting and error correcting codes. *Bell Syst. Tech. J.*, 29(2), 147-160 (cité p. 21).
- HARRISON, P. F. (2017). Varistran : Anscombe's variance stabilizing transformation for RNA-seq gene expression data. *J. Open Source Softw.*, 2(16), 257 (cité p. 78).
- HASHIMOTO, T., GIFFORD, D., & JAAKKOLA, T. (2016). Learning Population-Level Diffusions with Generative RNNs. *Proceedings of The 33rd International Conference on Machine Learning*, 48, 2417-2426 (cité p. 18).
- HEID, C. A., STEVENS, J., LIVAK, K. J., & WILLIAMS, P. M. (1996). Real time quantitative PCR. *Genome Res.*, 6(10), 986-994 (cité p. 11).
- HERBACH, U. (2021). Gene regulatory network inference from single-cell data using a self-consistent proteomic field. *arXiv preprint* (cités pp. 26, 27, 32, 56, 84).
- HERBACH, U. (2023). Harissa : Stochastic Simulation and Inference of Gene Regulatory Networks Based on Transcriptional Bursting. *Computational Methods in Systems Biology*, 97-105 (cités pp. 56, 57, 60).
- HERBACH, U., BONNAFFOUX, A., ESPINASSE, T., & GANDRILLON, O. (2017). Inferring gene regulatory networks from single-cell data : a mechanistic approach. *BMC Syst Biol*, 11(1) (cités pp. 25, 26, 32, 41, 55, 56).
- HERMANGE, G., COURNÈDE, P.-H., & PLO, I. (2022). Modélisation mathématique de l'hématopoïèse et des hé-mopathies : développement, dynamique et traitement. *Hématologie*, 28(4), 183-200 (cité p. 13).

- HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554-2558 (cité p. 17).
- HUANG, S. (2009). Non-genetic heterogeneity of cells in development : More than just noise. *Development*, 136, 3853-62 (cité p. 10).
- HUANG, S. (2010). Cell Lineage Determination in State Space : A Systems View Brings Flexibility to Dogmatic Canonical Rules. *PLOS Biol*, 8(5), e1000380 (cités pp. 17, 20, 33, 40, 54).
- HUANG, X.-T., LI, X., QIN, P.-Z., ZHU, Y., XU, S.-N., & CHEN, J.-P. (2018). Technical advances in single-cell RNA sequencing and applications in normal and malignant hematopoiesis. *Front. Oncol.*, 8, 582 (cité p. 12).
- HUANG, Z., LUO, S., WANG, Z., ZHANG, Z., JIANG, B., NIE, Q., & ZHANG, J. (2026). Deep learning linking mechanistic models to single-cell transcriptomics data reveals transcriptional bursting in response to DNA damage (cité p. 55).
- HUGUET, G., MAGRUDER, D. S., TONG, A., FASINA, O., KUCHROO, M., WOLF, G., & KRISHNASWAMY, S. (2022). Manifold Interpolating Optimal-Transport Flows for Trajectory Inference. *Advances in Neural Information Processing Systems*, 35, 29705-29718 (cité p. 78).
- HUGUET, G., TONG, A., ZAPATERO, M. R., TAPE, C. J., WOLF, G., & KRISHNASWAMY, S. (2023). Geodesic Sinkhorn For Fast and Accurate Optimal Transport on Manifolds. *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, 1-6 (cité p. 21).
- HUH, D., & PAULSSON, J. (2011). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature genetics*, 43(2), 95-100 (cité p. 10).
- HULETT, H. R., BONNER, W. A., BARRETT, J., & HERZENBERG, L. A. (1969). Cell sorting : automated separation of mammalian cells as a function of intracellular fluorescence. *Science*, 166(3906), 747-749 (cités pp. 7, 11).
- IKUTA, K., & WEISSMAN, I. L. (1992). Evidence that hematopoietic stem cells express mouse c-kit but do not depend on steel factor for their generation. *Proceedings of the National Academy of Sciences*, 89(4), 1502-1506 (cité p. 12).
- IWASA, Y. (2023). Mathematical modeling for developmental processes. *Development, Growth & Differentiation*, 65(5), 272-285 (cité p. 13).
- Ji, Z., & Ji, H. (2016). TSCAN : Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, 44(13), e117 (cités pp. 18, 19).
- JIANG, Q., ZHANG, S., & WAN, L. (2022). Dynamic inference of cell developmental complex energy landscape from time series single-cell transcriptomic data. *PLOS Computational Biology*, 18(1), 1-22 (cités pp. 20, 39).
- JOVIC, D., LIANG, X., ZENG, H., LIN, L., XU, F., & LUO, Y. (2022). Single-cell RNA sequencing technologies and applications : A brief overview. *Clin. Transl. Med.*, 12(3), e694 (cité p. 12).

