



HAL
open science

scBoolSeq: Linking scRNA-Seq Statistics and Boolean Dynamics

Gustavo Magaña López, Laurence Calzone, Andrei Zinovyev, Loïc Paulevé

► **To cite this version:**

Gustavo Magaña López, Laurence Calzone, Andrei Zinovyev, Loïc Paulevé. scBoolSeq: Linking scRNA-Seq Statistics and Boolean Dynamics. 2023. hal-04294917v1

HAL Id: hal-04294917

<https://hal.science/hal-04294917v1>

Preprint submitted on 20 Nov 2023 (v1), last revised 29 Aug 2024 (v2)





HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

SCBOOLSEQ: Linking scRNA-Seq Statistics and Boolean Dynamics

Gustavo Magaña López¹ , Laurence Calzone^{2,3,4} , Andrei Zinovyev⁵ , Loïc Paulevé^{1*} 

1 Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

2 Institut Curie, Université PSL, F-75005, Paris, France

3 INSERM, U900, F-75005, Paris, France

4 Mines ParisTech, Université PSL, F-75005, Paris, France

5 In silico R&D, Evotec, Toulouse, France

* loic.pauleve@labri.fr

Abstract

Boolean networks are largely employed to model the qualitative dynamics of cell fate processes by describing the change of binary activation states of genes and transcription factors with time. Being able to bridge such qualitative states with quantitative measurements of gene expressions in cells, as scRNA-Seq, is a cornerstone for data-driven model construction and validation. On one hand, scRNA-Seq binarisation is a key step for inferring and validating Boolean models. On the other hand, the generation of synthetic scRNA-Seq data from baseline Boolean models provides an important asset to benchmark inference methods. However, linking characteristics of scRNA-Seq datasets, including dropout events, with Boolean states is a challenging task.

We present SCBOOLSEQ, a method for the bidirectional linking of scRNA-Seq data and Boolean activation state of genes. Given a reference scRNA-Seq dataset, SCBOOLSEQ computes statistical criteria to classify the empirical gene pseudocount distributions as either unimodal, bimodal, or zero-inflated, and fit a probabilistic model of dropouts, with gene-dependent parameters. From these learnt distributions, SCBOOLSEQ can perform both binarisation of scRNA-Seq datasets, and generate synthetic scRNA-Seq datasets from Boolean trajectories, as issued from Boolean networks, using biased sampling and dropout simulation. We present a case study demonstrating the application of SCBOOLSEQ's binarisation scheme in data-driven model inference. Furthermore, we compare synthetic scRNA-Seq data generated by SCBOOLSEQ with BOOLODE from the same Boolean Network model. The comparison shows that our method better reproduces the statistics of real scRNA-Seq datasets, such as the mean-variance and mean-dropout relationships while exhibiting clearly defined trajectories in a two-dimensional projection of the data.

Author summary

The qualitative and logical modeling of cell dynamics has brought precious insight on gene regulatory mechanisms that drive cellular differentiation and fate decisions by predicting cellular trajectories and mutations for their control. However, the design and validation of these models is impeded by the quantitative nature of experimental measurements of cellular states. In this paper, we provide and assess a new

methodology, SCBOOLSEQ for bridging single-cell level pseudocounts of RNA transcripts with Boolean classification of gene activity levels. Our method, implemented as a Python package, enables both to *binarise* scRNA-Seq data in order to match quantitative measurements with states of logicals models, and to generate synthetic data from Boolean trajectories in order to benchmark inference methods. We show that SCBOOLSEQ accurately captures main statistical features of scRNA-Seq data, including measurement dropouts, improving significantly the state of the art. Overall, scBoolSeq brings a statistically-grounded method for enabling the inference and validation of qualitative models from scRNA-Seq data.

Introduction

Unveiling the mechanisms that regulate cellular decisions is a central task in systems biology. For instance, numerous efforts have been conducted to elucidate the core mechanisms that control differentiation and cell fate decision processes such as osteogenesis [1–3], haematopoiesis [4–7], dopaminergic neuron differentiation [8], early retinal development [9], and various cancer types [10–13].

The advent of single-cell RNA sequencing (scRNA-Seq) technologies has greatly enhanced the resolution with which these dynamic phenomena can be studied. As a preliminary step, most studies first determine cell identities via either clustering and subsequent manual annotation or via the direct classification of cells [14]. Furthermore, trajectory reconstruction methods [15–17] allow visualising and hypothesising how gradual changes in gene expression eventually lead to commitment to specific lineages and phenotypes. A tremendous challenge is then to identify regulatory mechanisms that control the identified dynamics of expression patterns and ultimately phenotypes.

Boolean networks are widely employed to model cellular differentiation [18–21] and fate decision [22, 23]. In these models, the activity of biological entities is represented as either active or inactive. This coarse-grained view of gene expression levels helps counter the varying levels of technical noise caused by sequencing technologies. The binary representation allows reasoning on the causal relationships between entities without having to estimate kinetic parameters or regulation thresholds, while ensuring consistency with underlying quantitative models [24]. Boolean models can predict trajectories and conclude on the impossibility of certain behaviours, optionally subject to mutations, and can encompass thousands of genes. They revealed to be a powerful and relevant modelling approach to predict combinations of genetic perturbations to control cell fate decision [25, 26].

