One model fits all: combining inference and simulation of gene regulatory networks

Elias Ventre[1,2,3,6], Ulysse Herbach[4,6], Thibault Espinasse[2,3], Gérard Benoit[1,5], Olivier Gandrillon[1,2,*]

[1]*Laboratoire de Biologie et Modelisation de la Cellule, Ecole Normale Superieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Universite Claude Bernard Lyon 1, 46 allee d'Italie F-69364 Lyon, France.*
[2]*Inria Center Grenoble Rhône-Alpes, Equipe Dracula, Villeurbanne, France*
[3]*Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, Villeurbanne, France*
[4]*Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France*
[5]*Present address; Université de Rennes 1, IGDR - CNRS UMR6290 - Inserm 1305, 35042 Rennes Cedex, France*
[6]*Co-first author*
[*]*Corresponding author: olivier.gandrillon@ens-lyon.fr*

Abstract:
The rise of single-cell data highlights the need for a nondeterministic view of gene expression, while offering new opportunities regarding gene regulatory network inference. We recently introduced two strategies that specifically exploit time-course data, where single-cell profiling is performed after a stimulus: HARISSA, a mechanistic network model with a highly efficient simulation procedure, and CARDAMOM, a scalable inference method seen as model calibration. Here, we combine the two approaches and show that the same model driven by transcriptional bursting can be used simultaneously as an inference tool, to reconstruct biologically relevant networks, and as a simulation tool, to generate realistic transcriptional profiles emerging from gene interactions. We verify that CARDAMOM quantitatively reconstructs causal links when the data is simulated from HARISSA, and demonstrate its performance on experimental data collected on *in vitro* differentiating mouse embryonic stem cells. Overall, this integrated strategy largely overcomes the limitations of disconnected inference and simulation.

Keywords: gene regulatory networks, causal inference, data simulation, transcriptional bursting, stochastic gene expression, single-cell transcriptomics, time-course profiles, lineage commitment

Introduction

Cell decision making as a response to exogenous or endogenous stimuli (e.g., differentiation, proliferation, cell death or biological activity modulation) is often supported by time-dependent modulation of gene expression upon stimulation. Understanding how and why gene expression changes as a function of time in response to specific stimuli is therefore critical to understand the underlying biological processes.

The "how" question can now be approached using single-cell-based technologies, offering an unprecedented resolution and a much finer view than population-based measures [1, 2]. The "why" question relates to the functioning of an underlying gene regulatory network (GRN) which describes interactions between genes through their expression products. GRNs are thus a central notion for understanding and predicting cellular behavior, but their construction from literature is a very laborious task, sometimes even impossible due to the lack of knowledge.

Reconstructing most-likely GRNs from transcriptomic datasets has therefore become a major goal in systems biology [3] but is also notoriously difficult, especially in the case of single-cell transcriptomic data. Indeed, the bursty synthesis of mRNAs, now clearly evidenced [4, 5], gives rise to highly variable and non-Gaussian expression data [1, 6], and current GRN inference methods employ a wide range of statistical and modeling tools [7]. Methods based on a specific mechanistic model have the great advantage of providing biological interpretability, since each inferred interaction between genes can be understood in terms of model behavior. Moreover, such approach generally provides interactions with their direction and intensity, which is not the case for most purely statistical methods.

However, the results of a method based on a mechanistic model can only be considered relevant if the model is able to correctly reproduce single-cell datasets. For instance, it is now widely accepted that the transcriptional bursting phenomenon is associated to specific patterns of gene expression products [8, 9], making continuous single-cell data close to Gamma distributions [10] and discrete data close to negative binomial distributions [11], the latter being themselves mathematically equivalent to Poisson distributions with Gamma-distributed random parameters. Thus, executable network models should at least be able to generate these patterns in their marginal distributions. In any case, the use of a mechanistic model-based method requires prior strong evidence that the underlying model is relevant for simulating realistic single-cell transcriptomic data sets.

We recently developed several methods for inferring GRNs from single-cell data based on a particular mechanistic network model, defined as a 'multi-agent' generalization of the well-known two-state stochastic model of gene expression [8] where genes are now being described by interacting two-state models [6]. These methods are well suited for single-cell RNA-seq (scRNA-seq) time-course data, each dataset being considered as a partial observation of the model at a certain time. Crucially, they do not require the observation of cell trajectories, whose inference is a problem in itself [12, 13], but only that the cells sampled at each timepoint are driven by the same dynamical process, i.e., resulting from the same GRN. Our first proposal was called WASABI [14], which uses a divide-and-conquer approach where the problem of GRN inference is solved one gene at a time. Although able to propose relevant GRNs, this approach suffered from two

drawbacks: it required days of computation for a GRN with 50 genes, and proposed a potentially long list of candidate networks. We therefore developed two other methods: HARISSA [6], a GRN simulation algorithm based on the mechanistic model together with a proof-of-concept inference method derived from likelihood maximization, and CARDAMOM [15], a simplified and scalable alternative for the GRN inference part that crucially exploits the notions of landscape and metastability.

In this work, we sought to investigate the benefits of using this model as an integrated tool for both GRN inference and data simulation. We therefore assessed its ability to allow for efficient network reconstruction from time-course scRNA-seq data, while accurately reproducing the dataset main features from the functioning of the inferred network. Note that to the best of our knowledge, this is not performed by existing GRN-based simulation tools, which are generally based on more phenomenological than mechanistic models [16, 17], meaning that gene expression patterns, and especially transitions between cell types, are hard-coded instead of emerging from interactions between genes.

After introducing the setup of our benchmark made from *in silico* datasets generated with the mechanistic model, we first evaluate the performances of HARISSA and CARDAMOM together with four state-of-the-art GRN inference algorithms: GENIE3 [18], PIDC [19], SINCERITIES [20] and SCRIBE [21]. We study the limits of the different categories of inference methods in the case of transcriptional bursting, and verify that the two model-based methods perform better than the others on these datasets. CARDAMOM appears as the best performing algorithm during this benchmark step, which only considers network structures. Importantly, the output of this algorithm is not only a matrix of interaction scores, but also a set of quantitative parameters that can be plugged into the GRN model for simulations.

In a second step, we use CARDAMOM to calibrate the model with a real time-stamped scRNA-seq dataset of differentiating mouse embryonic stem (ES) cells [22]. We demonstrate the ability of the model to reproduce the global features of real time-course transcriptomic profiles. We also show that most of the inferred interactions are indeed supported by biological evidence such as ChIP-seq experiments, although this evidence was not used during the inference process. Altogether, these results establish the ability of an executable network model not only to simulate realistic single-cell datasets, but also to provide an effective reverse-engineering algorithm capable of reproducing the main gene expression patterns of an experimental dataset as emergent properties of the underlying GRN.

Results

HARISSA simulates single-cell datasets from a mechanistic GRN model

We first wanted to benchmark the ability of the different inference algorithms to reconstruct correct network structures from *in silico* generated datasets, i.e., when the ground truth is known. For this, we used the simulation module of HARISSA [23], which generates trajectories of a mechanistic model describing gene expression dynamics (both mRNA and the corresponding proteins) within a single cell, these dynamics being influenced by an underlying GRN and driven

by transcriptional bursting (see Methods and Figure S1). As shown in previous work, this model is indeed able to generate scRNA-seq datasets with realistic marginal distributions [6, 23].

We simulated nine types of datasets corresponding to different network structures (Figure 1): a network of 4 genes with a branching structure and inhibition feedback loop (FN4); a network of 5 genes with a cycling structure (CN5); a network of 8 genes with multiple branching structure and feedback loops (FN8); a network of 8 genes with branching trajectories (BN8); networks with a tree structure of 5, 10, 20, 50, and 100 genes (Trees). These networks represent the main types of network structures that have been used for benchmarking GRN inference algorithms [17]. Overall, the objective was to reproduce time-course experiments in which single-cell profiling is performed after a given stimulus, typically a change of medium [22, 24, 25]. This stimulus was therefore taken into account in all the simulations, in the form of a virtual gene defined as being inactive before the beginning of the experiment and fully activated afterwards.

For each network structure (Figure 1A), the transcriptional bursting model implies that typical single-cell trajectories do not follow a diffusion-like process (at least in the space of mRNA levels), and differ strongly from the more usual and intuitive population-average trajectory (Figure 1B). The practical datasets were obtained by sampling independent cells at a specific sequence of timepoints, therefore not keeping the real cell trajectories but rather considering different cells at each timepoint, forming time-stamped snapshots (Figure 1C). Interestingly, both feedback networks (FN4 and especially FN8) produce a recognizable "differentiation trajectory" across the UMAP space with a clear temporal order of cells. Due to the stochastic nature of cell trajectories generated by the mechanistic model, branching trajectories in snapshots only appear in specific cases, generally when a toggle-switch is dominating the GRN structure and then generates distinct branches in the UMAP representation (see BN8 in Figure 1).

As mentioned previously, HARISSA consists of two modules for performing respectively simulation and inference. Whereas the original inference module of HARISSA was limited to a few genes [6], it recently integrated an effective CARDAMOM-inspired simplification [23] that allows to infer networks with a much larger number of genes. We therefore also benchmarked this method along with the others.

CARDAMOM quantitatively reconstructs causal GRN links

We then inferred GRN structures from the *in silico* generated datasets using the six algorithms presented in the Methods section (HARISSA, CARDAMOM, GENIE3, PIDC, SINCERITIES, and SCRIBE). Note that neither GENIE3 nor PIDC are able to use the temporal information (except for the stimulus state information, which they are also provided with), giving them a disadvantage compared to the other algorithms. They were nevertheless used in the benchmark as they are considered to be among the best algorithms for single-cell data, and given that very few algorithms are specifically adapted to time-stamped datasets. Indeed, most methods are limited to static data, and those that are not (such as SCRIBE) require temporally-ordered cell trajectories instead of independent snapshots, thus requiring a pre-processing step that can itself be subject to errors. Moreover, it was not known how they would fare in a time-course setting with transcriptional bursting, which was an interesting question per se.

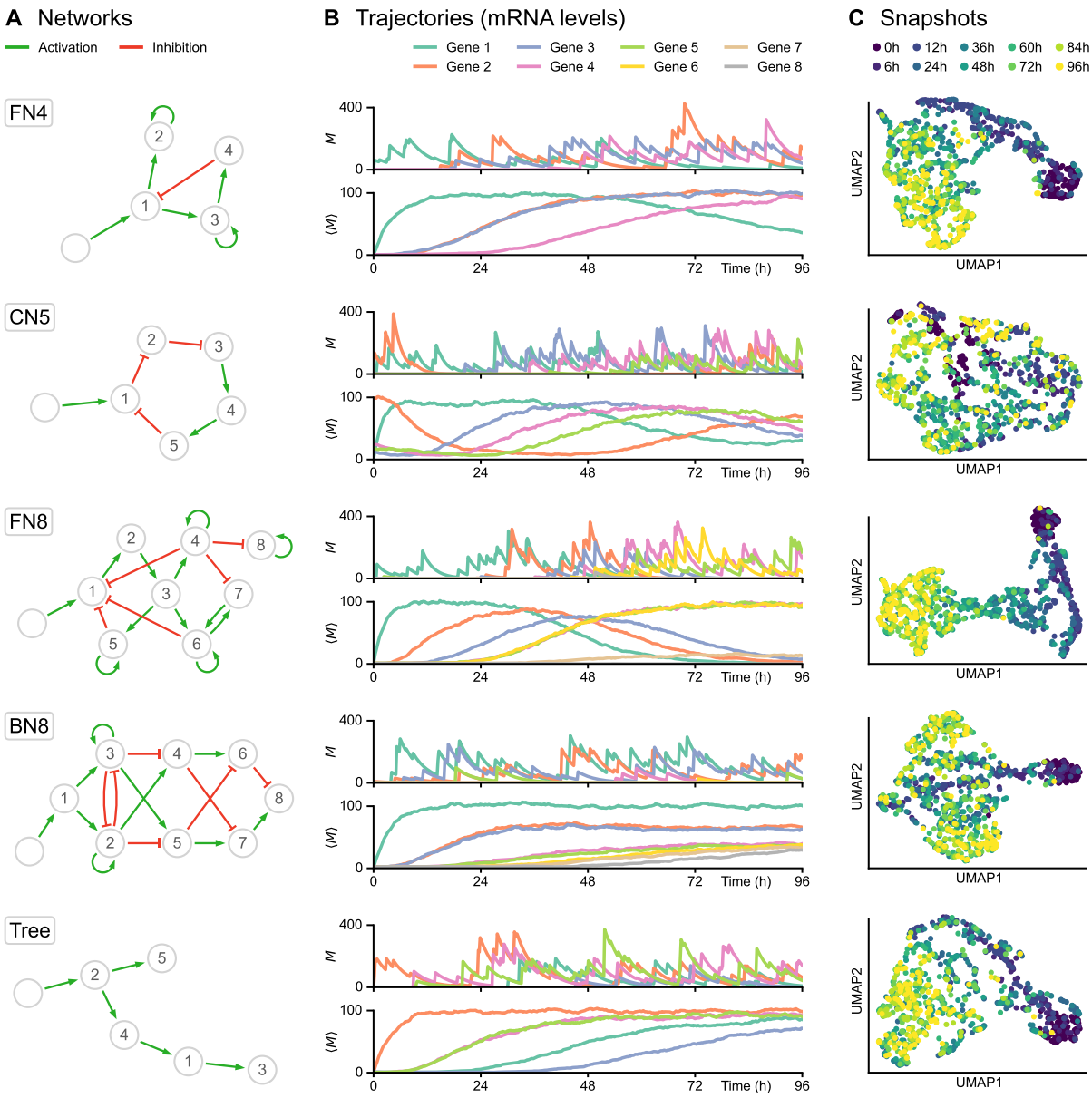

Figure 1: Single-cell data simulation using HARISSA. (A) Networks used for subsequent tests, including feedback loop networks (FN), a cycling network (CN) and a branching network (BN). Genes and stimulus are represented by numbered nodes and an empty node, respectively, while green arrows indicate activation and red blunt arrows indicate inhibition. **(B)** Corresponding trajectories, defined as time-dependent mRNA levels (in copies per cell). For each network, the first plot shows one example of single-cell trajectory M while the second plot shows the population average $\langle M \rangle$ from 1000 cells. The transcriptional bursting model underlying HARISSA implies that every single-cell trajectory differs strongly from the more usual population average. **(C)** Two-dimensional UMAP representations of corresponding single-cell snapshots, defined as mRNA levels sampled at 10 timepoints in different cells from 0h to 96h, with 100 cells per timepoint. Such snapshots are called *time-stamped data* in the text and are fundamentally different from single-cell trajectories, which are currently not available experimentally.