- KAERN, M., ELSTON, T. C., BLAKE, W. J., & COLLINS, J. J. (2005). Stochasticity in gene expression : from theories to phenotypes. *Nat. Rev. Genet.*, 6(6), 451-464 (cités pp. 10, 14).
- KANTOROVICH, L. (1942). On the Translocation of Masses. *Comptes Rendus (Doklady) de l'Academie des Sciences de l'URSS*, 37, 199-201 (cité p. 19).
- KAUFFMAN, S. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215), 177-178 (cités pp. 13, 26).
- KIEL, M. J., YILMAZ, Ö. H., IWASHITA, T., YILMAZ, O. H., TERHORST, C., & MORRISON, S. J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell*, 121(7), 1109-1121 (cité p. 12).
- KIM, J. K., & MARIONI, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol*, 14(1), R7 (cité p. 55).
- KIRK, J., ORR, J., & HOPE, C. (1968). A mathematical analysis of red blood cell and bone marrow stem cell control mechanisms. *British journal of haematology*, 15(1), 35-46 (cité p. 13).
- KLEIN, A. M., MAZUTIS, L., AKARTUNA, I., TALLAPRAGADA, N., VERES, A., LI, V., PESHKIN, L., WEITZ, D. A., & KIRSCHNER, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5), 1187-1201 (cités pp. 12, 38).
- KLEIN, D., PALLA, G., LANGE, M., KLEIN, M., PIRAN, Z., GANDER, M., MENG-PAPAXANTHOS, L., STERR, M., SABER, L., JING, C., BASTIDAS-PONCE, A., COTA, P., TARQUIS-MEDINA, M., PARIKH, S., GOLD, I., LICKERT, H., BAKHTI, M., NITZAN, M., CUTURI, M., & THEIS, F. J. (2025). Mapping cells through time and space with moscot. *Nature*, 638(8052), 1065-1075 (cité p. 20).
- Ko, M. (1991). A stochastic model for gene induction. *Journal of Theoretical Biology*, 153, 181-94 (cités pp. 24, 40).
- KOBAYASHI-KIRSCHVINK, K. J., COMITER, C. S., GADDAM, S., JOREN, T., GRODY, E. I., OUNADJELA, J. R., ZHANG, K., GE, B., KANG, J., XAVIER, R. J., So, P. T. C., BIANCALANI, T., SHU, J., & REGEV, A. (2024). Prediction of single-cell RNA expression profiles in live cells by Raman microscopy with Raman2RNA. *Nat. Biotechnol.*, 42(11), 1726-1734 (cités pp. 12, 87).
- KRASNYAKOV, I. (2024). Mathematical modeling of evolution of cell networks in epithelial tissues. *Quant. Biol.*, 12(3), 286-300 (cité p. 5).
- KULLBACK, S., & LEIBLER, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86 (cité p. 33).
- KUPIEC, J. J. (1997). A Darwinian theory for the origin of cellular differentiation. *Mol. Gen. Genet.*, 255(2), 201-208 (cité p. 10).
- KUPIEC, J.-J., GANDRILLON, O., MORANGE, M., & SILBERSTEIN, M. (2011). *Le hasard au cœur de la cellule*. Editions Matériologiques. (Cité p. 10).
- LA MANNO, G., SOLDATOV, R., ZEISEL, A., BRAUN, E., HOCHGERNER, H., PETUKHOV, V., LIDSCHREIBER, K., KASTRITI, M. E., LÖNNERBERG, P., FURLAN, A., FAN, J., BORM, L. E., LIU, Z., van BRUGGEN, D., GUO, J., HE, X.,