Nevertheless, linking qualitative gene activation states with their quantitative measurements, such as count of RNA transcripts, is a delicate task with high stakes for Boolean modelling. We present SCBOOLSEQ, which, given a reference dataset, provides a bidirectional link between scRNA-Seq and Boolean activation states.

The binary coarse-graining of scRNA-Seq, we refer to as *binarisation*, consists in assigning a qualitative active or inactive state to a gene, from one single-cell or a pool of single-cell measurements. The pools of cells usually correspond to phenotypes and other important cellular states. As Boolean models aim at predicting stability and trajectories between such cellular states, binarised data are crucial to assess their fitness with trajectories and steady states. One can easily note that the binary classification may be irrelevant in some cases, e.g., when in intermediate activation levels, or because of lacking statistical support. Therefore, it is important that binarisation methods actually result in three possible outcomes of the gene state: activate, inactive, or undetermined. However, numerous methods fully binarise transcriptome data with no regard for uncertainty or intermediate expression and the diversity of empirical pseudocount distributions [27]. REFBOOL [28] provided an important effort for quantifying statistical

uncertainty for the binarisation and allowing intermediate states. Their approach aims at exploiting a user-defined gene expression library which serves as a proxy to take into consideration the context of the global gene expression landscape when coarse-graining data. Unfortunately this approach is only available for bulk RNA-Seq data.

The inverse operation of binarisation consists in generating RNA pseudocounts from Boolean activation states. Coupled with simulations of Boolean models, this enables generating synthetic datasets from Boolean models subject to ranges of combinations of perturbations, simulating gene knock-out or constitutive activation, for instance. Resulting synthetic scRNA-Seq data can then serve as a basis to evaluate inference methods, such as gene regulatory networks inference, trajectory inference, and Boolean model inference.

Generating single-cell and bulk RNA-Seq data has been addressed by count simulators [29–31]. With different underlying assumptions, count simulators reproduce the statistical characteristics of real datasets via parametric and semi-parametric approaches. They are capable of simulating a wide variety of scenarios and even batch effects, but generally fail at integrating information from GRN known a priori. Efforts have been made to integrate knowledge about GRNs into count simulators [32]. However, this method requires the GRN to be a directed acyclic graph, which might not be the case in general. Alternative methods rely on translating Boolean networks into non-linear Ordinary Differential Equations (ODEs). A first work in this line was ODEFY which presented a canonical way of transforming Boolean into continuous models [33]. More recently, BOOLODE was presented in the context of GRN inference method benchmarking [34,35], introducing the addition of noise terms to make the ODEs stochastic. By building on top of Boolean networks, these approaches enable to capture the logical and dynamical relationships among the regulators. BOOLODE uses Hill functions to reflect the modulation of gene expression [36–38]. However, this approach relies on a considerable amount of parameters such as mRNA transcription and degradation rates, Hill thresholds and coefficients, signalling timescales, and interaction strengths. Determining these parameters is an important bottleneck as they can hardly be estimated from experimental scRNA-Seq data and need therefore to be set arbitrarily or randomly sampled. Moreover, these ODE-based generators fail to produce data with statistical properties comparable to those of real scRNA-Seq datasets.

We believe it is crucial that generated count data resemble as much as possible scRNA-Seq data to obtain fair inference benchmarks, which implies mimicking dropouts and other statistical features. SCBOOLSEQ relies on the learning of gene-wise RNA pseudocount statistics from a reference dataset. This learning is performed in three steps: (i) the classification of empirical gene pseudocount distributions; (ii) the use of Gaussian Mixtures with up to two components as a parametric model; and (iii) the simulation of dropout events with probabilities that are inversely proportional to the expression value. SCBOOLSEQ requires the reference dataset to be constituted of only highly variable genes (HVGs). Functions to perform this filtering are available on major scRNA-Seq analysis distributions such as STREAM [15] and SCANPY [17]. By selecting HVGs after quality control, normalisation, and batch correction, one ensures that SCBOOLSEQ’s reference reflects the underlying biological variation rather than technical noise. In addition to HGVs which are automatically selected by the designated functions in scRNA-Seq analysis environments, differentially expressed genes (DEGs) and known markers can also be incorporated to SCBOOLSEQ’s reference in order to have a fuller image of the transcriptional landscape of the dynamic phenomenon of interest.

Thus, from the preprocessed reference dataset, SCBOOLSEQ is able to perform two distinct complementary operations: the binarisation of a scRNA-Seq dataset with respect to the reference dataset, and the generation of synthetic scRNA-Seq from Boolean activation states, as illustrated by Fig. 1.