We also emphasize that among these algorithms, only CARDAMOM and HARISSA have the significant advantage of providing biological interpretability, thanks to the mechanistic model on which they are based: here the network parameters are not mere interaction scores, but quantitative parameters that can be plugged into the model for simulations. From this perspective, even similar performances compared to the other algorithms would be satisfying.

Inference was performed on ten independent datasets for each condition, and the results were merged into the area under the precision-recall curve (AUPR) which measures the quality of the inferred GRN structure. We also compared the inferred GRNs with a naive method consisting in assigning to each edge of the network the value given by the Pearson correlation coefficient between the corresponding genes (abbreviated as PEARSON): this comparison with Pearson coefficients makes it possible to verify, when the algorithms show good performances, that these are not only due to highly correlated data which are thus not difficult to analyze. The results are presented in Figure 2A-B for the first five algorithms. We present the results for SCRIBE separately in Figure 2C because this algorithm requires temporally or pseudo-temporally ordered trajectories, and the results then depend on the pre-processing that is applied on the time-stamped data.

CARDAMOM and HARISSA appeared to outperform the other algorithms for most of these datasets. In particular, in terms of directed interactions, these two methods always clearly performed better than the others. The undirected networks for which GENIE3 and PIDC have similar performances (CN5 and BN8) correspond to cases where the Pearson correlation method is also accurate.

Also, if GENIE3 and PIDC represent an improvement over the Pearson correlation method, they seem to perform poorly when the correlation between genes is not sufficient to infer a reliable GRN. More precisely, we observe that GENIE3 and PIDC are accurate for tree-like networks (Trees), even with bifurcating trajectories (BN8) and cycling (CN5), which was not the case in [17]. On the contrary, SINCERITIES performs very poorly for these type of networks, but seems however competitive for networks with feedback loops (FN4 and FN8) where GENIE3 and PIDC have lower performances. These networks are more difficult to reconstruct. Indeed, as visible in Figure 1, the population-average trajectories of some genes are completely similar. Some genes also have the same marginal distribution of mRNA levels: for example in the network FN4, gene 2 and gene 3 have the same input (gene 1), so their marginal distributions evolve similarly at each timepoint. Then SINCERITIES, which bases the inference procedure on the approximate distribution for each gene, fails to make this subtle distinction, illustrating the improvement that is typically expected from HARISSA and CARDAMOM. On all the networks, GENIE3 fails to infer reliably the direction of the interaction, i.e., to distinguish the interaction $i \to j$ from the interaction $j \to i$. On the contrary, because of their mechanistic assumptions, CARDAMOM, HARISSA and SINCERITIES have always quite similar results for directed and undirected inference. Finally, we observed that CARDAMOM outperforms HARISSA on most of the networks.

Regarding SCRIBE, we tested its performances for 3 types of data (Figure 2C):

1. When we have access to real trajectories (*Real traj.*): each cell at each timepoint is being associated to a real ancestor at the previous timepoint and a real descendant at the

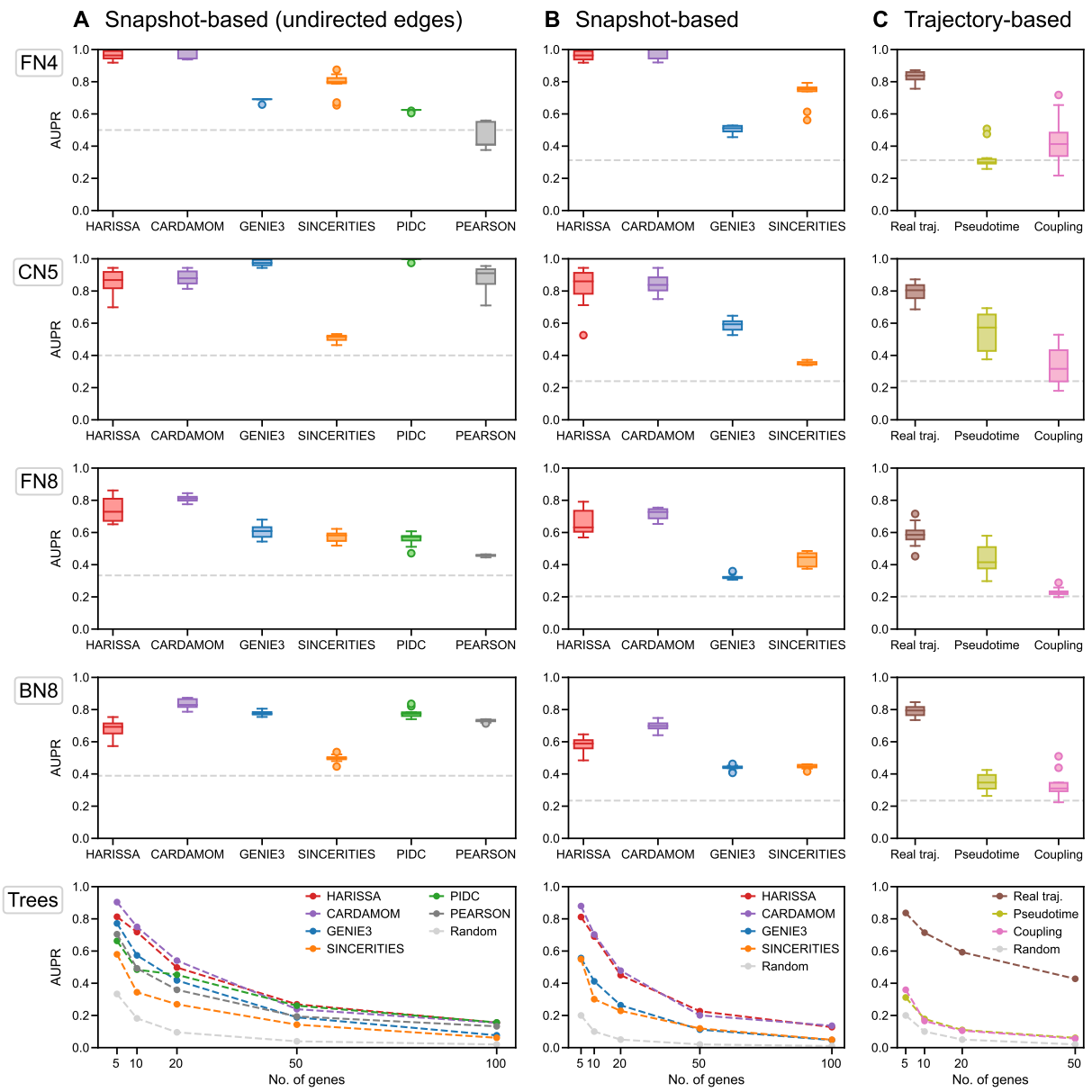

Figure 2: Benchmark of inference methods for five different network structures. For each network, inference is performed on ten independently simulated datasets, each dataset containing the same 10 timepoints with 100 cells per timepoint (1000 cells sampled per dataset). The performance on each dataset is then measured as the area under the precision-recall curve (AUPR), based on the unsigned inferred weights of edges. Finally, the performance of each method is summarized as a box plot of the corresponding AUPR values, or the average AUPR value for the tree-structure activation networks (*Trees*). For each plot, the dashed gray line indicates the average performance of the random estimator (assigning to each edge a weight 0 or 1 with 0.5 probability). For the *Trees* networks, each dataset corresponds to a random tree structure of fixed size (5, 10, 20, 50, and 100 genes) sampled from the uniform distribution over trees of this size. **(A)** Performance of all methods when considering only undirected interactions. **(B)** Performance of the methods able to infer directed interactions. **(C)** Performance of the SCRIBE inference method for the same networks, in three conditions: when one has access to real single-cell trajectories (in brown), when pseudo-trajectories are reconstructed from time-stamped data using a coupling method similar to Waddington-OT (in pink), and when a single pseudo-trajectory is reconstructed using the pseudotime algorithm SLINGSHOT (in light green).

following one. Such knowledge can of course only be accessed with *in silico* generated datasets or *in vitro* for a very limited number of genes by using live-cell imaging of short-lived transcriptional reporters [26];

2. When we do not have access to real trajectories, and each cell at each timepoint is associated to a pseudo-ancestor at the previous timepoint and a pseudo-descendant at the following one, using the Waddington-OT method described in [27] (*Coupling*);

3. When we do not have access to real trajectories, and the algorithm SLINGSHOT [28] is used for reconstructing a pseudo-temporally ordered trajectory (*Pseudotime*).

We observed that SCRIBE performs well when the datasets are built with the method 1, but poorly when the datasets are built with the methods 2 and 3, at least on the tested networks (Figure 2C). These poor performances are due to the loss of temporal coupling between measurements of genes that interact. They suggest that neither optimal coupling nor pseudotime reconstruction are sufficiently efficient for GRN inference in case of transcriptional bursting. Concerning the optimal coupling method, we notice that this might be due to the "movement by diffusion" assumption on which the method Waddington-OT is built, which does not takes into account the constraints on the trajectories imposed by the GRN.

When computing the average runtime of each algorithm on the tree-like networks, we observed that except for SCRIBE, all algorithms are suitable for inferring GRN with a realistic number of genes (see Table S1). Thus, due to this computational limit and its poor performances when using time-stamped data, we did not consider SCRIBE for further analysis.

We then investigated the limit performances with respect to the number of cells and/or timepoints. We observed that the performances of the first five algorithms decrease for the tree-like networks when the number of genes increases (Figure 2). This can be due to three main factors:

1. A sequence of timepoints too coarse in relation to the dynamics would directly lead to a lack of inference accuracy;

2. A sequence of timepoints which is too restricted may not allow to see interactions involving some genes that are regulated late in the process. For example, in Figure 2, we observe that the inference on the Trees networks is very poor for more than ten genes: it comes from the fact that some genes are never activated before 96h;

3. The number of cells at each timepoint can simply not be enough to infer a reliable GRN.

We therefore investigated the effects of these three factors on the accuracy of the algorithms by studying their performances in terms of AUPR for ten datasets generated from ten randomly-generated tree-like network of ten genes, when varying the number of cells at each timepoint (Figure 3A), the length of the interval for a fixed time gap between each timepoint (Figure 3B), and the density of the sequence of timepoints for an interval with fixed length (Figure 3C). As anticipated, all these conditions have an impact on the quality of the inference: augmenting their values tends to produce a better quality of inference. We also observed that the number of sampled cells seems less critical than the other factors, confirming that few cells at a sequence of timepoints which is dense and long-enough is preferable to many cells on a sequence of

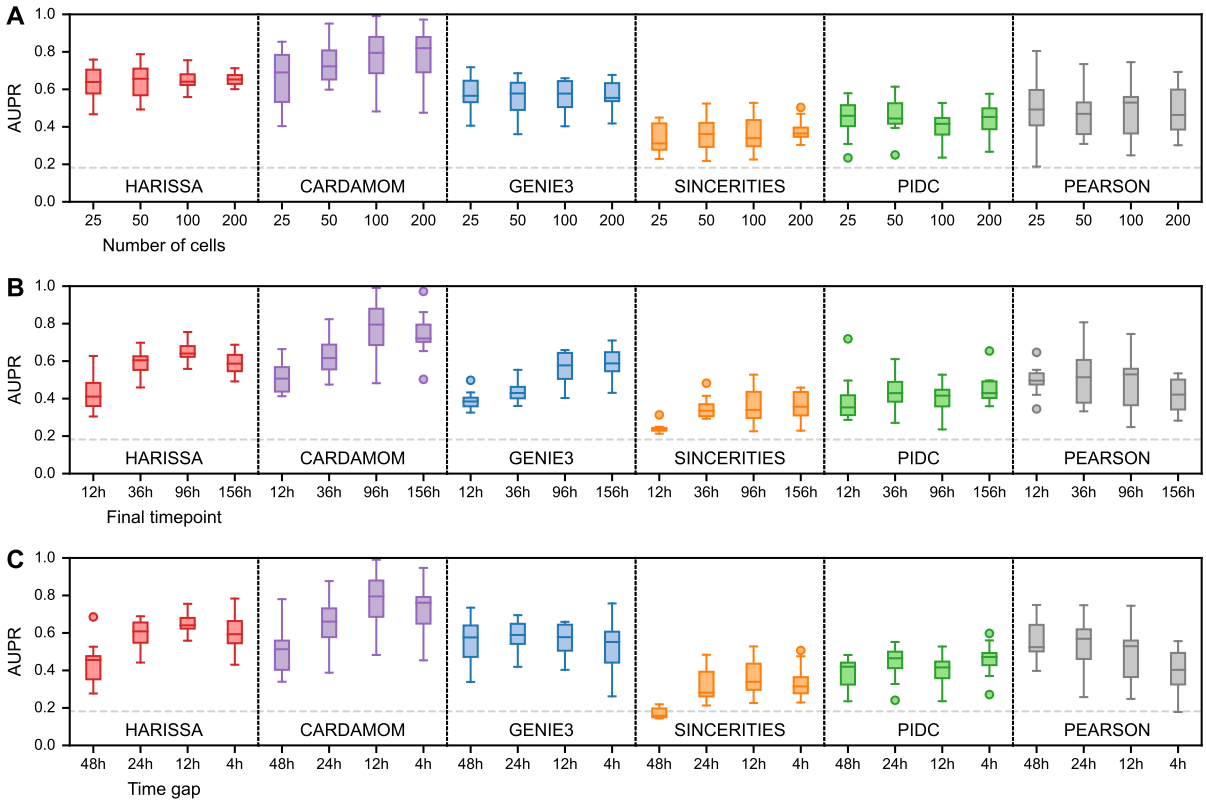

Figure 3: Dependence of inference methods on data collection parameters. For simplicity, only the case of undirected interactions is considered here and the datasets are restricted to 10-gene tree-structure networks (see Figure 2 for the general benchmark). Inference is performed for each method and condition on ten independently simulated datasets and summarized by box plots of AUPR values as in Figure 2. **(A)** Performance as a function of the number of cells per timepoint, while keeping the same timepoints. **(B)** Performance as a function of the length of the measurement period, while keeping the same gap between timepoints and the same total number of cells. **(C)** Performance as a function of the density of the measurements, while keeping the same final timepoint and the same total number of cells.

timepoints which is too coarse and/or too short. This should be kept in mind when designing single-cell transcriptomics experiments aiming at GRN inference.