- BARKER, R., SUNDSTRÖM, E., CASTELO-BRANCO, G., ... KHARCHENKO, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719), 494-498 (cité p. 19).
- LACOMBE, F., MAYNADIÉ, M., CHATELAIN, B., & BÉNÉ, M.-C. (2017). Évolution des outils informatiques d'analyse des données en cytométrie en flux. *Revue Francophone des Laboratoires*, 2017(495), 57-61 (cités pp. 8, 12).
- LAFORGE, B., GUEZ, D., MARTINEZ, M., & KUPIEC, J.-J. (2005). Modeling embryogenesis and cancer : an approach based on an equilibrium between the autostabilization of stochastic gene expression and the interdependence of cells for proliferation. *Progress in Biophysics and Molecular Biology*, 89(1), 93-120 (cité p. 10).
- LAMBERT, S. A., JOLMA, A., CAMPITELLI, L. F., DAS, P. K., YIN, Y., ALBU, M., CHEN, X., TAIPALE, J., HUGHES, T. R., & WEIRAUCH, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4), 650-665 (cité p. 9).
- LANGE, M., BERGEN, V., KLEIN, M., SETTY, M., REUTER, B., BAKHTI, M., LICKERT, H., ANSARI, M., SCHNIERING, J., SCHILLER, H., PE'ER, D., & THEIS, F. (2022). CellRank for directed single-cell fate mapping. *Nat Methods*, 19, 159-170 (cité p. 52).
- LANGE, M., PIRAN, Z., KLEIN, M., SPANJAARD, B., KLEIN, D., JUNKER, J. P., THEIS, F. J., & NITZAN, M. (2024). Mapping lineage-traced cells across time points with moslin. *Genome Biology*, 25(1) (cités pp. 20, 39, 40, 87).
- LARSON, D. R. (2011). What do expression dynamics tell us about the mechanism of transcription? [Developmental mechanisms, patterning and evolution]. *Current Opinion in Genetics & Development*, 21(5), 591-599 (cité p. 25).
- LATCHMAN, D. S. (1996). Inhibitory transcription factors. *The International Journal of Biochemistry & Cell Biology*, 28(9), 965-974 (cité p. 9).
- LAURENTI, E., & GÖTTGENS, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689):418-426. (Cités pp. 8, 16).
- LAVENANT, H., ZHANG, S., KIM, Y.-H., & SCHIEBINGER, G. (2024). Towards a mathematical theory of trajectory inference. *The Annals of Applied Probability*, 34(1A), 428-500 (cités pp. 20, 39, 62).
- LIU, R., BALSUBRAMANI, A., & ZOU, J. (2020). Learning transport cost from subset correspondence (cités pp. 20, 21, 39, 40).
- LODISH, H., BERK, A., KAISER, C. A., KRIEGER, M., BRETSCHER, A., PLOEGH, H., AMON, A., MARTIN, K. C., YAFFE, M. B., MASSON, P. L., & SANLAVILLE, C. (2022). *Biologie moléculaire de la cellule* (5e édition). De Boeck supérieur. (Cité p. 8).
- LÖNNBERG, T., SVENSSON, V., JAMES, K. R., FERNANDEZ-RUIZ, D., SEBINA, I., MONTANDON, R., SOON, M. S. F., FOGG, L. G., NAIR, A. S., LILIGETO, U., STUBBINGTON, M. J. T., LY, L.-H., BAGGER, F. O., ZWIESSELE, M., LAWRENCE, N. D., SOUZA-FONSECA-GUIMARAES, F., BUNN, P. T., ENGWERDA, C. R., HEATH, W. R., ... TEICHMANN, S. A. (2017). Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Sci. Immunol.*, 2(9) (cités pp. 18, 19).
- LOTFOLLAHI, M., KLIMOVSKAIA SUSMELJ, A., DE DONNO, C., HETZEL, L., JI, Y., IBARRA, I. L., SRIVATSAN, S. R., NAGHIPOURFAR, M., DAZA, R. M., MARTIN, B., SHENDURE, J., MCFALINE-FIGUEROA, J. L., BOYEAU, P.,

- WOLF, F. A., YAKUBOVA, N., GÜNNEMANN, S., TRAPNELL, C., LOPEZ-PAZ, D., & THEIS, F. J. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), e11517 (cités pp. 20, 21, 39).
- LUMMERTZ DA ROCHA, E., ROWE, R. G., LUNDIN, V., MALLESIAIAH, M., JHA, D. K., RAMBO, C. R., LI, H., NORTH, T. E., COLLINS, J. J., & DALEY, G. Q. (2018). Reconstruction of complex single-cell trajectories using CellRouter. *Nat. Commun.*, 9(1), 892 (cités pp. 18, 19).
- LUO, S., WANG, Z., ZHANG, Z., ZHOU, T., & ZHANG, J. (2023a). Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics. *Nucleic Acids Res*, 51(1), 68-83 (cité p. 55).
- LUO, S., ZHANG, Z., WANG, Z., YANG, X., CHEN, X., ZHOU, T., & ZHANG, J. (2023b). Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model. *R Soc Open Sci*, 10(4), 221057 (cité p. 55).
- MACARTHUR, B. D., MA'AYAN, A., & LEMISCHKA, I. R. (2009). Systems biology of stem cell fate and cellular reprogramming. *Nature reviews Molecular cell biology*, 10(10), 672-681 (cité p. 16).
- MACARTHUR, B. D., PLEASE, C. P., & OREFFO, R. O. C. (2008). Stochasticity and the Molecular Mechanisms of Induced Pluripotency. *PLOS ONE*, 3(8), 1-11 (cité p. 16).
- MACKEY, M. C., & GLASS, L. (1977). Oscillation and Chaos in Physiological Control Systems. *Science*, 197(4300), 287-289 (cité p. 14).
- MACKEY, M. C., M., T.-K., & R., Y. (2011). Molecular distributions in gene regulatory dynamics. *Journal of Theoretical Biology*, 274(1), 84-96 (cité p. 26).
- MACOSKO, E. Z., BASU, A., SATIJA, R., NEMESH, J., SHEKHAR, K., GOLDMAN, M., TIROSH, I., BIALAS, A. R., KAMITAKI, N., MARTERSTECK, E. M., TROMBETTA, J. J., WEITZ, D. A., SANES, J. R., SHALEK, A. K., REGEV, A., & MCCARROLL, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5), 1202-1214 (cités pp. 12, 38).
- MAGAÑA-LÓPEZ, G., CALZONE, L., ZINOVYEV, A., & PAULEVÉ, L. (2024). scBoolSeq : Linking scRNA-seq statistics and Boolean dynamics. *PLOS Computational Biology*, 20(7), 1-25 (cité p. 40).
- MAINI, P., & BAKER, R. (2012). *Developmental Biology : Mathematical Modelling of Development* (cité p. 14).
- MAJIMA, K., KOJIMA, Y., MINOURA, K., ABE, K., HIROSE, H., & SHIMAMURA, T. (2024). LineageVAE : reconstructing historical cell states and transcriptomes toward unobserved progenitors. *Bioinformatics*, 40(10), btac520 (cités pp. 20, 40).
- MAR, J. C. (2019). The rise of the distributions : why non-normality is important for understanding the transcriptome and beyond. *Biophys. Rev.*, 11(1), 89-94 (cité p. 26).
- MARCINIAK-CZOCZRA, A., STIEHL, T., HO, W., A. and Jäger, & WAGNER, W. (2008). Modeling of Asymmetric Cell Division in Hematopoietic Stem Cells—Regulation of Self-Renewal Is Essential for Efficient Repopulation. *Stem cells and development*, 18, 377-85 (cité p. 14).