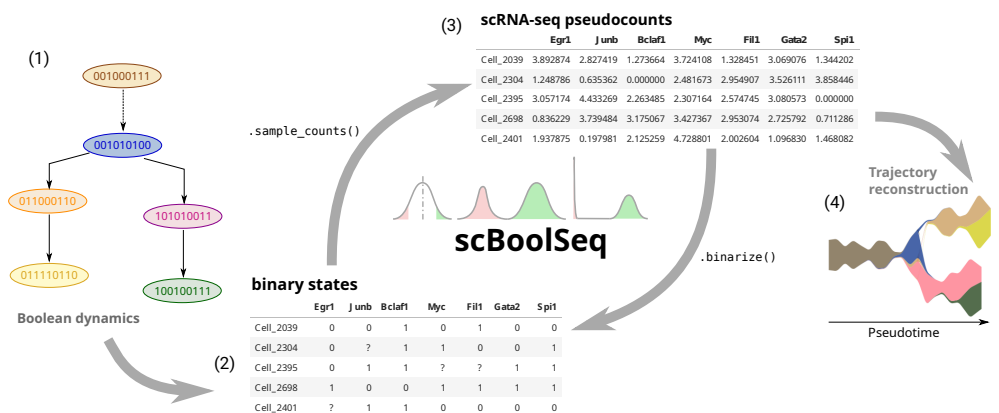


Fig 1. From left to right: (1) A branching trajectory constructed by merging two Boolean simulations, each one leading to a different stable state. (2) A binarised expression matrix, having genes as columns and samples as rows. (3) A pseudocount matrix (same format as the Boolean matrix). (4) A STREAM-plot reconstructing the branching trajectory from synthetic data generated from the Boolean traces [15]. SCBOOLSEQ can be used to go from gene expression matrices (such as 3) to Boolean matrices and vice-versa.

We first show that our 3-distribution model of scRNA-Seq counts and dropouts is able to accurately reproduce the statistical characteristics of a range of scRNA-Seq datasets. For the binarisation of scRNA-Seq data, we first apply our method to a publicly available scRNA-Seq dataset of early retinogenesis. We show that SCBOOLSEQ correctly identifies the different cell types described in the original study, defined by a minimal set of marker genes. These identities can subsequently be used in order to label cell groups found by the louvain clustering algorithm [39]. Going beyond cell type identification, we use the Boolean gene activity values determined by SCBOOLSEQ in order to prune a mouse regulon database [40]. The resulting GRN is validated via Gene Set Enrichment Analysis performed using METASCAPE [41] which yielded numerous relevant Gene Ontology terms related to the kept genes.

Finally, we show that SCBOOLSEQ's synthetic scRNA-Seq data generated from Boolean traces produces both discernible trajectories when applying dimensionality reduction techniques and statistics that comparable to those of real datasets.

Overall, SCBOOLSEQ provides an efficient method to learn statistics of a scRNA-Seq dataset and derive binarisation and synthetic generation procedures with few parameters. SCBOOLSEQ has been implemented as an open source Python package available at github.com/bnediction/scBoolSeq.

Results

In the following, we assume that scRNA-Seq data is preprocessed as log pseudocounts $x_{c,g} = \log(x_{c,g}^{\text{norm}} + 1)$, where c and g refer to cells and genes, respectively. Any size-factor based normalisation can be used, as long as it is of the form $x_{c,g}^{\text{norm}} = \frac{x_{c,g}^{\text{raw}}}{\alpha}$ where α is a constant. For instance $\alpha_c = \sum_g x_{c,g}$ would represent the standard library size normalisation, yielding counts/reads per million (CPM/RPM). Our methodology is applicable to alternative normalisations such as TPM (transcripts per kilobase per million reads) or RPKM (reads per kilobase per million reads). The log transformation is necessary in order to ensure the validity of the underlying parametric distributions.

Classification of Pseudocount Distributions and Dropout Model

scBOOLSEQ builds on the ideas presented in [42] which seek to capture the different expression patterns across bulk RNA-Seq samples of cancer patients. By computing a series of statistical criteria, they proposed to classify empirical pseudocount distributions as bimodal, zero-inflated, or unimodal. This choice of distributions reflects the underlying hypotheses of gene activity: bimodal genes exhibit two distinct expression patterns for the absence and presence of their corresponding encoded proteins. For unimodal genes, we suppose that only cells lying at tails of the distribution can be confidently inferred to be active or inactive. It also appeared that several genes show a high proportion of zeros, which are then classified as zero-inflated. Their classification method employs statistics such as mean, median, variance, dropout rate, amplitude, dip test's p-value [43], kurtosis, density peak, and Bimodality Index [44]. In a first step, genes which do not exhibit a high enough variability or have excessive dropout rates are filtered out. Then, bimodal patterns are searched within kept genes, using a combination of statistics. Afterwards, genes with no bimodal patterns are tested for zero-inflation by looking at the empirical distributions' density peaks. Remaining genes are classified as unimodal.

With scBOOLSEQ, we generalized and improved this approach to account for the specificities of scRNA-Seq data, notably their potential high dropout in gene counts, and to enable the sampling of count for reconstructed distributions in order to generate synthetic scRNA-Seq datasets from Boolean activation states. As we illustrate in S2 Fig, when applied to scRNA-Seq, the PROFILE classification algorithms show two shortcomings: (1) for genes classified as bimodal and unimodal, the dropout tends to artificially decrease their mean and inflate their variance, impeding a good characterisation of their empirical pseudocount distributions via Gaussian or two-component Gaussian Mixtures; (2) for zero-inflated genes, the classification does not result in a parametric distribution, which complicates sampling. We improved the algorithm by computing the statistics on non-zero data and propose a novel probabilistic model for dropouts in order to capture the proportion of zeros. By modelling the probability of a dropout occurring as a function of the expression level with gene-dependent parameters, we were able to reproduce the per-gene dropout rates of different reference datasets. Furthermore, we observed that, when sampling from the aforementioned parametric distributions and applying our dropout model, the zero-inflation character of certain genes as well as the excess kurtosis and skewness of unimodal and bimodal genes were globally recovered (S3 Fig).