Hence, both CARDAMOM and HARISSA, with a benefit for using CARDAMOM, allowed to efficiently reconstruct network structures by reverse engineering the generative model on which they are based. We then needed to test its ability to reproduce an experimental dataset from the literature after network inference.

Application to a real dataset yields a biologically relevant network

As a test case, we used a time-stamped *in vitro* dataset from Semrau et al. [22] obtained by scRNA-seq of a retinoic acid (RA)-induced differentiation of mouse ES cells (see Methods). This well-characterized model system of *in vitro* differentiation recapitulates the transition from pluripotent embryonic stem cells towards two cellular lineages (ectoderm- and extraembryonic endoderm-like cells), all characterized by well-established molecular markers that were further used in GRN inference.

In order to interpret the resulting GRN, we sought to assess whether the inferred interactions

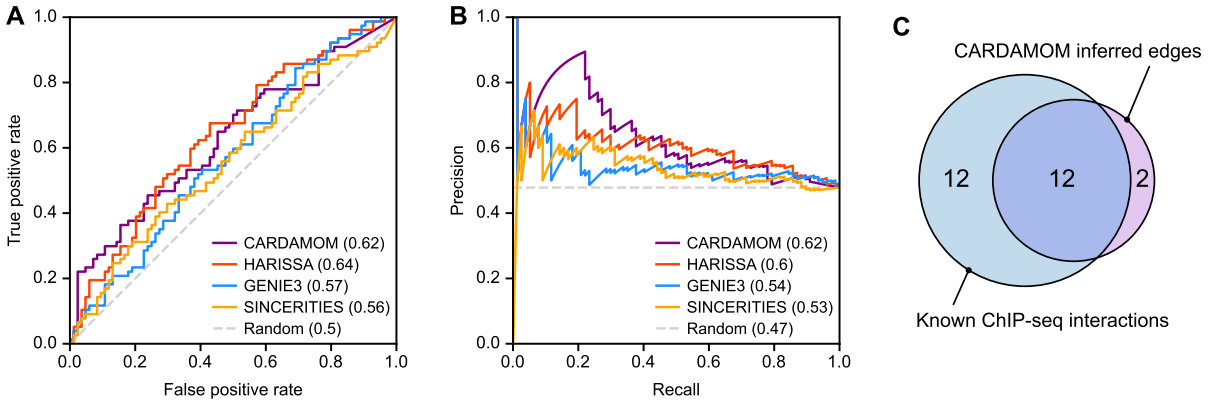

Figure 4: Comparison between inference methods and physical interactions derived from ChIP-seq. The four directed GRN inference methods were applied to the experimental dataset from [22] restricted to a panel of 41 marker genes identified by the authors, and a reference network was obtained independently from edges supported by ChIP-seq data. As we only have access to physical interactions involving the retinoic acid (RA) stimulus or genes Pou5f1, Sox2, and Jarid2, the comparison only considers the related edges. **(A)** Receiver operating characteristic (ROC) curve and corresponding area under the curve (AUROC) for each inference method. **(B)** Precision-recall (PR) curve and corresponding area under the curve (AUPR) for each inference method. **(C)** Venn diagram showing the overlap, for interactions involving the RA stimulus, between directed edges predicted by CARDAMOM and known physical interactions identified by ChIP-seq analysis.

are supported by known biochemical evidence of physical interaction between regulators and regulated genes (Figure 4). For this, we annotated the inferred edges coming from genes encoding known transcription regulators (i.e., transcription factors and cofactors) included in the network and for which ChIP-seq data are currently available in ES or the closely related embryonic carcinoma (EC) cell system. Since the RA stimulus exerts its differentiating effect mainly through the members of the RA-activated nuclear receptors subfamily RAR (NR1B) that encompass 3 paralogs (i.e., RARα, RARβ and RARγ), the annotation of the interaction edges linking the stimulus and the regulated genes was based on the presence/absence of ChIP-seq peaks for any RAR paralogs at less than 10 kb upstream or downstream of the annotated transcription start site (TSS) in RA-stimulated ES or EC cells [29, 30]. Although arbitrary, the chosen distance between TSS and DNA binding site for the indicated transcription regulator is relatively conservative as transcriptional effect could be exerted from greater distance up to megabases [31] and the absence of supporting peak as defined should not be interpreted as a proof of absence of any direct modulating effect. Similarly, the edges supported by physical interactions data for Sox2 and Pou5f1, or Jarid2 were extracted from [32, 33].

Using these known physical interactions as a ground-truth, we compared the receiver operating characteristic (ROC) and precision-recall (PR) curves related to the network structures inferred by the four algorithms (Figure 4A-B). We observed that, in accordance with the previous results, CARDAMOM and HARISSA appear as the top-ranked algorithms, displaying both a very close ability to infer known edges.

We then examined the structure of the network inferred by CARDAMOM (Figure 5). Importantly, in agreement with its differentiating effect in ES/EC cell systems, we observed that the RA stimulus is densely connected with genes involved in pluripotency maintenance as supported by multiple

biological analyses [29, 30, 34] and to a lesser extent with gene nodes corresponding to genes associated with specific cell fates, this latter observation likely reflecting how the stimulus is modeled (see Methods). Notably, these last nodes also exhibit a relatively high interconnectivity (e.g., endodermal differentiation) as compared to intergroup connectivity. Although biologically interesting, these observations illustrate our previous conclusion and likely mirrors the unbalanced experimental design characterized by a dense sequence of timepoints during the early phase (0h to 36h) of the differentiation process analysed and a coarser sequence of timepoints in the mid (36h to 48h) and late (48h to 96h) phases of the process [22].

Most notably, the overwhelming majority (0.85%) of the inferred edges that involve the RA stimulus are supported by biochemical evidence (Figure 4C). Similarly, the edges inferred from Pou5f1, Sox2, and Jarid2 nodes are globally supported by physical interaction (2/3 for Pou5f1, 2/3 for Sox2, and 1/1 for Jarid2).

We also observed that some inferred edges are not supported by documented physical interactions, as expected for genes encoding proteins unable to directly interact with DNA. As an example of such node, Sparc (also known as Osteonectin) appears highly connected to genes associated with all four cell states despite its inability to directly interact with gene basal transcription machinery (i.e., RNA polymerase complex). However, the inferred edges are clearly in agreement with its documented role in promoting endodermal differentiation [35]. Additionally, unsupported inferred edges may mirror the lag time between the expression and therefore the physical interaction between regulator and regulated genes and the observed transcriptional effect. By contrast with TFs that establish contact with the transcription machinery, modifying cofactors often catalyse deposition/erasure of epigenetic marks (e.g., acetylation/methylation of histones, DNA methylation) that will likely modulate transcription in a longer lasting manner. In this respect it is interesting to note that the Dnmt3b gene negatively interacts with many other genes in the network, which mirrors the fact that Dnmt3b is a *de novo* DNA methyltransferase, and has an indirect effect on gene regulation through CpG methylation, a well documented epigenetic mark generally associated with gene expression silencing. Altogether, this illustrates that our GRN model does incorporate various epigenetic information or indirect effects and is not restricted to physical interactions between transcription factors and their target genes.

While most inferred edges involving genes that encode TFs appear to be supported by physical interactions, many physical interactions detected by ChIP-seq are missed by CARDAMOM (e.g., 50% for RA, see Figure 4C). This observation is however not necessarily the sign of a lack of accuracy of the inferred GRN, since the detection of a physical interaction is not *per se* the hallmark of a modulating effect on the transcription level of the target gene [30]. Additionally, some specific regulatory structures are notoriously difficult to infer as illustrated by the high failure rate (96%) in inferring edges from Jarid2, a component of a repressive complex expressed in the pluripotent state and directly involved in the silencing of differentiation-associated genes. Interestingly, the interaction between Jarid2 and most of its physical targets presented in our GRN were instead wrongly detected as an inhibitory effect of the regulated genes on their regulators. This is due to the fact that CARDAMOM works by going forward in time, and thus fails to capture an inhibition that has an effect at the beginning of the process and which can be detected only further: instead, it would be prone to interpret the increase of the repressed genes by the effect

of other intermediate genes, and the decrease of the repressor by an inhibitory effect of the repressed genes. We discuss further such bias of the algorithm in the Discussion section.

For a better understanding of the inferred GRN dynamics, we also examined a dynamical network representation, where each edge appears at the timepoint for which it was detected with the strongest intensity by the inference algorithm (Figure 6). Unsurprisingly, the RA stimulus is detected at the earliest timepoint of the response (6h) and then ceases to influence the signal, which propagates in waves through the network as we described in a previous study [14]. For example, we do clearly observe the late increase of interactions for genes involved in specifying the extraembryonic endoderm.

Simulation of the inferred network reproduces the original dataset

While inferring the GRN structure, CARDAMOM also inferred all the other parameters of the model, as described in Methods, except the mRNA degradation rates $d_{0,i}$ for each gene i. These parameters $d_{0,i}$ are not negligible as they scale the dynamics of the process. To address this problem, we used values from the literature that can be found in [36] (see Methods and Table S2). Once the model has been calibrated, one can simulate an *in silico* dataset and sample the nine timepoints corresponding to the *in vitro* experiment. More precisely, we simulated two different datasets after calibrating the mechanistic model: one actually using the inferred network interactions, and one corresponding to the "null network" defined by removing the interactions (i.e., all genes individually calibrated with the same parameters but kept independent).

We first decided to verify, as advocated by Soneson et al. [37], the suitability of our generated synthetic data. We used countsimQC [37], a recent tool for comparing multivariate single-cell datasets, already used for benchmarking synthetic scRNA-seq data [38] (see Methods for more details). The synthetic dataset indeed mimics experimental data for a large number of tested characteristics (Figure S2). However, we observed that except for correlations, the features considered by countsimQC are also well reproduced by the dataset simulated with the null network. This suggests that countsimQC features are not sufficient for measuring the accuracy of the dataset reproduction.

We then explored the ability of the synthetic dataset to recapitulate more sophisticated dimensions of the experimental data. At that stage, the critical question concerns the temporality of the synthetic data. Indeed, as developed in [15] and [23], none of the algorithms used for the benchmark (and presented in Methods) allows to take into account the real temporality of the data: the only information they use is the order of the sequence of timepoints at which the cells are measured. It is therefore not necessarily expected that a dataset simulated with the network inferred by CARDAMOM can reproduce the data distribution exactly at the same timepoints. The temporality is taken into account in a second time, by setting the value of the degradation rates from the literature. However, the hypothesis that these degradation rates are not time-dependent (which of course oversimplifies the biological reality [39]) may prevent us from being able to perfectly fit the time-dependent evolution of the data.

We observed that this hypothesis indeed limits our ability to simulate the true dynamics at the last timepoint. In particular, the process seems to accelerate between 72h and 96h and the model