- MARIANI, M. (2018). A Gamma-convergence approach to large deviations. *arXiv preprint* (cités pp. 24, 63).
- MARTIN, G. R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12), 7634-7638 (cités pp. 4, 6).
- MAUGÉ, Y., & VENTRE, E. (2026). CardamomOT : a mechanistic optimal transport-based framework for gene regulatory network inference, trajectory reconstruction and generative modeling. *bioRxiv* (cité p. 85).
- MAXIMOW, A. (1909). Der Lymphozyt als gemeinsame Stammzelle der verschiedenen Blutelemente in der embryonalen Entwicklung und im postfetalen Leben der Säugetiere. *Folio Haematologica* (cité p. 7).
- MCADAMS, H. H., & ARKIN, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3), 814-819 (cité p. 10).
- MCCANN, R. J. (1997). A Convexity Principle for Interacting Gases. *Advances in Mathematics*, 128 (cité p. 19).
- MCCULLAGH, E., FARLOW, J., FULLER, C., GIRARD, J., LIPINSKI-KRUSZKA, J., LU, D., NORIEGA, T., ROLLINS, G., SPITZER, R., TODHUNTER, M., & EL-SAMAD, H. (2009). Not all quiet on the noise front. *Nat Chem Biol*, 5(10), 699-704 (cité p. 40).
- MERKT, S., FUHRMANN, L., DUDKIN, E., SCHLITZER, A., NIETHAMMER, B., & HASENAUER, J. (2025). A dynamic model for Waddington's landscape accounting for cell-to-cell communication. *Mathematical Biosciences*, 390, 109537 (cité p. 88).
- MONGE, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale. (Cité p. 19).
- MORRISON, S. J., SHAH, N. M., & ANDERSON, D. J. (1997). Regulatory Mechanisms in Stem Cell Biology. *Cell*, 88(3), 287-298 (cité p. 4).
- MORRISON, S. J., & SPRADLING, A. C. (2008). Stem cells and niches : mechanisms that promote stem cell maintenance throughout life. *Cell*, 132(4), 598-611 (cités pp. 4, 88).
- NGUYEN, T. N. T., MARTIN, M., ARPIN, C., BERNARD, S., GANDRILLON, O., & CRAUSTE, F. (2024). In silico modelling of CD8 T cell immune response links genetic regulation to population dynamics. *Immunoinformatics*, 15, 100043 (cité p. 15).
- NIEPEL, M., SPENCER, S. L., & SORGER, P. K. (2009). Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr Opin Chem Biol* (cité p. 40).
- NISSEN, S. B., PERERA, M., GONZALEZ, J. M., MORGANI, S. M., JENSEN, M. H., SNEPPEN, K., BRICKMAN, J. M., & TRUSINA, A. (2017). Four simple rules that are sufficient to generate the mammalian blastocyst. *PLoS biology*, 15(7), e2000737 (cité p. 15).
- NOVICK, A., & WEINER, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences*, 43(7), 553-566 (cité p. 10).