Probabilistic Simulation of Dropout Events

Dropouts arise from both biological (lack of transcription at measurement time) [45] and technical causes (sampling and amplification bias) [46]. For this reason, we built a probabilistic model aiming to: (i) reproduce the distribution of dropout rates across genes in the studied reference datasets; (ii) have a minimal set of gene-dependent parameters; and (iii) have a physical interpretation that accounts for the biological and technical causes of dropouts. Dropout parameters are estimated on a gene-dependent basis because empirical sampling rates exhibit gene-specific bias rather than being uniform random samples of mRNA molecules present in the cell [47]. By modelling this gene-dependent biases and simulating dropout events after sampling from parametric distributions, our dropout method mimics the physical phenomena that give rise to dropout events and generates data that reproduces the statistics of scRNA-Seq data, as illustrated by Fig. 2.

Dropout model Under the hypothesis that the probability of not observing counts for a certain gene within any given cell is inversely proportional to its relative abundance, the relationship is defined as an exponential decay which has been shown to describe the mean-dropout relationship in several scRNA-Seq datasets [48]. We denote by $x_{c,g}$ the prior pseudocount of gene g in cell c and by $x_{c,g}^{\text{obs}}$ the measured pseudocount. The mathematical formulation of the proposed dropout model is of the following form:

$$P(x_{c,g}^{\text{obs}} = 0 \mid x_{c,g}) = \beta_g e^{-\lambda_g x_{c,g}} \quad (1)$$

When simulating dropout events based on these probabilities, the number of dropout events for a given gene across all cells follows a Poisson-binomial distribution [49], that is the discrete probability distribution of a series of independent Bernoulli trials whose success (dropout) probabilities are not necessarily identical. This reflects our hypotheses on dropouts: for any given gene, having a dropout event for cell i is independent of the dropout in cell j , and two cells having comparable relative transcript abundances of any given gene will have similar probabilities of this gene being observed or dropped-out.

Rate parameter The rate parameter λ_g determines the shape of the exponential and thus how rapidly the dropout probabilities decay with the expression value. This parameter is learnt from the reference dataset, independently for each gene, in order to reflect the aforementioned gene-dependent sampling bias. It is calculated by setting the half-life of equation 1 to the gene's empirical non-zero mean as follows, for each gene g of the reference dataset:

$$\lambda_g = \frac{\ln(2)}{\hat{\mu}_{\text{NZ}}(g)} \quad (2)$$

where $\hat{\mu}_{\text{NZ}}(g)$ is the mean of non-zero pseudocounts of gene g in the reference dataset.

Normalisation constant The normalisation constant β_g is computed from sampled prior pseudocounts as the optimum value minimising the quadratic deviation between the expected dropout rate of the synthetic sample $E[\tau_g]$ and the reference dropout rate for that gene τ_g^{ref} (proportion of zero entries in the reference dataset):

$$\beta_g = \frac{n\tau_g^{\text{ref}}}{\sum_{c=1}^n e^{-\lambda_g x_{c,g}}} \quad (3)$$

where n is the number of sampled cells.

This optimum is derived analytically from the expected value of a Poisson-binomial distribution. This ensures that for the same underlying non-zero distribution the dropout rate will, on average, be close to that of the reference.

S1 Fig shows an example of the distribution of rate parameters and the obtained dropout probabilities over the range of expression of a typical log-normalised scRNA-Seq dataset. Overall, we observe a trend depending on the gene pseudocount distribution category: for the same sampled value, zero-inflated genes have the highest probability of dropout, followed by bimodal genes. Genes presenting a unimodal distribution have the lowest dropout rates (and highest non-zero means) and therefore will be seldom dropped-out.

Validation

We validated our model by sampling from the learnt parametric distributions and simulating dropouts with our exponential model of Eq. (1). We found that our method reproduces extensive statistics of these datasets, specially the gene mean-variance and mean-dropout relationships which characterise scRNA-Seq data (Fig. 2). Furthermore,