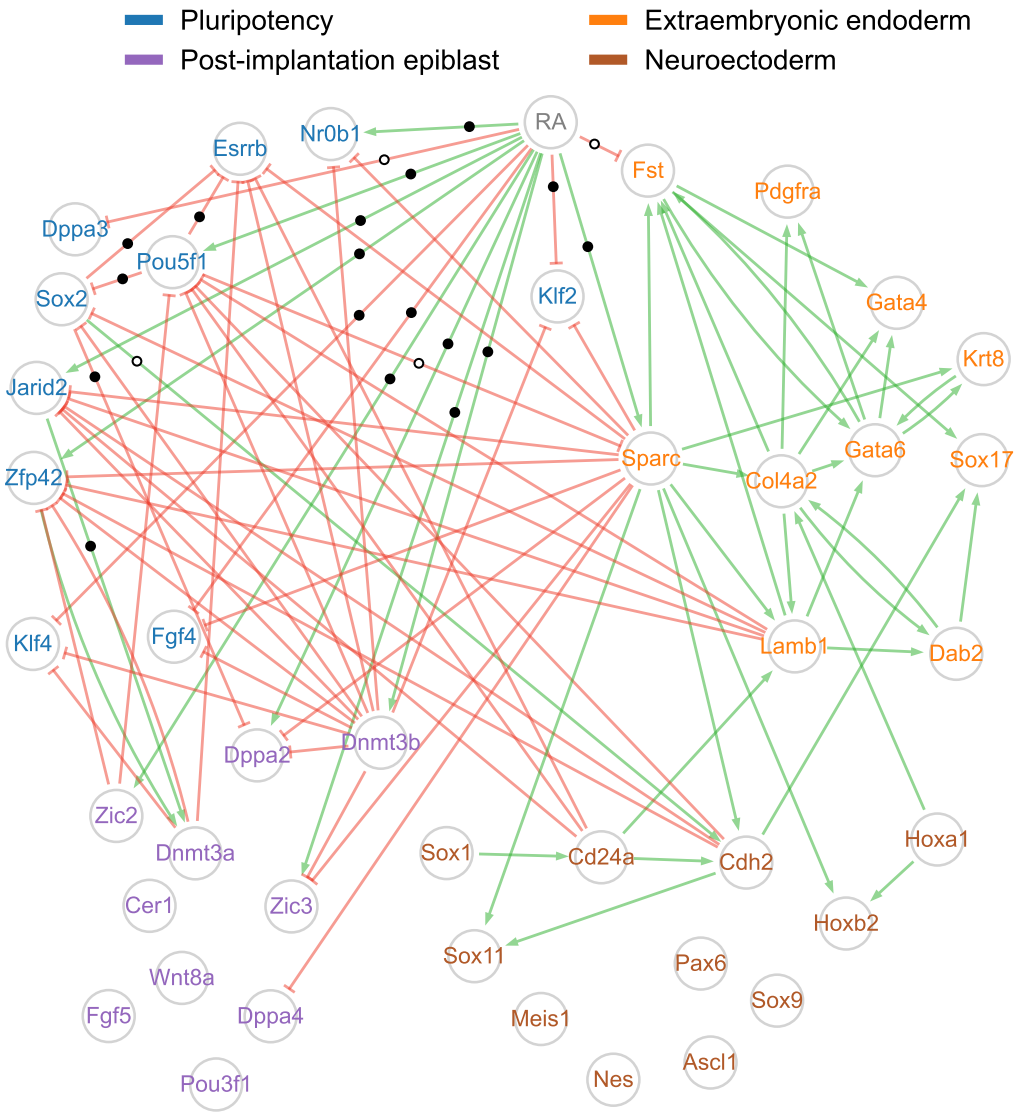

Figure 5: Network inferred by CARDAMOM from a real time-stamped scRNA-seq dataset. The CARDAMOM inference method was applied to the experimental dataset from [22] restricted to a panel of 41 marker genes identified by the authors. The network structure is obtained by keeping only the 5% strongest activations (green arrows) and inhibitions (red blunt arrows) acting on each gene. Genes are colored according to four groups related to different cell states (pluripotency, post-implantation epiblast, neuroectoderm, extraembryonic endoderm) following the proposed classification of [22]. Edges supported by a ChIP-seq interaction are marked with black dots (see main text for the definition of what is considered as an interaction) and edges that are not supported are marked with white dots. Edges for which we have no reliable information have no mark.

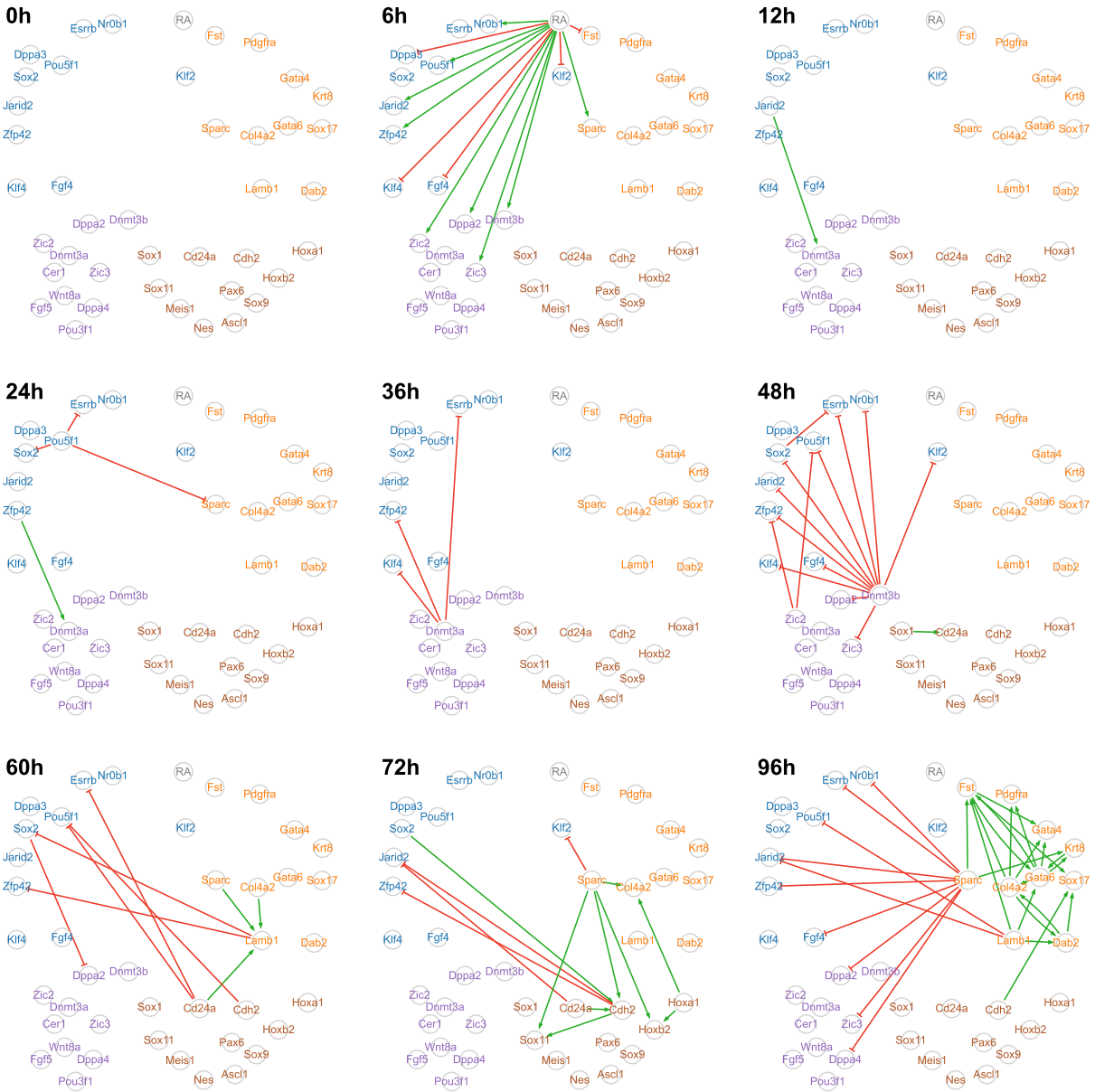

Figure 6: Time decomposition of the network inferred by CARDAMOM from a real time-stamped scRNA-seq dataset. Decomposition of the network shown in Figure 5, where each edge appears at the timepoint for which it was detected with the strongest intensity. This dynamic representation highlights a consistent flow of information coming from the stimulus. Gene positions and colors as well as activation and inhibition representations are the same as in Figure 5.

cannot be in adequacy with both the dynamics between 0h and 72h, and between 72h and 96h with the same degradation rates. It is important to note that such a global variation in degradation rates has been observed experimentally during the differentiation of chicken erythroid progenitors (see [24] and data at https://osf.io/k2q5b/). We thus decided to multiply the degradation rates by a scaling factor after 72h, allowing the process to reach its final state in time.

We compared these datasets using different metrics. First, we examined the extent to which the simulations matched the experimental marginal distributions of each gene. In Figure 7A, we represent the time-dependent evolution of the p-value of a Kolmogorov-Smirnov test between the GRN-generated distributions and the experimental dataset. We can see that some genes are better fitted than others, as exemplified by the Esrrb gene that is correctly fitted while the Sparc gene seems more difficult to catch (see Figure 7B). We nevertheless observe that for most genes and timepoints, the p-value is above 5%, meaning that the marginal distributions of the experimental data are quite well reproduced by the GRN model. We compare in Figure 7C the mean Earth Mover Distance (EMD), and in Figure 7D the mean p-value of the Kolmogorov-Smirnov test applied on the 41 genes at each timepoint, between the empirical distributions of the experimental dataset and the two simulated datasets (with and without GRN). We observe that without GRN, the distance between the distributions generated by the model are constantly increasing, that is diverging from the experimental datasets (Figure 7C). This is corroborated by the fact that the mean p-values are decreasing monotonically (i.e. the model's output is more and more significantly different from the experimentally observed distributions, see Figure 7D). The behavior of the GRN-simulated dataset is much closer to the experimental one, as seen from a smaller (and constant) EMD distance as well as a larger mean p-value. However, we observe in Figure 7A that the mid-timepoints (the central portion of the dynamics) seems to be the most difficult to capture, since mid-time p-value are often higher that at the beginning or end of the kinetics. This is corroborated by Figure S3, where we plotted the temporal marginal distributions of six genes, and where Sox2 and Sparc in particular appear to have a final distribution close to the experimental one but not the correct transient behavior.

Finally, we were interested in how well we could capture the joint distributions. For this, we compared UMAP representations of the experimental dataset (Figure 7E) and the datasets simulated from the inferred network (Figure 7F) and from the null network (Figure 7G). These three datasets were projected on the same pseudo-axes based on the UMAP computed from the experimental dataset, using the methodology described by McInnes et al. [40]. This common projection allows a better side-by-side comparison between datasets. It is immediately evident that our GRN-generated data points are very closely mimicking the actual experimental data points, and that this resemblance is completely lost if all interactions are removed. The fact that UMAP is not linear requires some precautions, that we discussed in the Methods section.

To conclude, we observed that the mechanistic model can reproduce the major characteristics of the gene expression patterns observed during a differentiation process examined at the single cell level. It also appears clearly that simulating the model with the network inferred by CARDAMOM significantly improves the fit to the experimental dataset compared to the simulation with the null network (see Figure 7 and Figure S5).

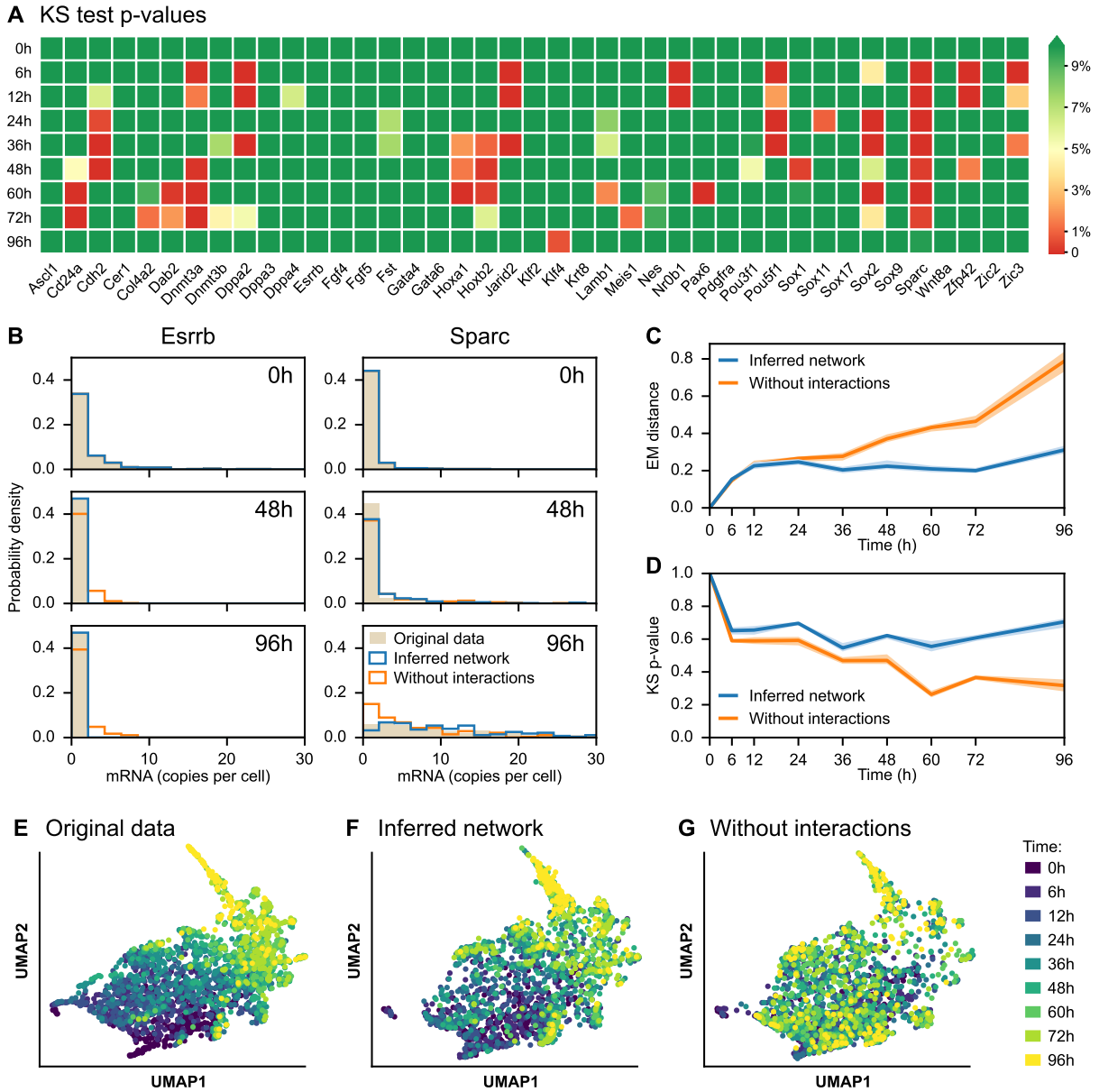

Figure 7: Inferred network simulations compared to the original dataset. (A) Heatmap of p-values associated with Kolmogorov–Smirnov (KS) tests between real and simulated mRNA distributions, for each of the 41 genes of the network and for each timepoint. The green color indicates p-values greater than 5%, implying that the model output is not significantly different from the experimental dataset. **(B)** Time-dependent distributions of Esrrb and Sparc genes for the experimental dataset (*original data*, in beige) and datasets simulated after calibrating the mechanistic model, one including interactions (*inferred network*, in blue) and one obtained after removing interactions (*without interactions*, in orange). **(C-D)** Average earth's mover (EM) distance **(C)** and average KS p-value **(D)** between real and simulated distributions, for the inferred network and without interactions. The dispersion corresponds to the first and ninth decile from ten simulations. **(E-F-G)** Two-dimensional UMAP representations of the original dataset **(E)** and the datasets simulated from the inferred network **(F)** and without interactions **(G)**. In these three plots, Sparc is removed from the genes represented in Figure 5 as its dynamics are not well captured by the mechanistic model, so the three datasets consist of 2449 cells with mRNA levels of 40 genes.