- OGURO, H., DING, L., & MORRISON, S. J. (2013). SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell stem cell*, 13(1), 102-116 (cité p. 12).
- OKADA, S., NAKAUCHI, H., NAGAYOSHI, K., NISHIKAWA, S.-I., MIURA, Y., & SUDA, T. (1992). In vivo and in vitro stem cell function of c-kit-and Sca-1-positive murine hematopoietic cells. *Blood*, 80(12), 3044-3050 (cité p. 12).
- ORKIN, S. H., & ZON, L. I. (2008). Hematopoiesis : an evolving paradigm for stem cell biology. *Cell*, 132(4), 631-644 (cité p. 8).
- OSAWA, M., HANADA, K.-i., HAMADA, H., & NAKAUCHI, H. (1996). Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science*, 273(5272), 242-245 (cité p. 12).
- PÁJARO, M., ALONSO, A. A., OTERO-MURAS, I., & VÁZQUEZ, C. (2017). Stochastic modeling and numerical simulation of gene regulatory networks with protein bursting. *Journal of Theoretical Biology*, 421, 51-70 (cité p. 26).
- PAPP, B., & PLATH, K. (2013). Epigenetics of reprogramming to induced pluripotency. *Cell*, 152(6), 1324-1343 (cité p. 8).
- PASSEGUÉ, E., WAGERS, A. J., GIURIATO, S., ANDERSON, W. C., & WEISSMAN, I. L. (2005). Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates. *The Journal of experimental medicine*, 202(11), 1599-1611 (cité p. 8).
- PAULSSON, J., BERG, O. G., & EHRENBERG, M. (2000). Stochastic focusing : fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97(13), 7148-7153 (cité p. 10).
- PEARSON, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175 (cité p. 82).
- PECCOUD, J., & YCART, B. (1995). Markovian Modelling of Gene Product Synthesis [p40]. *Theoretical population biology*, 48, 222-234 (cités pp. 24, 40).
- PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., & DUCHESNAY, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830 (cité p. 30).
- PERFETTO, S., CHATTOPADHYAY, P., & ROEDERER, M. (2004). Seventeen-colour flow cytometry : unravelling the immune system. *Nature Reviews Immunology*, 4(8), 648-655 (cités pp. 8, 12).
- PEYRÉ, G., & CUTURI, M. (2019a). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 355-607 (cités pp. 19, 39, 63).
- PEYRÉ, G., & CUTURI, M. (2019b). Computational Optimal Transport with Applications to Data Sciences. *Foundations and Trends in Machine Learning*, 11(5-6), 355-607 (cités pp. 28, 29).

- PIJUAN-SALA, B., GRIFFITHS, J. A., GUIBENTIF, C., HISCOCK, T. W., JAWAID, W., CALERO-NIETO, F. J., MULAS, C., IBARRA-SORIA, X., TYSER, R. C. V., HO, D. L. L., REIK, W., SRINIVAS, S., SIMONS, B. D., NICHOLS, J., MARIONI, J. C., & GÖTTGENS, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745), 490-495 (cité p. 6).
- PRASAD, N., YANG, K., & UHLER, C. (2022). Optimal Transport using GANs for Lineage Tracing. *arXiv preprint* (cités pp. 20, 21, 39).
- PUJADAS, E., & FEINBERG, A. P. (2012). Regulated Noise in the Epigenetic Landscape of Development and Disease. *Cell*, 148(6), 1123-1131 (cité p. 17).
- PUJO-MENJOUET, L. (2016). Blood Cell Dynamics : Half of a Century of Modelling. *Math. Model. Nat. Phenom.*, 11(1), 92-115 (cité p. 13).
- RAJ, A., PESKIN, C. S., TRANCHINA, D., VARGAS, D. Y., & TYAGI, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol*, 4(10), e309 (cités pp. 24, 25, 40).
- RAJ, A., & van OUDENAARDEN, A. (2008). Nature, nurture, or chance : stochastic gene expression and its consequences. *Cell*, 135(2), 216-26 (cité p. 40).
- RAMALHO-SANTOS, M., & WILLENBRING, H. (2007). On the origin of the term “stem cell”. *Cell stem cell*, 1(1), 35-38 (cité p. 7).
- RAO, S. S. P., HUNTLEY, M. H., DURAND, N. C., STAMENOVA, E. K., BOCHKOV, I. D., ROBINSON, J. T., SANBORN, A. L., MACHOL, I., OMER, A. D., LANDER, E. S., & AIDEN, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665-1680 (cité p. 10).
- RASER, J. M., & O'SHEA, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678), 1811-1814 (cité p. 25).
- RASER, J. M., & O'SHEA, E. K. (2005). Noise in gene expression : origins, consequences, and control. *Science*, 309(5743), 2010-2013 (cité p. 10).
- REIK, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143), 425-432 (cité p. 4).
- RICHARD, A., BOULLU, L., HERBACH, U., BONNAFOUX, A., MORIN, V., VALLIN, E., GUILLEMIN, A., PAPILI GAO, N., GUNAWAN, R., COSETTE, J., ARNAUD, O., KUPIEC, J.-J., ESPINASSE, T., GONIN-GIRAUD, S., & GANDRILLON, O. (2016). Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLOS Biology*, 14(12), 1-35 (cités pp. 11, 16, 25).
- RIPANI, L. (2017). *Le problème de Schrödinger et ses liens avec le transport optimal et les inégalités fonctionnelles* [thèse de doct., Université de Lyon]. (Cité p. 24).
- ROEDER, I., & GLAUCHE, I. (2006). Towards an understanding of lineage specification in hematopoietic stem cells : a mathematical model for the interaction of transcription factors GATA-1 and PU. 1. *Journal of theoretical biology*, 241(4), 852-865 (cité p. 15).