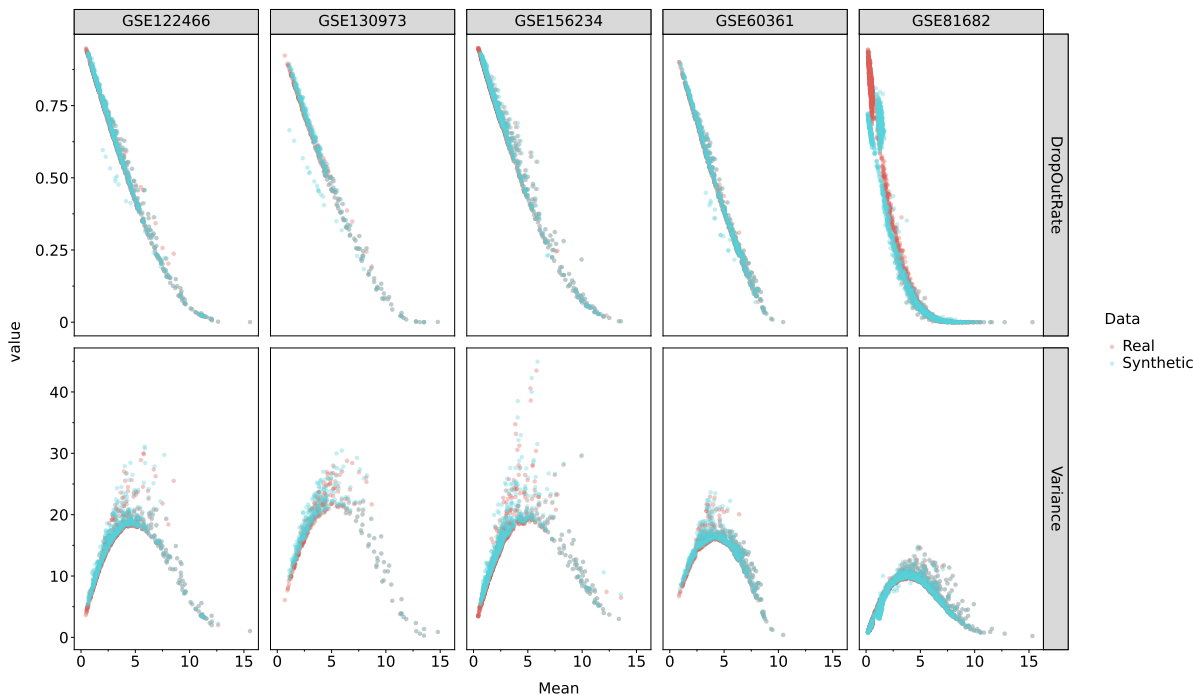


Fig 2. Mean - Variance, and Mean - Dropout Rate relationships of HVGs in different datasets. Each blue dot represents the average of 100 samples for a given gene.

the correlation profile between all combinations of mean, variance, skewness, and excess kurtosis is globally recovered (S3 Fig). We find that these correlations are only recovered when applying our dropout simulation method.

Binarisation of scRNA-seq data

The coarse-graining scheme of SCBOOLSEQ is based on the classification of pseudocount distribution from a reference dataset, as illustrated by Fig. 3. For each gene, cells whose expression level is high (respectively low) enough to classify it as True/active (resp. False/inactive) will be binarised whilst cells whose expression level is ambiguous will be left as undefined. As shown in Fig. 3, the category-dependent binarisation strategy causes each distribution type to have different proportions of False, True, and undetermined values.

Bimodal genes are binarised using their corresponding univariate two-component Gaussian Mixture Model (GMM), whose parameters are estimated on the reference dataset. The GMM's density is given by Eq. 4. The model has two components denoted C_i which are characterised by their parameters $(\phi_i, \mu_i, \sigma_i^2)$. In the following, it always holds that $\mu_2 > \mu_1$, for every bimodal gene. Therefore, we have two components which represent cells whose transcript level can be classified as active C_2 or inactive C_1 .

$$p(x) = \phi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + \phi_2 \mathcal{N}(x|\mu_2, \sigma_2^2) \quad s.t. \quad \phi_1 + \phi_2 = 1 \quad (4)$$

The probabilities of observation x belonging to each one of the two components are first calculated as detailed in Eq. (5):

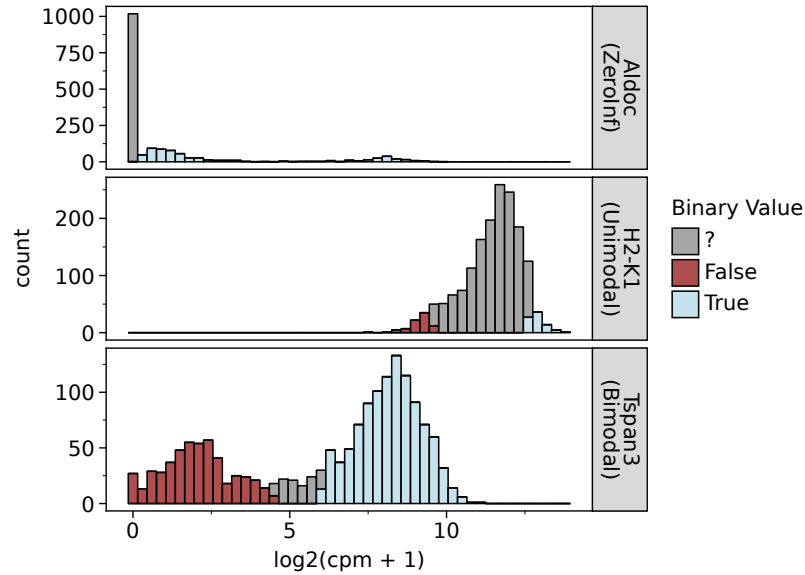


Fig 3. Illustration of the category-dependent binarisation allows accounting for different shapes in empirical pseudocount distributions. For each category, plots show the empirical distribution for a selected gene in the GSE81682 dataset, and the part of the values which are binarised with parameters $z = “?”$ for zero-inflated case, $q = 0.05$ and $\alpha = 0$ for unimodal and $\theta = 0.95$ for bimodal.