Discussion

The major interest of the method we proposed in this work is that it uses the same model for both inferring and simulating a gene regulatory network. The simulation part is not a novelty: a growing number of algorithms are proposed for simulating realistic single-cell gene expression datasets [38], and some have already been used for benchmarking GRN reconstruction methods [16, 17, 41].

Part of the success of our GRN model lies in its ability to reproduce the main characteristics of cell-cell heterogeneity observed in experimental datasets by making stochasticity an inherent part of the model, instead of adding an adapted noise a posteriori (see Methods). This is not the case of most algorithms that are used to simulate gene expression data associated to a regulatory network for benchmarking purposes [42, 43], even when they are based on underlying mechanistic models [16]. This point is also crucial for developing analytical results able to be used for the reverse engineering of the model, as it has been done for HARISSA and CARDAMOM.

Among the recently described algorithms, SERGIO [16] is the closest to our work. Nevertheless, we want to highlight some key differences:

1. SERGIO's mechanistic model is based on SDEs, treating noise as a Gaussian white noise. This is clearly insufficient to capture the *biological zeros* and Gamma-shaped variability, which in SERGIO arise only from the addition of technical noise. In our modeling scheme, these features arise naturally from the transcriptional bursting phenomenon.

2. As stressed by the authors, SERGIO seeks to simulate data with an explicit GRN as an input, a forward simulation goal, rather than attempt to estimate it from data, a reverse engineering goal. This is a fundamental difference with our work where we seek *simultaneously* to infer and to simulate data. To the best of our knowledge, the ability to do so using a mechanistic model is a true novelty of our work.

3. Our modeling scheme is amenable to in-depth mathematical analysis [6, 15, 23, 44].

4. The use of a specific module to add technical noise, as well as SERGIO's ability to generate both spliced and unspliced versions of mRNAs are welcome innovation that we will consider in future versions of our work.

We also mention that there has been a recent surge of interest in using generative adversarial networks (GANs) for producing realistic new single-cell transcriptomics data [45]. Although it can be an efficient strategy for data generation and augmentation, its behaves as a black box regarding the underlying biology. Our main added value here is that our model is based on the biophysical reality of the cell and provides a clear materialistic explanation for generated data. Since we have shown that this generative model can be calibrated from single-cell datasets, it can also be used for control purposes, aiming at controlling the cellular phenotype by interfering with the GRN behavior.

While the test case was made using a dataset obtained during a differentiation sequence, one should note that our approach can be applied to any biological process for which time-stamped single-cell transcriptomic data are obtained after applying a given stimulus. When

such time-stamped snapshots are not available, the algorithm could in principle take as an input time-reconstructed data (i.e., artificially ordered snapshots). In that case, the quality of the inference will strictly depend on the effectiveness of the time reconstruction algorithm.

Although efficient and promising, the model and the method we presented here have some limitations that are clearly identified, and that should guide future research efforts. First, we consider that the burst frequencies, which are critical parameters of the regulation, are sigmoid functions of protein levels. This implies that the distribution of a gene generally cannot have more than 2 modes [15], one associated to a low frequency of bursts and one to a high frequency of bursts, which correspond to the regions of the gene expression space where the sigmoid is relatively flat. Thus, if the distributions of some genes appear more complex than a mixture of these 2 modes, the model is not expected to reproduce accurately its dynamics, since some minor regulatory interactions should be not detected (in particular if the "hidden" mode is much closer to one of the two main modes than the other). This seems to be the case for example for the slight decrease of Sparc at t = 24h, that would have been better capture by adding an extra mode. To solve these kind of errors, it could be necessary to complexify the model, for example by modeling the burst rate functions by a multi-layer perceptron rather than a sigmoid as proposed in [15].

Second, CARDAMOM uses the temporality linearly, by taking into account the timepoints one after the other without possibility of backward step. This explains why the algorithm is unable to detect the known fact that a gene like Jarid2 inhibits some Extraembryonic and Neuroectoderm genes: indeed these inhibitions could only be detected by going up the arrow of time when the target genes see their expression increase, in order to find the real cause of this effect. Instead, we have seen that the algorithm interprets it as an activation of some other genes. We believe that it could be tackled by taking inspiration from the Recurrent Neural Network (RNN) theory, but it should be achieved while keeping the interpretability of the results. Note that we already developed a similar analogy for the regression step at each timepoint, which can be interpreted as the learning step of a perceptron [15].

From that point of view, if the information was indeed transmitted forward, the network would be supposed to be completed step by step until reaching its complete form. However, two types of incompatibilities may still occur:

1. A direct incompatibility occurs when an edge which has been inferred at a certain timepoint is chosen to be reset to a value close to 0 or even to change sign at another one.

2. An indirect incompatibility corresponds to the case where the effect of an edge $(j \to i)$, which has been inferred at a certain timepoint, is compensated by another edge $(k \to i)$ at a following timepoint, but that the gene k expression products were high enough at the previous timepoint to thwart the effect of the relation $(j \to i)$.

This explains why, for the experimental dataset presented in the Results section, the model is not able to reproduce the behavior of some genes. One good example is Pou5f1 between t = 0h and t = 24h: indeed, the edges that could generate the slight increase of Pou5f1 at t = 6h are thwarted by the edges inferred at the following timepoints which lead to the strong decrease of Pou5f1 after t = 48h. However, these incompatibilities could also have a biological meaning, and

be impossible to solve by modifying the regression problems. To go further, it is necessary to discuss the notions of structure and states of the network, which is related to the importance of possibly hidden variables. If the network incorporates all critical nodes, then the structure should not change. But it is different if there are some hidden variables, like genes the level of which are not measured, which results in modifying the network structure. For example, the problem of Pou5f1 that we have presented above could be explained in the following way: at t = 0h, an hidden variable may act on the interaction by preventing its possibility, before the hidden variable disappear at t = 24h. So if the hidden variable was integrated, then the network structure would not change anymore, but it is likely that we should in our case consider a modification of the network structure at t = 24h.

Third, the model does not allow the synthesis and degradation rates to vary over time. We have already mentioned that this was a problem for simulating the passage from t = 72h to t = 96h, and we decided to speed up the last time step for the model to reach its stationary distribution at 96h. We believe that most of the errors observed in the simulation with respect to the experimental dataset could be solved by finding an appropriate degradation rate. Thus, a significant improvement would consist in taking into account at each regression step the size of the time interval, and not only its order in the series of timepoints as it is now the case. This could allow to find a most-likely GRN in accordance to the degradation rate at each timepoint, or even to infer a most-likely degradation rate for each timepoint. Note however that if the latter case could both provide us a new information on the variation of degradation rates during differentiation and a better accuracy on the relative importance of the interactions in the GRN, it could also accentuate the problems of identifiability, and should therefore be studied carefully.

Fourth, the model does not take into account proliferation nor apoptosis while studying the stochasticity of the differentiation processes, nor the regulation of the proliferation rate by gene expression products. When sampling a distribution of n cells at a time t, the initial condition is built by sampling n cells under the uniform distribution among the set of cells at 0h, and to simulate its evolution during t hours. However, if some cells are supposed to have a higher death rate, and others to have a higher division rate, the process should evolve preferably in a certain direction in the gene expression space, which is going to be ignored in the current version of CARDAMOM. Taking into account these characteristics is a notoriously difficult task: a significant improvement has been recently achieved with Waddington-OT [27, 46], where a stochastic diffusion process models gene expression dynamics. Extending this kind of approach for the mechanistic model will be the subject of future works.

Finally, future versions of our method may consider additional biological features such as spatial cell-cell communication as the advent of multiomics datasets should provide data allowing to analyze the effect of these processes on differentiation, which is not possible in the case of scRNA-seq data without seriously compromising identifiability. We believe that the work presented here could serve as a basis for developing multiscale approaches to differentiation processes.

Methods

Mechanistic model of gene regulatory networks

The model used throughout this article is based on a hybrid version of the well-established two-state model of gene expression [6], where a gene is described by the state of a promoter, which can be either *on* or *off*. If the promoter is *on*, mRNAs are being transcribed at a rate s_0, which are then translated into proteins at a rate s_1. Degradation of both mRNAs and proteins occurs at a rate d_0 and d_1, respectively. The transitions between the on and off states occur at times of rates k_{on} and k_{off}. We consider the *bursty* regime of this model ($k_{\text{on}} \ll k_{\text{off}}$), corresponding to short active periods with high transcription rates, as experimentally observed [47–50]. In this regime, mRNA is then transcribed by bursts of tens to hundreds of molecules. The random times at which these bursts occur are still described by an exponential distribution of parameter k_{on}, and their random size by an exponential distribution with mean s_0/k_{off}. This model is compatible with experimental single-cell data, as steady-state mRNA levels follow for each gene a Gamma distribution, in line with continuous single-cell data [10].

The key idea is to incorporate this model into a network: the burst rate for each gene i is given by a gene-specific function $k^{\theta}_{\text{on},i}(P)$, where P is the vector of protein quantities (Figure S1). This function depends on proteins through a GRN, represented by an n-by-n matrix $\theta = (\theta_{ij})$ where n is the number of genes in the network. The value of $k^{\theta}_{\text{on},i}(P)$ then corresponds to the transcriptional burst frequency of gene i given protein levels P. Each parameter θ_{ij} encodes the interaction $j \to i$ with its direction, sign, and intensity. Recent work suggests that burst sizes are smaller and more uniform than previously anticipated [49] therefore leaving more room for burst frequency modulation [51] as a mechanism for gene expression regulation. We therefore consider that interactions come mainly from the modulation of burst frequencies $k^{\theta}_{\text{on},i}$ and that for any gene i, the rates $k_{\text{off},i}$ do not depend on P. The burst frequencies can be represented by sigmoid functions [23] as a simplification of the mechanistic form used in [6, 14]:

$$k^{\theta}_{\text{on},i}(P) = k_{0,i} + (k_{1,i} - k_{0,i}) \left(1 + \exp\left(-\beta_i - \sum_{j=1}^{n} \theta_{ij} P_j\right)\right)^{-1}$$

where $k_{0,i}$ (resp. $k_{1,i}$) is the minimal (resp. maximal) burst frequency of gene i and β_i is the basal activity of gene i, which can be also considered as the constant activity of a set of genes that are not measured but act on the network.

Simulation of time-stamped datasets

In order to simulate the mechanistic model, we used the simulation module of the HARISSA package [23]. One computational advantage of this method, which consists in sampling burst times with maximum rate and then deciding with an appropriate rule which ones to keep, is that it is guaranteed to be exact without requiring any numerical integration.

To simulate discrete "count" data that are produced by current scRNA-seq technologies, each mRNA level is generated by sampling from a Poisson distribution whose mean is the simulated expression level. The resulting cell profiles are exactly (resp. approximately) distributed according

to the discrete-valued "Gillespie" version of the mechanistic model in the absence (presence) of interactions between genes [52].

In order to reproduce *in vitro* experiments for a specific GRN, we use the following method: (1) let the model run for t $<$ 0h until its (stochastic) steady state is reached; (2) introduce at t = 0h a virtual *stimulus* gene with a constant maximal value for its protein. Such stimulus represents a perturbation in the environment of the cells, inducing them to evolve towards a new (stochastic) steady state. For example, in the case of mouse ES cell differentiation, this corresponds to the addition of all-trans RA in the medium. A time-stamped dataset corresponds to the sampling of independent cells at a specific sequence of timepoints (therefore "killing" sampled cells at each timepoint) starting from t = 0h. Namely, for the benchmark of Figure 2, the sequence of 10 timepoints was set to 0, 6, 12, 24, 36, 48, 60, 72, 84, and 96h.

Relevance to biological data

The exact probability distribution associated to the mechanistic model remains unknown for general networks. However, the analysis developed in [15] suggests that the marginal on mRNAs of the distribution at each time t can be reasonably approximated by a Gamma mixture:

$$M_t \sim \sum_{z \in Z} \mu_t(z) \prod_{i=1}^{n} \text{Gamma}\left(\frac{k_{z,i}}{d_{0,i}}, \frac{k_{\text{off},i}}{s_{0,i}}\right), \tag{1}$$

where Z denotes the set of cell types seen as the basins of attraction of the deterministic limit, $k_{z,j}$ denotes the mode of bursts frequency associated to a gene j within a basin $z \in Z$, and μ_t is a probability vector describing the relative weight of the basins in the process at time t.

The Poissonian layer transforms the Gamma distributions into negative binomial (NB) distributions, which gives:

$$M_t \sim \sum_{z \in Z} \mu_t(z) \prod_{i=1}^{n} \text{NB}\left(\frac{k_{z,i}}{d_{0,i}}, \frac{k_{\text{off},i}}{k_{\text{off},i} + s_{0,i}}\right). \tag{2}$$

Such mixture distributions are known to be compatible with continuous single-cell data [1, 10]. In particular, we recover the second order relationship between pairs of variables that are characteristic of experimental datasets. Indeed, we remark that the mean of a negative binomial distribution NB(a, b) is $m = \frac{a(1-b)}{b}$ and its variance $v = \frac{a(1-b)}{b^2}$, which implies that:

$$CV^2 = \frac{v}{m^2} = \frac{1}{a(1-b)} = \frac{1}{b}\frac{1}{m}.$$

Thus, for every gene i, by replacing b by $\frac{k_{\text{off},i}}{k_{\text{off},i}+s_{0,i}}$, we see that the relation $CV^2 \sim \frac{1}{m}$, which is characteristic of cell-cell heterogeneity in single-cell data is well verified by the mechanistic model, provided that $k_{\text{off},i}$ does not depend on the protein field. This also argues in favor of the assumption that a GRN does not affect significantly the bursts size. Note however that following this criteria, any model generating negative binomial distributions could be considered as realistic. Such criteria is therefore not sufficient for characterizing the accuracy of a model of gene expression, in particular when a synthetic noise well adapted is added to be in accordance with experimental datasets [16].