- ROEDER, I., HORN, K., SIEBURG, H.-B., CHO, R., MULLER-SIEBURG, C., & LOEFFLER, M. (2008). Characterization and quantification of clonal heterogeneity among hematopoietic stem cells : a model-based approach. *Blood, The Journal of the American Society of Hematology*, 112(13), 4874-4883 (cit   p. 15).
- ROEDER, I., KAMMINGA, L. M., BRAESEL, K., DONTJE, B., de HAAN, G., & LOEFFLER, M. (2005). Competitive clonal hematopoiesis in mouse chimeras explained by a stochastic model of stem cell organization. *Blood*, 105(2), 609-616 (cit   p. 15).
- ROSSANT, J., & TAM, P. P. L. (2017). New insights into early human development : lessons for stem cell derivation and differentiation. *Cell Stem Cell*, 20(1), 18-28 (cit   p. 8).
- SAELENS, W., CANNODT, R., TODOROV, H., & SAEYS, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, 37(5), 547-554 (cit   p. 18).
- S  EZ, M., BRISCOE, J., & RAND, D. A. (2022). Dynamical landscapes of cell fate decisions. *Interface Focus*, 12(4), 20220002 (cit   p. 17).
- SAMARAN, J., PEYR  , G., & CANTINI, L. (2026). CHAMPOLLION : Robust Multi-Omics Integration via Inverse Optimal Transport Using Paired Cells. *bioRxiv* (cit   p. 20).
- S  NCHEZ,   ., & KONDEV, J. (2008). Transcriptional control of noise in gene expression. *Proceedings of the National Academy of Sciences*, 105(13), 5081-5086 (cit   p. 24).
- SCHEMA, M., SHALON, D., DAVIS, R. W., & BROWN, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235), 467-470 (cit   p. 11).
- SCHIEBINGER, G., SHU, J., TABAKA, M., CLEARY, B., SUBRAMANIAN, V., SOLOMON, A., GOULD, J., LIU, S., LIN, S., BERUBE, P., LEE, L., CHEN, J., BRUMBAUGH, J., RIGOLLET, P., HOCHEDLINGER, K., JAENISCH, R., REGEV, A., & LANDER, E. S. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4), 928-943.e22 (cit  s pp. 19, 20, 35, 39, 52, 63, 66, 78, 79).
- SCHR  DINGER, E. (1932). Sur la th  orie relativiste de l'  lectron et l'interpr  tation de la m  canique quantique. *Annales de l'institut Henri Poincar  *, 2(4), 269-310 (cit  s pp. 21, 63).
- SEMRAU, S., GOLDMANN, J. E., SOUMILLON, M., MIKKELSEN, T. S., JAENISCH, R., & van OUDENAARDEN, A. (2017). Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.*, 8(1), 1096 (cit  s pp. 17, 28, 34, 35, 50, 52, 55, 65, 67, 77, 78).
- SHA, Y., QIU, Y., ZHOU, P., & NIE, Q. (2023). Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nature Machine Intelligence*, 6(1), 25-39 (cit  s pp. 20, 21, 39).
- SHALEK, A. K., SATIJA, R., SHUGA, J., TROMBETTA, J. J., GENNERT, D., LU, D., CHEN, P., GERTNER, R. S., GAUBLomme, J. T., YOSEF, N., SCHWARTZ, S., FOWLER, B., WEAVER, S., WANG, J., WANG, X., DING, R., RAYCHOWDHURY, R., FRIEDMAN, N., HACHOEN, N., ... REGEV, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505), 363-369 (cit   p. 88).
- SHAPIRO, H. M. (2003). *Practical flow cytometry*. John Wiley & Sons, Ltd. (Cit   p. 12).

- SHI, J., TESCHENDORFF, A. E., CHEN, W., CHEN, L., & LI, T. (2020). Quantifying Waddington's epigenetic landscape : a comparison of single-cell potency measures. *Briefings in Bioinformatics*, 21(1), 248-261 (cit   p. 17).
- SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 13(2), 238-241 (cit   p. 11).
- SINKHORN, R. (1967). Diagonal Equivalence to Matrices with Prescribed Row and Column Sums. *The American Mathematical Monthly*, 74(4), 402-405 (cit  s pp. 19, 21, 33, 39, 40, 63).
- SPANGRUDE, G. J., HEIMFELD, S., & WEISSMAN, I. L. (1988). Purification and characterization of mouse hematopoietic stem cells. *Science*, 241(4861), 58-62 (cit   p. 7).
- SPUDICH, J. L., & KOSHLAND JR, D. E. (1976). Non-genetic individuality : chance in the single cell. *Nature*, 262(5568), 467-471 (cit   p. 10).
- STREET, K., RISSO, D., FLETCHER, R. B., DAS, D., NGAI, J., YOSEF, N., PURDOM, E., & DUDOIT, S. (2018). Slingshot : cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1), 477 (cit  s pp. 18, 19).
- STUMPF, P. S., SMITH, R. C. G., LENZ, M., SCHUPPERT, A., M  LLER, F.-J., BABTIE, A., CHAN, T. E., STUMPF, M. P. H., PLEASE, C. P., HOWISON, S. D., ARAI, F., & MACARTHUR, B. D. (2017). Stem cell differentiation as a non-Markov stochastic process. *Cell Syst.*, 5(3), 268-282.e7 (cit   p. 15).
- SUTER, D. M., MOLINA, N., GATFIELD, D., SCHNEIDER, K., SCHIBLER, U., & NAEF, F. (2011). Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science*, 332(6028), 472-474 (cit  s pp. 24, 25, 40, 41).
- SVENSSON, V., VENTO-TORMO, R., & TEICHMANN, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, 13(4), 599-604 (cit   p. 12).
- TAJER, P., PIKE-OVERZET, K., ARIAS, S., HAVENGA, M., & STAAL, F. J. T. (2019). Ex Vivo Expansion of Hematopoietic Stem Cells for Therapeutic Purposes : Lessons from Development and the Niche. *Cells*, 8(2) (cit   p. 7).
- TAKAHASHI, K., TANABE, K., OHNUKI, M., NARITA, M., ICHISAKA, T., TOMODA, K., & YAMANAKA, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5), 861-872 (cit   p. 6).
- TAKAHASHI, K., & YAMANAKA, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4), 663-676 (cit   p. 6).
- TANAY, A., & REGEV, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637), 331-338 (cit  s pp. 12, 38).
- TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B. B., SIDDIQUI, A., LAO, K., & SURANI, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5), 377-382 (cit   p. 12).