$$p(C_i|x) = \frac{p(C_i)p(x|C_i)}{\sum_{j=1}^2 p(C_j)p(x|C_j)} = \frac{\phi_i \mathcal{N}(x|\mu_i, \sigma_i^2)}{\sum_{j=1}^2 \phi_j \mathcal{N}(x|\mu_j, \sigma_j^2)} \quad (5)$$

Then, the binary classification is performed according to a given confidence threshold θ , with $0.5 < \theta \leq 1$:

$$b_{\text{bimodal}}(x) = \begin{cases} 0 & \text{if } p(C_1|x) \geq \theta \\ 1 & \text{if } p(C_2|x) \geq \theta \\ ? & \text{otherwise} \end{cases} \quad (6)$$

For genes classified as unimodal, we use symmetric thresholds based on two parameters: a *margin* quantile q (0.05 by default) and a multiplier α for the interquartile range IQR. These thresholds are based on Tukey’s fences for outlier detection [50], with modified defaults to binarise a small fraction of observations. Note that in Eq. (7), $Q(q)$ represents the q -th quantile of the gene’s empirical distribution.

$$b_{\text{unimodal}}(x) = \begin{cases} 0 & \text{if } x < Q(q) - \alpha \text{IQR} \\ 1 & \text{if } x > Q(1 - q) + \alpha \text{IQR} \\ ? & \text{otherwise} \end{cases} \quad (7)$$

Finally, genes whose empirical pseudocount distribution is classified as zero-inflated use a zero-or-not binarisation scheme [45]. Genes having non-zero counts are classified as True whilst zero entries are classified as undetermined (parameter $z = “?”$) to reflect the uncertainty regarding the technical/biological causes of this zero, or as False (parameter $z = 0$) if considered as a signal, as suggested by [45].

$$b_{\text{zero-inflated}}(x) = \begin{cases} 1 & \text{if } x > 0 \\ z & \text{otherwise} \end{cases} \quad (8)$$

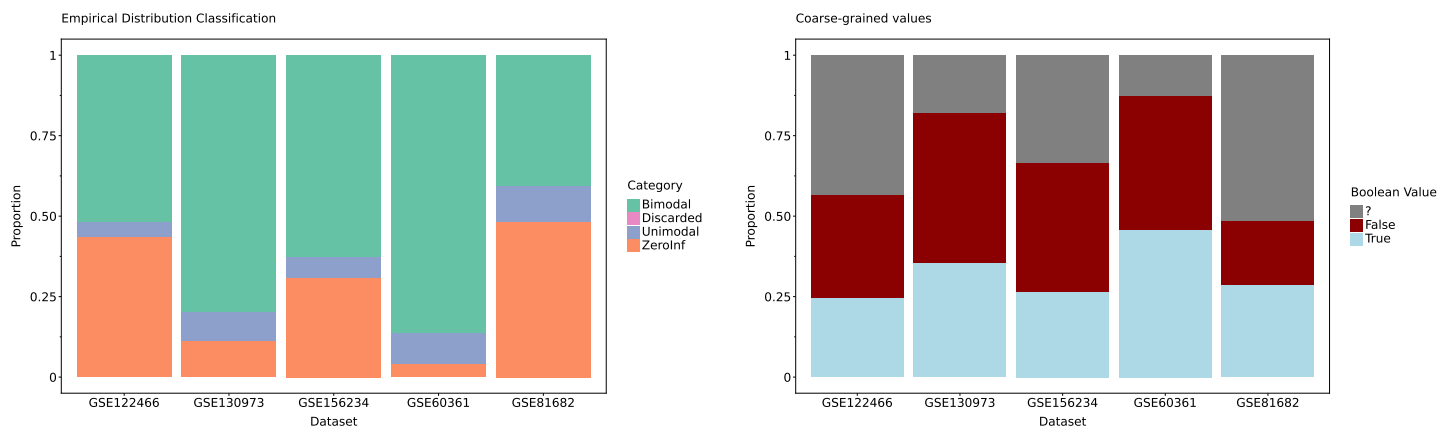


Fig 4. Left: Distribution of categories among the studied datasets. Right: Proportion of binarised values across datasets using the default parameters for each distribution type. These proportions are both determined by the categories and the specified thresholds. These were obtained using parameters $z = ?$ for zero-inflated case, $q = 0.05$ and $\alpha = 0$ for unimodal, and $\theta = 0.95$ for bimodal. The dropout rate threshold for marking a gene as discarded was set to 0.99.