Tested algorithms

The six algorithms that we are going to use for the benchmark represent together the main categories of GRN inference methods presented in [7]:

- GENIE3 [18], which computes the regulatory network for each gene independently, using tree-based ensemble methods to predict the expression profile of each target gene from the profiles of all the other genes;
- PIDC [19], which infers undirected network using the notion of mutual information;
- SINCERITIES [20], which uses Granger causality after computing temporal changes in gene expression through the distance between two consecutive timepoints of the marginal distributions;
- SCRIBE [21], which is based on the notion of conditioned Restricted Directed Information and ideally needs real cells trajectories, which is unrealistic experimentally. We then pseudo-temporally order the time-stamped synthetic data used for the benchmark with two methods, one using a pseudotime algorithm and the other using an optimal coupling method with optimal transport, following the idea developed in [27]. We also tested with a dataset with real trajectories in order to compare the performances. The results of this algorithm are presented separately, due to the difference in the information that is needed.
- HARISSA [15] and CARDAMOM [23] which are based on the mechanistic model presented above, and both compute the network by solving a set of regression problems. They are nevertheless based on distinct mathematical analyses of the same model. HARISSA aims to solve a maximum likelihood problem on the protein distributions after computing a most-probable position for the protein concentration in each cell, CARDAMOM reconstructs the GRN by comparing the function k_{on} to the modes associated to a joint mRNA distribution previously inferred.

Measuring algorithm performance for the benchmark

We evaluated the GRN inference algorithms on simulated datasets using the area under the precision-recall curve (AUPR). Since inferring these coefficients is a notoriously difficult task [17], we do not take into account diagonal coefficients of the GRN matrix, which correspond to self-regulations. Note also that we chose precision-recall (PR) curves rather than receiver operating characteristic (ROC) curves because of the well-known class imbalance problem. Indeed, the sparsity hypothesis suggests that the number of interactions expected for a network of size n is smaller than half of the total number of possible interactions (n^2): it is then natural to focus on minimizing false positives (interactions that are detected but not present) rather than false negatives (interactions that are present but not detected), which explains the preference of PR over ROC.

Experimental dataset

We used data collected from a differentiation experiment of mouse embryonic stem cells induced by all-trans retinoic acid treatment [22]. This scRNA-seq dataset consists of 9 timepoints (0,

6, 12, 24, 36, 48, 60, 72, and 96h), each timepoint containing between 137 and 335 sampled cells after pre-processing (272 on average, for a total number of 2449 cells). To limit artificial correlations between genes (due to a multiplicative cell-specific technical factor mainly related to the reverse-transcription step), we selected cells with a total number of UMI counts ≥ 2000 in line with Semrau et al. [22], which resulted in keeping only 2449 out of 3456 measured cells.

On the other hand, we did not normalize cells by their respective total UMI counts and argue that this type of sample normalization is hazardous in the case of single-cell data. Indeed, such "library sizes" are small compared to bulk data (because here 1 sample = 1 cell) and are in fact biologically fluctuating, likely reflecting the transcriptional bursting phenomenon (this is easily seen when simulating "perfect" data from the mechanistic model, see Figure S2). In practice, since the CARDAMOM inference method starts with a binarization step (applying a specific, statistically derived threshold to each gene based on the mechanistic model), a multiplicative factor on each cell should not have too much impact as long as the number of cells is large enough. More generally, we argue that such normalization of cells should rather be "soft-coded" as a random factor to be estimated within a statistical framework.

The total number of genes measured in this experiment is 17452, which is much larger than in our benchmark. As they are unlikely to all be important in characterizing the differentiation process, we decided to restrict our analysis to a panel of 41 genes that had previously been identified as key marker genes for pluripotency, post-implantation epiblast, neuroectoderm and extraembryonic endoderm [22]. This number of genes allows to infer a network rich enough to make cell types emerge in a non-trivial way, while keeping a reasonable statistical power regarding the number of sampled cells. Note that the speed of the algorithms (Table S1) would allow a much larger subset of genes to be used: the limiting factor here is not computational speed but statistical power resulting from the number of cells available (see Figure 2).

Calibration of the mechanistic model

The principle of CARDAMOM is based on a two-step procedure :

- In a first step, we find the set of parameters α defining the mixture of negative binomial distributions (2) which fits well the data. The only parameter which is allowed to vary at each timepoint is the mixture parameters (allowing to estimate the values of the typical modes associated to the functions $k_{\text{on},i}$, for every gene i). Note that every parameters are computed except the degradation rates, which are constant for each gene and scale the dynamics of gene expression;

- In a second step, we calibrate the mechanistic model in order to obtain the distribution which is the closest to this Gamma mixture distribution. The network θ is then actualized at each timepoint in order to fit the mixture parameters.

The degradation rates are then fixed at the values that are found in tables from the literature [36]. Since many genes do not appear in these tables, we decide to set the same value for the degradation rates of all the genes belonging to the same functional group identified in [22] (see Table S2). The detail concerning these two steps can be found in [15]. Note that the first step has been modified since the original publication in order to replace the MCMC algorithm, which

was used to find the parameters of the negative binomial distributions associated to each cell type, by a simpler variational method.

As mentioned previously, the model cannot be in adequacy with both the dynamics from 0h to 72h and from 72h to 96h with the same degradation rates for the experimental dataset used in Results, due to the acceleration of the process between 72h and 96h. For this reason, we decided for our simulation to multiply degradation rates before the last timepoint by a factor $f = 6$, large enough to reach the steady state between 72h and 96h.

Comparing datasets using countsimQC

We used countsimQC [37] to compare the experimental dataset, the dataset simulated with the inferred network and the dataset simulated with the null network (Figure S2). Although there are no significant difference between the first two datasets, we observe in Figure S2F, that the correlations between genes are not perfectly reproduced (they are clearly more accurate than for the dataset simulated with the null network). This gap between the correlations between genes is also illustrated in Figure S4, which compares joint distributions between the simulated dataset (with the inferred network) and the experimental one for three pairs of genes at the final timepoint. We observe that if the global form of the correlation is respected, they are not as strong in the simulated dataset as in the experimental dataset. This suggests that the inferred GRN recovers the true correlation but not with the right intensity, which may be due to the sensitivity of model to the value of its parameters.

The fact that except for the correlations (sample-sample and feature-feature), the statistical characteristics explored by countsimQC are also well reproduced by the dataset simulated with the null network, suggests that they are generally not sufficient for measuring the accuracy of a dataset reproduction. This is partly due to the fact that any calibration of the model with the right scaling parameters but not the right GRN should matches most of these characteristics. In that meaning, the successes of simulation algorithms prior to our work are limited when measured with similar criteria. Our methodology, for which we used distinct criteria which are particularly well illustrated in Figure S5 and Figure 7, then appears as a significant improvement in the field of executable GRN inference.

Comparing datasets using UMAP

Since UMAP is not linear, the projections of datasets shown in Figure S5 are likely to force the projected data to be artificially close to the reference dataset. Thus, we decide to present two figures similar to Figure 7E-F-G, but where instead of projecting the simulated dataset on the pseudo-axis corresponding to the projection of the experimental dataset, we project both datasets together and show separately the cells corresponding to each dataset. Using this methodology, we represented in Figure S5A the UMAP projection of the experimental dataset and in Figure S5B the one of the dataset simulated with the inferred network. We did the same for the experimental dataset (Figure S5C) and the dataset simulated with the null network (Figure S5D). Then, although they have different representations, Figures Figure S5A and Figure S5C represent the same dataset, and the difference comes from the second dataset (the simulated one) with which the reduction has been performed. This allows to emphasize that the representation of a

distribution of cells with UMAP is very sensitive to the choice of the data that are integrated in the projection. Once again, we observed that the dataset simulated using the inferred network does seem much closer to the experimental dataset than the one simulated with the null network: in particular, Figure S5B demonstrates that the UMAP projection of the dataset with network is close to the one of the experimental dataset both in the arrangement of the cells between the different timepoints and in the general form of the subspace occupied by the cells, which is not the case for the UMAP projection of the dataset simulated with the null network, represented in Figure S5D.

Code availability

CARDAMOM is available at `https://gitbio.ens-lyon.fr/eventr01/cardamom`.
HARISSA is available at `https://github.com/ulysseherbach/harissa`.

Author contributions

Conceptualization, E.V., U.H., T.E., and O.G.; Formal Analysis, E.V., U.H., and G.B.; Funding Acquisition, O.G.; Investigation, All; Methodology, E.V., U.H., and T.E.; Software, E.V. and U.H.; Supervision, O.G. and T.E.; Validation, All; Visualization, E.V., U.H., and O.G.; Writing – Original Draft, E.V., G.B., and O.G.; Writing – Review and Editing, All.

Acknowledgments

This work was supported by funding from French agency ANR (SingleStatOmics; ANR-18-CE45-0023-03). We would like to thank especially Christophe Arpin, Thomas Lepoutre, Anton Crombach and Arnaud Bonnafoux for critical reading of the manuscript. We also thank all members of the SBDM and Dracula teams, and of the SingleStatOmics project, for providing such stimulating working environment. We finally thank the BioSyL Federation, the LabEx Ecofect (ANR-11-LABX-0048) and the LabEx Milyon of the University of Lyon for inspiring scientific events.

References

[1] J. C. Mar. “The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond”. In: *Biophysical Reviews* 11 (2019), pp. 89–94.

[2] A. F. Coskun, U. Eser, and S. Islam. “Cellular identity at the single-cell level”. In: *Mol Biosyst* 12 (2016), pp. 2965–2979.

[3] V. A. Huynh-Thu and G. Sanguinetti. “Gene Regulatory Network Inference: An Introductory Survey”. In: *Methods Mol Biol* 1883 (2019), pp. 1–23.

[4] C. Zong, L.-H. So, L. A Sepúlveda, S. O Skinner, and I. Golding. “Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene”. In: *Molecular systems biology* 6 (2010), p. 440.

[5] H. Ochiai, T. Sugawara, T. Sakuma, and T. Yamamoto. “Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells”. In: *Scientific reports* 4 (2014), pp. 1–9.

[6] U. Herbach, A. Bonnaffoux, T. Espinasse, and O. Gandrillon. “Inferring gene regulatory networks from single-cell data: a mechanistic approach”. In: *BMC Systems Biology* 11 (2017), p. 105.

[7] K. Akers and T.M. Murali. “Gene regulatory network inference in single cell biology”. In: *Current Opinion in Systems Biology* (2021).

[8] V. Shahrezaei and P. S. Swain. “Analytical distributions for stochastic gene expression”. In: *PNAS* 105 (2008), pp. 17256–17261.

[9] N. Friedman, L. Cai, and X S. Xie. “Linking stochastic dynamics to population distribution: an analytical framework of gene expression”. In: *Phys Rev Lett* 97 (2006).

[10] C. Albayrak, C. A. Jordi, C. Zechner, J. Lin, C. A. Bichsel, M. Khammash, and S. Tay. “Digital Quantification of Proteins and mRNA in Single Mammalian Cells”. In: *Molecular Cell* 61 (2016), pp. 914–924.

[11] Z. S. Singer, J. Yong, J. Tischler, J. A. Hackett, A. Altinok, M. A. Surani, L. Cai, and M. B. Elowitz. “Dynamic heterogeneity and DNA methylation in embryonic stem cells”. In: *Mol Cell* 55 (2014), pp. 319–31.

[12] G. Schiebinger. “Reconstructing developmental landscapes and trajectories from single-cell data”. In: *Current Opinion in Systems Biology* 27 (2021), p. 100351.

[13] L. Deconinck, R. Cannoodt, W. Saelens, B. Deplancke, and Y. Saeys. “Recent advances in trajectory inference from single-cell omics data”. In: *Current Opinion in Systems Biology* 27 (2021), p. 100344.

[14] A. Bonnaffoux, U. Herbach, A. Richard, A. Guillemin, S. Gonin-Giraud, P.-A. Gros, and O. Gandrillon. “WASABI: a dynamic iterative framework for gene regulatory network inference”. In: *BMC Bioinformatics* 20 (2019), p. 220.

[15] E. Ventre. “Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics”. In: *In Silico Biology* 14 (2021), pp. 89–113.

[16] P. Dibaeinia and S. Sinha. “SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks”. In: *Cell systems* (2020).

[17] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. M. Murali. “Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data”. In: *Nat Methods* 17 (2020), pp. 147–154.

[18] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. “Inferring regulatory networks from expression data using tree-based methods”. In: *PLOS One* 5 (2010).

[19] T. E Chan, M. P H Stumpf, and A. C Babtie. “Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures”. In: *Cell Systems* 5 (2017).

[20] N. Papili Gao, S. M. M. Ud-Dean, O. Gandrillon, and R. Gunawan. “SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles”. In: *Bioinformatics* (2017).

[21] X. Qiu, A. Rahimzamani, L. Wang, B. Ren, Q. Mao, T. Durham, J.L. McFaline-Figueroa, L. Saunders, C. Trapnell, and S. Kannan. “Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe”. In: *Cell Systems* 10 (2020), pp. 1–10.

[22] S. Semrau, J. E. Goldmann, M. Soumillon, T. S. Mikkelsen, R. Jaenisch, and A. van Oudenaarden. “Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells”. In: *Nat Commun* 8 (2017), pp. 1–16.

[23] U. Herbach. “Gene regulatory network inference from single-cell data using a self-consistent proteomic field”. In: *arXiv* 2109.14888 (2021).

[24] A. Richard, L. Boullu, U. Herbach, A. Bonnafoux, V. Morin, E. Vallin, A. Guillemin, N. Papili Gao, R. Gunawan, J. Cosette, O. Arnaud, J. J. Kupiec, T. Espinasse, S. Gonin-Giraud, and O. Gandrillon. "Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process". In: *PLoS Biol* 14 (2016), e1002585.