- TASK, K., JARAMILLO, M., & BANERJEE, I. (2012). Population Based Model of Human Embryonic Stem Cell (hESC) Differentiation during Endoderm Induction. *PLOS ONE*, 7(3), 1-10 (cit   p. 16).
- THOM, R. (1989). *Structural stability and morphogenesis*. CRC press. (Cit   p. 17).
- THOMSON, J. A., ITSKOVITZ-ELDOR, J., SHAPIRO, S. S., WAKNITZ, M. A., SWIERGIEL, J. J., MARSHALL, V. S., & JONES, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. *Science*, 282(5391), 1145-1147 (cit  s pp. 4, 6).
- TILL, J. E., & MCCULLOCH, E. A. (1961). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation Research*, 14, 213-222 (cit   p. 7).
- TILL, J. E., MCCULLOCH, E. A., & SIMINOVITCH, L. (1964). A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proceedings of the National Academy of Sciences*, 51(1), 29-36 (cit   p. 14).
- TONG, A., HUANG, J., WOLF, G., van DIJK, D., & KRISHNASWAMY, S. (2020). TrajectoryNet : A Dynamic Optimal Transport Network for Modeling Cellular Dynamics. *Proceedings of the 37th International Conference on Machine Learning*, 119, 9526-9536 (cit  s pp. 20, 39, 78).
- TOSENBERGER, A., GONZE, D., BESSONNARD, S., COHEN-TANNOUDJI, M., CHAZAUD, C., & DUPONT, G. (2017). A multiscale model of early cell lineage specification including cell division. *NPJ Syst. Biol. Appl.*, 3(1), 16 (cit   p. 16).
- TRAPNELL, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.*, 25(10), 1491-1498 (cit  s pp. 12, 18).
- TRAPNELL, C., CACCHIARELLI, D., GRIMSBY, J., POKHAREL, P., LI, S., MORSE, M., LENNON, N. J., LIVAK, K. J., MIKKELSEN, T. S., & RINN, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4), 381-386 (cit  s pp. 16, 18).
- TRONSTAD, M., KARLSSON, J., & DAHLIN, J. S. (2025). MultistageOT : Multistage optimal transport infers trajectories from a snapshot of single-cell data. *Proceedings of the National Academy of Sciences*, 122(50), e2516046122 (cit  s pp. 20, 39).
- TURING, A. M. (1953). The chemical basis of morphogenesis. *Bull. Math. Biol.*, 52(1-2), 153-97 (cit   p. 14).
- UDOMLUMLEART, T., HU, S., & GARG, S. (2021). Lineages of embryonic stem cells show non-Markovian state transitions. *iScience*, 24(8), 102879 (cit   p. 15).
- VENTRE, E. (2021). Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics. *In Silico Biol.*, 14(3-4), 89-113 (cit  s pp. 26, 33, 54-56).
- VENTRE, E., HERBACH, U., ESPINASSE, T., BENOIT, G., & GANDRILLON, O. (2023). One model fits all : Combining inference and simulation of gene regulatory networks. *PLOS Computational Biology*, 19(3), 1-28 (cit  s pp. 34, 35, 41, 50, 55, 65).