The proportion of observations classified as 0 or 1 can be approximated by Eq. (9) 240

$$\xi(1 - \tau) + \beta(p^*) + \eta(2q) \quad (9)$$

with the average proportions of binarised observations for each category normalized by the proportion of genes classified as zero-inflated, bimodal, and unimodal, denoted by ξ , β , and η , respectively, and where τ represents the average empirical dropout rate. 241
242
243

Fig. 4 gives statistics on the fraction of observations that are binarised across the selected evaluation datasets. In general, zero-inflated genes with a high dropout rate will only have a few observations binarised to 1 and most cells will be classified as undefined. Bimodal genes are binarised across most cells because the underlying Gaussian Mixture correctly describes the bimodal genes' empirical distributions. Finally, unimodal genes will have twice the margin quantile q fraction of observations binarised in the case of $\alpha = 0$ in Eq. (7). 244
245
246
247
248
249
250

Case study of binarisation: Early-born Retinal Neurons 251

We applied scBOOLSEQ to a publicly available scRNA-Seq dataset in order to binarise expression data and obtain a qualitative description of phenotypes. We show that the obtained qualitative profiles can serve as a basis to perform inference of Boolean networks, which can mimic the differentiation process and identify key genes and interaction involved in the dynamics. 252
253
254
255
256

The dataset originates from [9] (*GEO accession* GSE122466) which analysed how the diversity of cell types found in the early retina (from embryonic days 10 to 17) arises from a pool of progenitor cells. These neurons are retinal ganglion cells (RGCs), cone photoreceptors (cones), horizontal cells (HC) and amacrine cells (AC). The analyses extended previously known marker genes and showed how these appear to be organised in transcriptional waves of co-expression. Extending the original results with a mechanistic model could help formulate hypotheses regarding the underlying regulatory mechanisms of early retinogenesis. Here, we illustrate how to combine the statistical analysis of scBOOLSEQ to coarse-grain the expression data with prior knowledge data on transcription factor regulations publicly available in the mouse 257
258
259
260
261
262
263
264
265
266

Table 1. A list of all the cellular types of interest, as well as the Boolean markers (cells with those genes binarised to 1/True/active) used to detect cells matching belonging to them. *N. Unambiguous Cells* represents cells that exclusively expressed the given set of markers.

Cell Type	N. Unambiguous Cells	Perc. Total	Markers
RPC (Retinal Progenitor Cells)	249	98.03%	Sox2, Fos, Hes1
NB1 (Neuroblasts, first group)	23	85.19%	Top2a, Prc1, Sstr2, Penk, Btg2
NB2 (Neuroblasts, second group)	27	81.82%	Neurod4, Pax6, Pcdh17
RGC (Retinal Ganglion Cells)	191	94.55%	Isl1, Pou4f2, Pou6f2, Elavl4
AC (Amacrine Cells)	81	67.50%	Onecut2, Prox1
HC (Horizontal Cells)	3	10.71%	Onecut1, Prox1
Cones (Photoreceptors)	8	100%	Otx2, Crx, Thrb, Rbp4

regulon database DOROTHEA [40] in order to build logical models which reproduce the differentiation process. Our objective is to first evaluate how the binarisation preserves the cell type classification, and how the resulting qualitative description of phenotypes enables to identify core regulations that explain the Boolean differentiation process.

Discriminating cellular types using prior-knowledge markers The reference study [9] considered prior knowledge markers for the cellular types at different stages of differentiation. We classified each cell according to its binarised expression profile and the markers it contains. Then, for each cellular type, we computed how many cells have the matching marker, and among them, how many match only with that cellular type. As shown in Table 1, the majority of cells per group were unambiguously identified, except for Horizontal Cells. Notice that Horizontal Cells share one marker *Prox1* with Amacrine Cells. It should be noted, that in this case, a quarter of cells have been classified using their binarisation (S4 Fig). Moreover, our classification of cells based on their binarised pseudocounts and prior-knowledge markers enables to label Louvain clusters of scRNA-Seq data, which turned out to be consistent with labels obtained using differential expression analysis by [9] (S5 Fig).

Data-driven inference of Boolean models The binarisation of SCBOOLSEQ enables to specify Boolean dynamical properties that reflect the observed differentiation process: existence of trajectories linking (partially) binarised cellular states, including branches from pluripotent states to distinct differentiated states, as well as stability properties. Then, inference methods such as BONESIS [51, 52] can derive Boolean networks that reproduce the specified dynamics. The logical rules are derived from prior knowledge Gene Regulatory Networks (GRNs), typically extracted from TF-TF (transcription factor - transcription factor) interaction databases, possibly completed with statistical network inference from scRNA-Seq data. By employing combinatorial optimization method, BONESIS enables accessing to the sparsest models, i.e., requiring as few as possible genes to reproduce the desired trajectories and stable states.

Using clustering and trajectory reconstruction methods, we applied SCBOOLSEQ to determine a partial binary profile of 6 cellular types, namely RPC (progenitor), intermediate neuroblast types NB1 and NB2, and final Cones, RGC and AC types. Note that due to the low number of cells classified as HC and their apparent distance between each others, we omitted this cellular state. The dynamical specification consisted in the existence of a trajectory from the RPC state to NB1 and then to NB2. From the NB2 state, three different trajectories must exist towards each of the final stable states. Moreover, we extracted from the DOROTHEA database a core TF-TF regulatory network together with target genes which have been binarised. Focusing on the largest