[25] P.S. Stumpf, R.C.G. Smith, M. Lenz, A. Schuppert, F.-J. Müller, A Babtie, T.E. Chan, M.P.H. Stumpf, C.P. Please, S.D. Howison, F. Arai, and B.D. MacArthur. "Stem Cell Differentiation as a Non-Markov Stochastic Process". In: *Cell Systems* 5 (2017), pp. 268–282.

[26] N. E. Phillips, A. Mandic, S. Omidi, F. Naef, and D. M. Suter. "Memory and relatedness of transcriptional activity in mammalian cell lineages". In: *Nature Communications* 10 (2019).

[27] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. "Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming". In: *Cell* 176 (2019), 928–943 e22.

[28] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC Genomics* 19 (2018), p. 477.

[29] E. Moutier, T. Ye, M. A. Choukrallah, S. Urban, J. Osz, A. Chatagnon, L. Delacroix, D. Langer, N. Rochel, D. Moras, G. Benoit, and I. Davidson. "Retinoic acid receptors recognize the mouse genome through binding elements with diverse spacing and topology". In: *J Biol Chem* 287 (2012), pp. 26328–41.

[30] A. Chatagnon, P. Veber, V. Morin, J. Bedo, G. Triqueneaux, M. Semon, V. Laudet, F. d'Alche-Buc, and G. Benoit. "RAR/RXR binding dynamics distinguish pluripotency from differentiation associated cis-regulatory elements". In: *Nucleic Acids Res* (2015).

[31] E. H. Z. Chua, S. Yasar, and N. Harmston. "The importance of considering regulatory domains in genome-wide analyses - the nearest gene is often wrong!" In: *Biol Open* 11.4 (2022).

[32] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells". In: *Cell* 133 (2008), pp. 1106–17.

[33] G. Li, R. Margueron, M. Ku, P. Chambon, B. E Bernstein, and D. Reinberg. "Jarid2 and PRC2, partners in regulating gene expression". In: *Genes & development* 24 (2010), pp. 368–380.

[34] S. Mahony, E. O Mazzoni, S. McCuine, R. A Young, H. Wichterle, and D. K Gifford. "Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis". In: *Genome biology* 12 (2011), pp. 1–15.

[35] C. Hrabchak, M. Ringuette, and K. Woodhouse. "Recombinant mouse SPARC promotes parietal endoderm differentiation and cardiomyogenesis in embryoid bodies". In: *Biochemistry and Cell Biology* 86 (2008), pp. 487–499.

[36] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. "Global quantification of mammalian gene expression control". In: *Nature* 473 (2011), pp. 337–42.

[37] C. Soneson and M. D Robinson. "Towards unified quality verification of synthetic count data with countsimQC". In: *Bioinformatics* 34 (2018), pp. 691–692.

[38] H. L. Crowell, S. X. M. Leonardo, C. Soneson, and M. D. Robinson. "Built on sand: the shaky foundations of simulating single-cell RNA sequencing data". In: *bioRxiv* (2021).

[39] K. S. Manning and T. A. Cooper. "The roles of RNA processing in translating genotype to phenotype". In: *Nat Rev Mol Cell Biol* 18 (2017), pp. 102–114.

[40] L. McInnes, J. Healy, and J. Melville. “UMAP: uniform manifold approximation and projection for dimension reduction”. In: *arXiv* 1802.03426 (2020).

[41] R. Cannoodt, W. Saelens, L. Deconinck, and Y. Saeys. “Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells”. In: *Nature Communications* 12.1 (2021).

[42] A. Mizeranschi, H. Zheng, P. Thompson, and W. Dubitzky. “Evaluating a common semi-mechanistic mathematical model of gene-regulatory networks”. In: *BMC Systems Biology* 9 (2015), pp. 1–12.

[43] H. Matsumoto, H. Kiryu, C. Furusawa, M. S. Ko, S. B. Ko, N. Gouda, T. Hayashi, and I. Nikaido. “SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation”. In: *Bioinformatics* (2017).

[44] E. Ventre, T. Espinasse, C.-E. Bréhier, V. Calvez, T. Lepoutre, and O. Gandrillon. “Reduction of a stochastic model of gene expression: Lagrangian dynamics gives access to basins of attraction as cell types and metastabilty”. In: *Journal of Mathematical Biology* 83 (2021), pp. 1–63.

[45] M. Marouf, P. Machart, V. Bansal, C. Kilian, D. S. Magruder, C. F. Krebs, and S. Bonn. “Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks”. In: *Nat Commun* 11 (2020), p. 166.

[46] H. Lavenant, S. Zhang, Y.-H. Kim, and G. Schiebinger. “Towards a mathematical theory of trajectory inference”. In: *arXiv preprint arXiv:2102.09204* (2021).

[47] D. M Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. “Mammalian genes are transcribed with widely different bursting kinetics”. In: *Science* 332 (2011).

[48] D. Nicolas, N. E. Phillips, and F. Naef. “What shapes eukaryotic transcriptional bursting?” In: *Mol Biosyst* 13 (2017), pp. 1280–1290.

[49] J. Rodriguez and D. R. Larson. “Transcription in Living Cells: Molecular Mechanisms of Bursting”. In: *Annu Rev Biochem* 89 (2020), pp. 189–212.

[50] E. Tunnacliffe and J. R. Chubb. “What Is a Transcriptional Burst?” In: *Trends Genet* 36 (2020), pp. 288–297.

[51] C. Li, F. Cesbron, M. Oehler, M. Brunner, and T. Hofer. “Frequency Modulation of Transcriptional Bursting Enables Sensitive and Rapid Gene Regulation”. In: *Cell Syst* (2018).

[52] U. Herbach. “Stochastic gene expression with a multistate promoter: breaking down exact distributions”. In: *SIAM Journal on Applied Mathematics* 79 (2019), pp. 1007–1029.

Supplementary information

Table S1: Related to Figure 2. Average runtime for inferring a network from datasets simulated with tree-like networks for which the results of the inference are represented in Figure 2, for the six algorithms that are used in the benchmark. Timings measured on a 16-GB RAM, 2.4 GHz Intel Core i5 computer.

Runtime	PIDC	SINCERITIES	HARISSA	CARDAMOM	GENIE3	SCRIBE
5 genes	0.02 s	0.01 s	0.16 s	0.24 s	9.02 s	34.36 s
10 genes	0.04 s	0.03 s	0.28 s	0.64 s	17.27 s	136.51 s
20 genes	0.06 s	0.06 s	0.45 s	1.70 s	31.17 s	> 5 min
50 genes	0.18 s	0.24 s	0.96 s	7.87 s	83.18 s	> 1 h
100 genes	0.75 s	0.77 s	1.80 s	18.57 s	159.42 s	> 3 h

Table S2: Related to Figure 7. Numerical values of mRNA and protein degradation rates (in h^{-1}) used for data simulation (sim.), compared with experimental (exp.) measures from the literature. Group abbreviations: Pluri = Pluripotency, Epi = Post-implantation epiblast, Neuro = Neuroectoderm, Endo = Extraembryonic endoderm.

Gene	Sox2	Zfp42	Klf2	Dnmt3	Cdh2	Sparc	Col4a2	Lamb1
Group	Pluri			Epi	Neuro	Endo		
mRNA (exp.)	0.044	0.053	0.067	-	0.015	0.012	0.012	0.015
mRNA (sim.)	0.075			0.2	0.03	0.005		
Protein (exp.)	0.0073	0.05	0.0027	0.023	0.022	0.17	0.078	0.038
Protein (sim.)	0.0075			0.02	0.02	0.1		

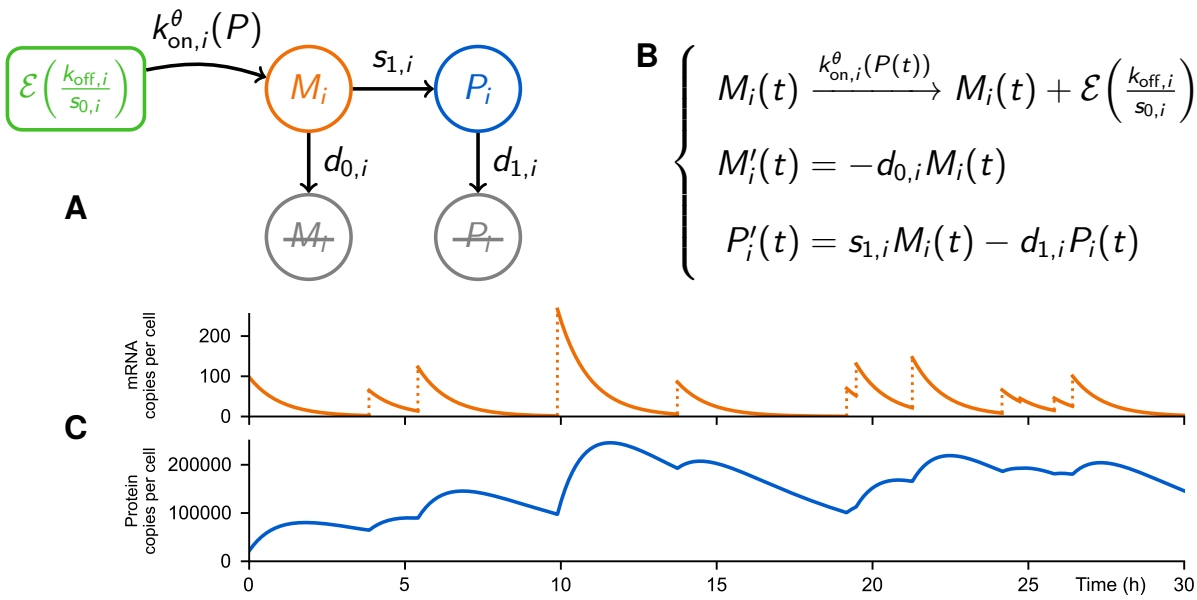

Figure S1: Related to Figure 1. Graphical (**A**) and mathematical (**B**) descriptions of the mechanistic model for the dynamics of a gene i. (**C**) Bursts of mRNA occur at random times with rate $k_{\text{on},i}$ and their size follows an exponential distribution $\mathcal{E}(k_{\text{off},i}/s_{0,i})$. The variables M_i and P_i describe respectively the mRNA and protein quantities associated to gene i in the cell. The vector of protein levels is denoted by $P = (P_1, \cdots, P_n)$ while θ denotes the GRN which couples the genes together through functions $k_{\text{on},i}$.

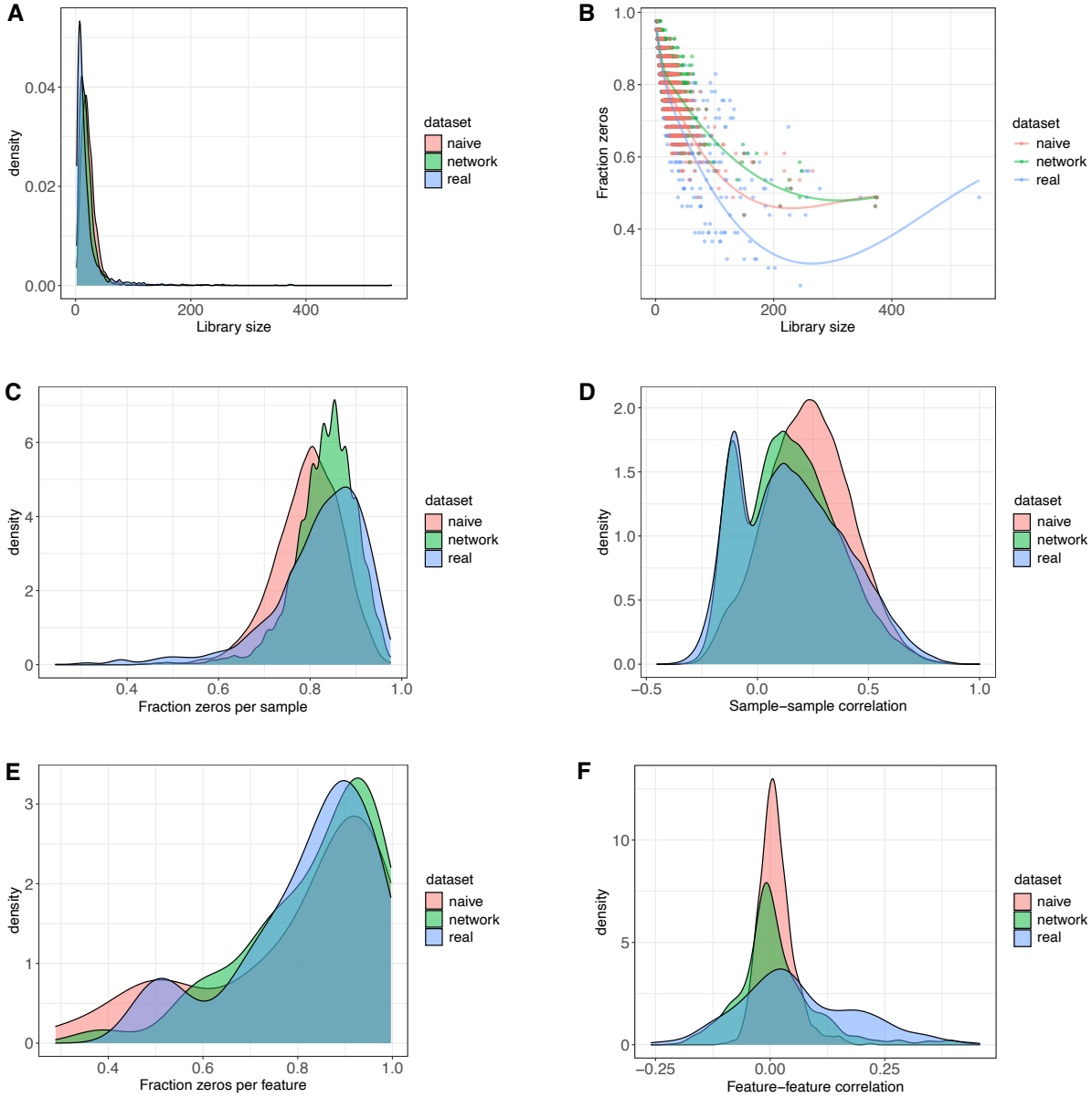

Figure S2: Related to Figure 7. Comparison of various statistical characteristics across datasets using the countsimQC package. Each plot shows the experimental dataset (*real*, in blue) and datasets simulated from the mechanistic model calibrated by CARDAMOM, including interactions (*network*, in green) and without interactions (*naive*, in red). Each dataset consists of 41 genes (*features*) measured in 2433 single cells (*samples*). (**A**) Distribution of "library sizes", defined as the total read count in each sample. (**B**) Association between the library size and the fraction of zeros observed per cell. (**C**) Distribution of the fraction of zeros observed per cell. (**D**) Distribution of cell-cell correlations, based on random cell pairings. (**E**) Distribution of the fraction of zeros observed per gene. (**F**) Distribution of gene-gene correlations, based on all possible gene pairings. Notably, the cell-cell correlation (**D**) bimodal pattern shows two possible pairings of cells: pairs with similar expression profiles (same genes *on*, same genes *off*) and therefore positively correlated, and pairs with opposite, "antinomic" profiles and therefore negatively correlated. This pattern is an indirect sign of the emergence of different cell types, a characteristic that is clearly not reproduced in the absence of interactions between genes (naive dataset).