- VIÑUELAS, J., KANEKO, G., COULON, A., BESLON, G., & GANDRILLON, O. (2012). Towards experimental manipulation of stochasticity in gene expression. *Progress in biophysics and molecular biology*, 110(1), 44-53 (cité p. 10).
- VIÑUELAS, J., KANEKO, G., COULON, A., VALLIN, E., MORIN, V., MEJIA-POUS, C., KUPIEC, J.-J., BESLON, G., & GANDRILLON, O. (2013). Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. *BMC biology*, 11(1), 15 (cités pp. 10, 25).
- WADDINGTON, C. H. (1957). *The Strategy of the Genes*. Allens & Unwin. (Cités pp. 5, 16, 54).
- WANG, Y., & NAVIN, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Mol. Cell*, 58(4), 598-609 (cité p. 12).
- WANG, Z., GERSTEIN, M., & SNYDER, M. (2009). RNA-Seq : a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1), 57-63 (cité p. 11).
- WAŻEWSKA-CZYŻEWSKA, M., & LASOTA, A. (1976). Mathematical problems in the dynamics of the red corpuscle system. *Mat. Stos*, 6, 23-40 (cité p. 14).
- WEINREB, C., WOLOCK, S., TUSI, B. K., SOCOLOVSKY, M., & KLEIN, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10), E2467-E2476 (cité p. 12).
- WILBREY-CLARK, A., ROBERTS, K., & TEICHMANN, S. A. (2020). *Cell atlas technologies and insights into tissue architecture* (T. 477). Biochemical Journal. (Cité p. 12).
- YACHIMURA, T., WANG, H., IMOTO, Y., YOSHIDA, M., TASAKI, S., KOJIMA, Y., YABUTA, Y., SAITOU, M., & HIRAOKA, Y. (2024). scEGOT : single-cell trajectory inference framework based on entropic Gaussian mixture optimal transport. *BMC Bioinformatics*, 25(1), 388 (cités pp. 20, 21, 39, 40).
- ZHU, J., ZHANG, K., ZHANG, Z., & KONG, D. (2025). Modeling Cell Type Developmental Trajectory using Multinomial Unbalanced Optimal Transport. *arXiv preprint* (cités pp. 20, 39).

RÉSUMÉ

La différenciation cellulaire est un processus biologique dynamique qui conduit une cellule à acquérir une identité cellulaire particulière. Récemment, le séquençage d'ARN en cellules uniques a permis de mesurer simultanément les niveaux d'expression génique à des instants précis de ce processus pour un grand nombre de cellules et de gènes. La répétition de ces mesures à différents moments permet d'accéder à la variation temporelle d'une distribution dans l'espace d'expression génique. La plupart des techniques de mesure étant destructives, l'inférence de trajectoires de différenciation cellulaire vise à reconstituer le processus de différenciation au sein de chaque cellule de la population. Les travaux existants supposent que les cellules évoluent dans l'espace d'expression génique par diffusion. Or, la dynamique réelle des gènes est bien plus complexe.

Dans mon travail de thèse, j'ai développé une méthode pour inférer une trajectoire de différenciation cellulaire à partir de données issues de séquençage d'ARN en cellules uniques, en m'appuyant sur la résolution du problème de Schrödinger et un modèle de gènes mécaniste et stochastique générant des dynamiques réalistes. La méthode repose sur la construction d'un processus de référence pour le problème de Schrödinger à partir du modèle de gènes appliqué à un réseau de régulation génique, appelé modèle Bursty.

En comparant les solutions du problème de Schrödinger obtenues avec un processus de référence diffusif et un processus de référence basé sur le modèle Bursty, j'ai établi la pertinence de cette nouvelle méthode, d'abord à partir de données simulées et d'un modèle de gènes simple, puis à partir de données expérimentales et d'un réseau de gènes complexe. Ce travail illustre donc l'intérêt d'intégrer une connaissance mécaniste du système biologique à la méthode d'inférence pour permettre une inférence plus réaliste et, *in fine*, plus pertinente de la différenciation cellulaire.

Mots-clés : Inférence de trajectoire, modèle d'expression génique, réseau de régulation génique, problème de Schrödinger, différenciation cellulaire, données en cellules uniques.

ABSTRACT

Cell differentiation is a dynamic biological process that leads a cell to acquire a specific cellular identity. Recently, single-cell RNA sequencing has enabled simultaneous measurement of gene expression at specific points in this process across a large number of cells and genes. Repeating these measurements at different times provides insight into the temporal variation of a distribution in the gene expression space. Since most measurement techniques are destructive, the goal of inferring cell differentiation trajectories is to reconstruct the differentiation process at the level of each cell in the population. Existing work assumes that cells evolve, within the gene expression space, via diffusion. However, the actual dynamics of gene expression are far more complex.

In my PhD thesis, I developed a method for inferring a cell differentiation trajectory from single-cell RNA sequencing data, based on the solution to the Schrödinger problem and a mechanistic, stochastic gene model that generates realistic dynamics. The method relies on constructing a reference process for the Schrödinger problem based on the gene model applied to a gene regulatory network, known as the Bursty model.

By comparing the solutions to the Schrödinger problem obtained with a diffusive reference process and a reference process based on the Bursty model, I established the validity of this new method, first using simulated data and a simple gene model, and then using experimental data and a complex gene network. This work thus illustrates the value of integrating mechanistic knowledge of the biological system into the inference method to enable a more realistic and, ultimately, more relevant inference of cell differentiation.

Keywords: Trajectory inference, gene expression model, gene regulatory network, Schrödinger problem, cell differentiation, single-cell data.