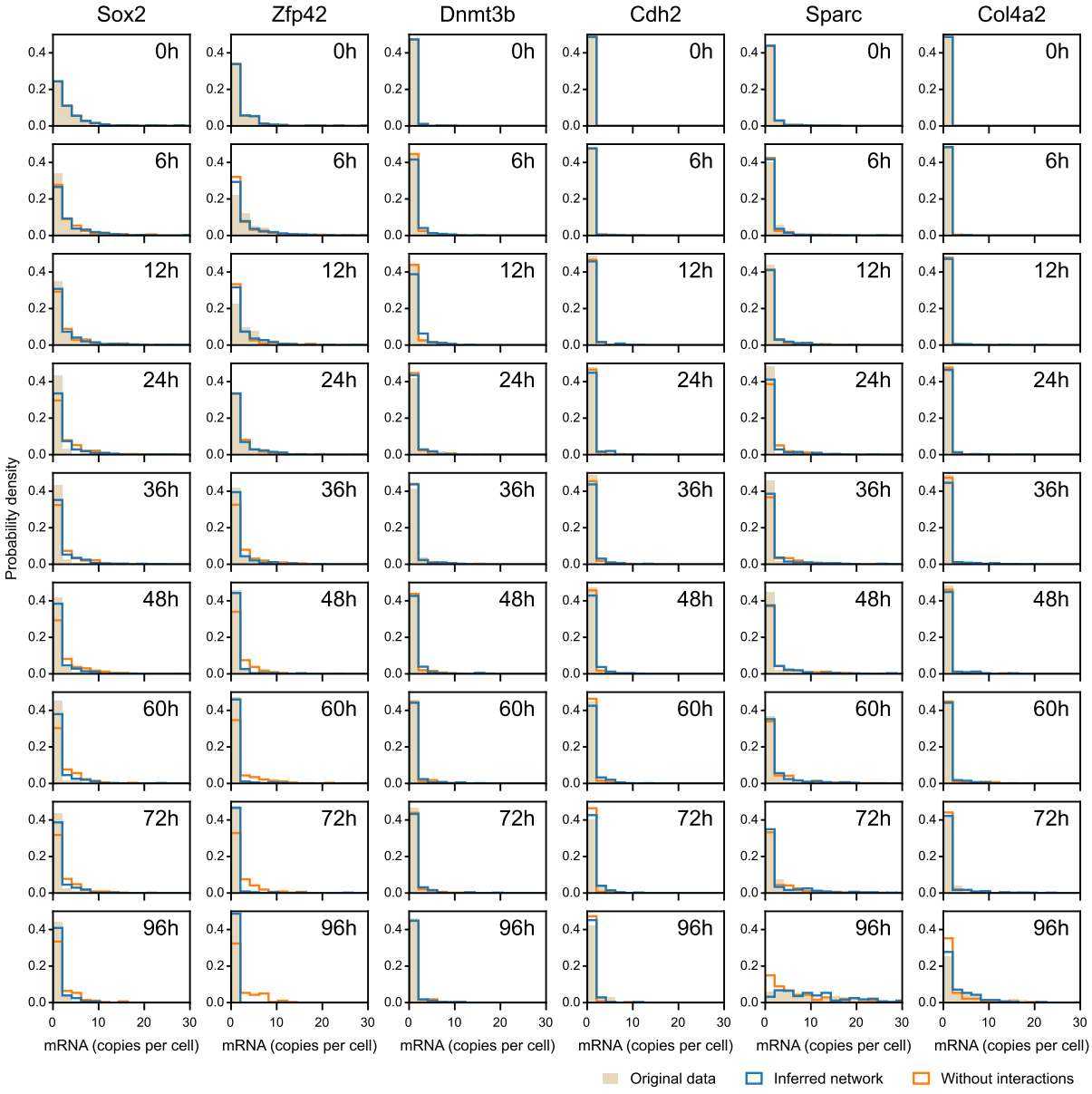

Figure S3: Related to Figure 7. Comparison between empirical distributions along timepoints, for six genes that have been found to play a key role in the regulation of the process (as visible in Figure 5). The experimental dataset (in beige), the dataset simulated from the inferred network (in blue) and the dataset simulated without interactions (in orange) correspond to Figure 7E, F and G, respectively.

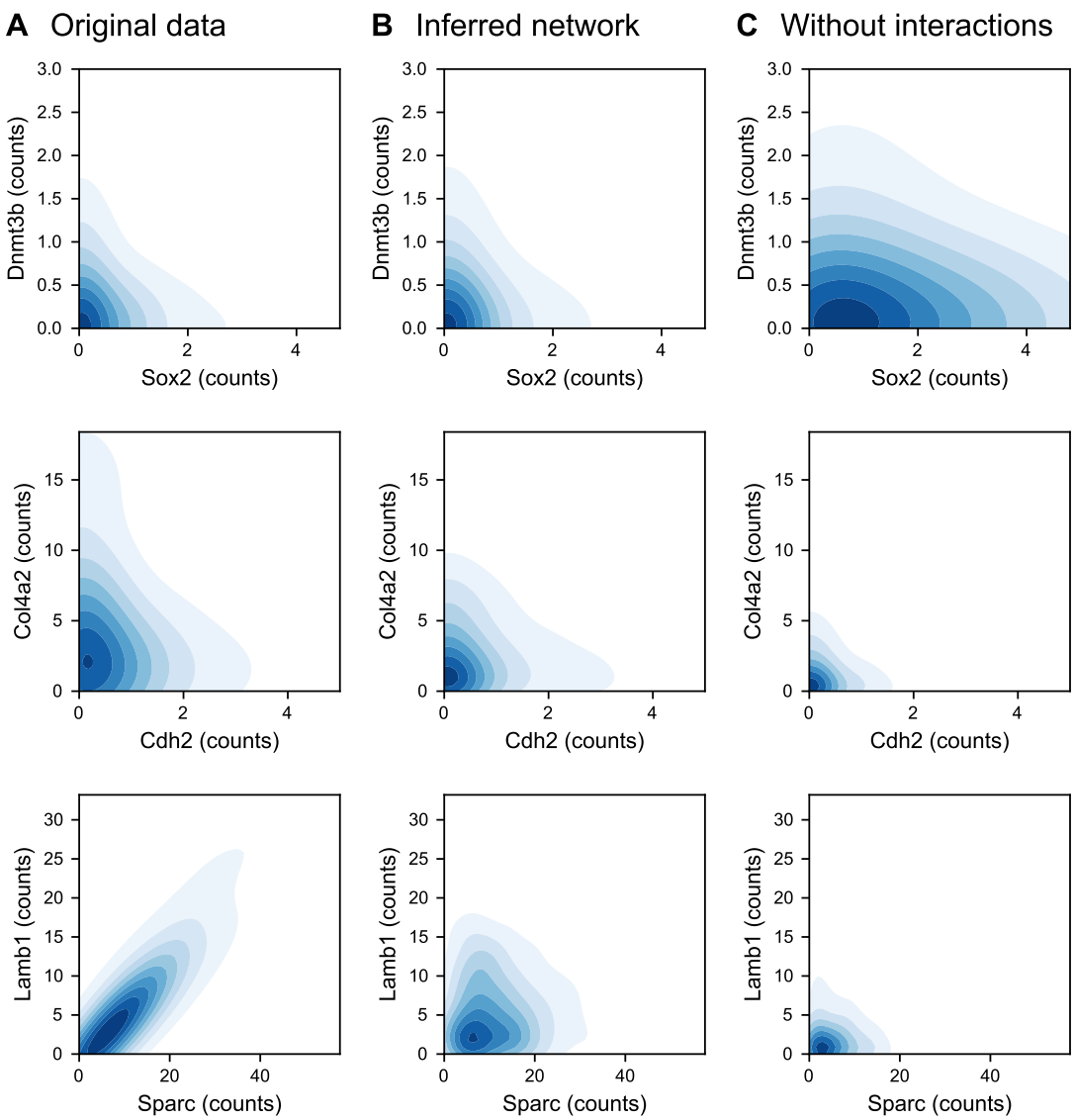

Figure S4: Related to Figure 7. Comparison of the joint distributions of three pairs of genes at the final timepoint between the experimental dataset (**A**) compared to the dataset simulated when the mechanistic model is calibrated by CARDAMOM (**B**) and the dataset simulated without interactions (**C**). The genes in each pair are expected to have interactions (direct or indirect) in the network represented in Figure 5.

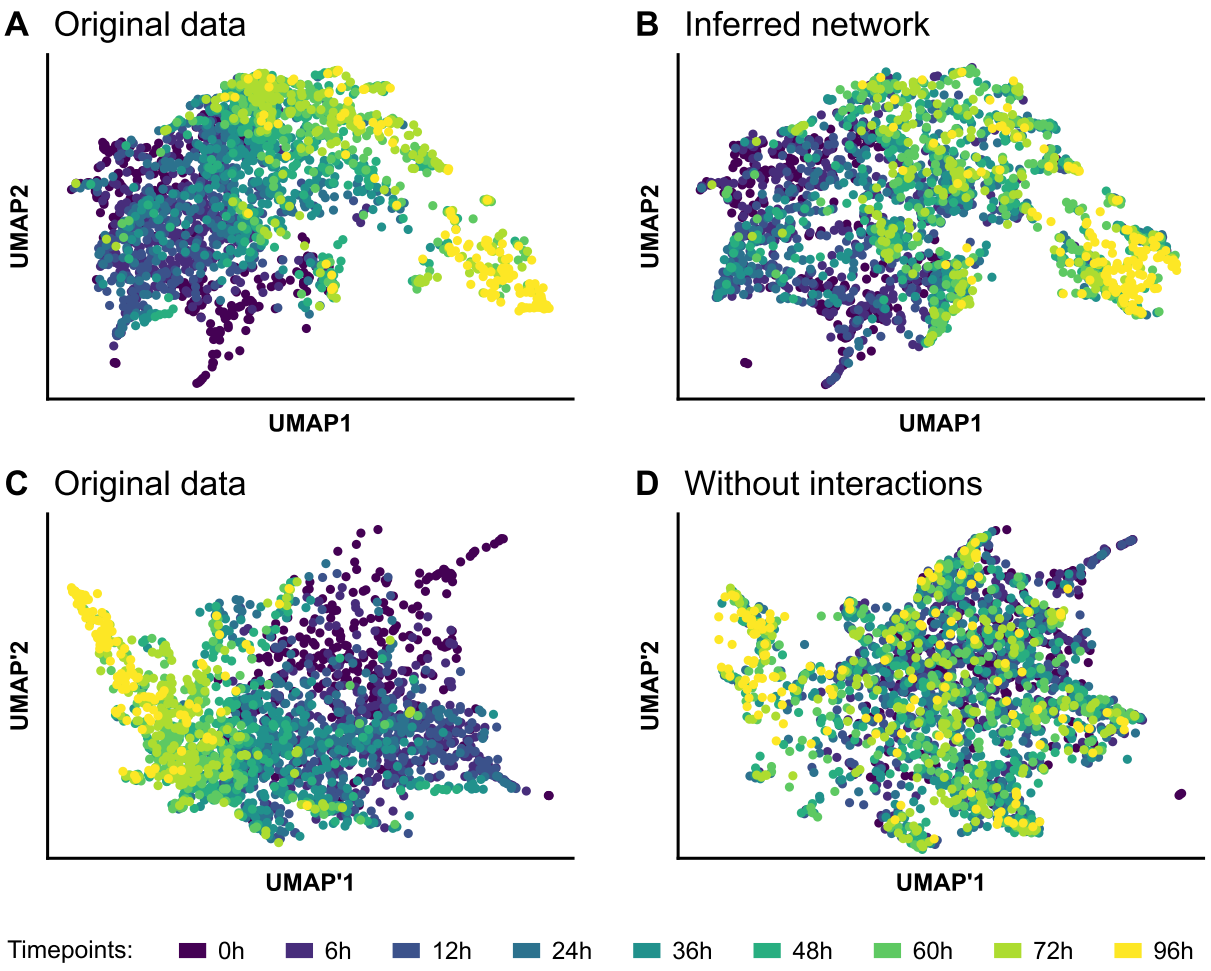

Figure S5: Related to Figure 7. Two-dimensional UMAP representations of the experimental dataset (*original data*) and datasets simulated after calibrating the mechanistic model with CARDAMOM, one including interactions (*inferred network*) and one obtained after setting $\theta = 0$ (*without interactions*). The four plots are based on two different projections, computed after merging the experimental dataset with (**A-B**) the simulated dataset including interactions or (**C-D**) the simulated dataset without interactions. Hence **A** and **C** are two representations of the exact same data, while **B** is to be compared with **A**, and **D** is to be compared with **C**.