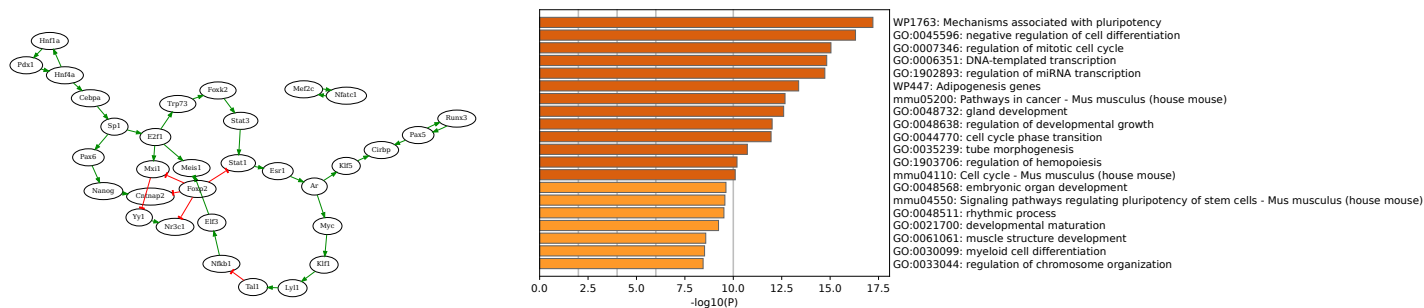


Fig 5. **Left:** Simplified view of the set of minimal TF-TF interactions employed in the Boolean models reproducing the differentiation process. For display, all leaf nodes with an in-degree of 1 were recursively removed from the GRN. The full filtered GRN obtained with BONESIS is provided in S6 Fig. **Right:** Top Gene Ontology Terms related to the 184 genes of the filtered GRN.

weakly connected component, it gave a GRN with 644 genes. Then, using BONESIS, we reconstructed Boolean networks that, using the input GRN interactions, are able to reproduce the desired trajectories and stable states. See Methods section, S7 Fig, and S1 Code for details. Because the binary profiles are partials, numerous genes have no imposed binary state in several cellular states. Using BONESIS, we identified models which rely as little as possible on the dynamics of those genes with undetermined states. It resulted in pruning the input GRN to 184 genes which suffice to explain the observed differentiation process. As shown in Fig. 5(Right), gene ontology enrichment analysis, performed using Metascape [41], shows many relevant ontology terms were found among the top hits, such as mechanisms associated with pluripotency, negative regulation of cell differentiation, regulation of mitotic cell cycle, gland development, regulation of developmental growth, and embryonic organ development. Obtained models can then serve as inputs for a more thorough systems biology analysis of the biological problem.

Synthetic scRNA-seq generation biased by Boolean states

As the inverse operation of binarisation, the parametric distributions and dropout model learned per genes from a reference dataset also enable generating synthetic pseudocounts corresponding to Boolean activation states. The main principle is to perform first biased sampling from distributions whose parameters are learnt on non-zero entries of the reference dataset. In a second step, dropout events are simulated according to the gene-dependant model of Eq. (1).

Biased sampling ensures that cells in which a gene is active will exhibit higher expression (pseudocounts) than those in which it is inactive. In the case the gene follows a unimodal distribution of median μ and variance σ , the pseudocounts are sampled from the half-normal distribution corresponding to the activation state ($\mathcal{HN}(\mu, \sigma^2)$ for active, and $\mu - \mathcal{HN}(0, \sigma^2)$ for inactive). In the case of bimodal distribution, composed of two normal distributions of median $\mu_1 < \mu_2$ and variance σ_1 and σ_2 , respectively, the sampling is performed from the mode corresponding to the activation state ($\mathcal{N}(\mu_1, \sigma_1^2)$ for inactive and $\mathcal{N}(\mu_2, \sigma_2^2)$ for active). Finally, in the case of zero-inflated genes, the learning from non-zero entries ensures falling back to one of the two aforementioned cases, and the dropout model learnt should reflect the inflation of zeros. The last step simulates dropouts in such a way that synthetic log-pseudocounts produced from the Boolean states will have gene-wise statistical properties closely resembling those of real scRNA-Seq data. The dropout event simulation can follow the

dropout model of Eq. (1) learnt per gene, or follow an arbitrary given distribution. 337

Application to artificial Boolean models The above steps enable generating 338
synthetic scRNA-Seq datasets from collections of binary states of genes, as it would be 339
typically generated from the simulation of Boolean networks [53, 54]. This generation 340
can then serve as a basis for benchmarking inference methods, by creating synthetic 341
datasets from fixed dynamical models and evaluate the ability of inference methods to 342
recover main features of the ground-truth model. This could notably be applied from 343
artificial Boolean models of different scale and topology. In that cases, however, node 344
are not directly referring to the genes of an experimental scRNA-Seq reference dataset, 345
and one need criteria to associate a reference gene to them. 346

A possible approach, proposed in SCBOOLSEQ, is to analyse the shape of the 347
node-wise distribution of Boolean values and assign genes having similar shape. 348
Intuitively, a gene is for instance active in most cells, it can be classified as Unimodal. 349
Subsequently, genes which vary considerably can be considered to be Bimodal. Genes 350
which are ubiquitously inactive with a couple exceptions (e.g., it is active in only one 351
state of the Boolean trace) would then be zero-inflated. SCBOOLSEQ uses scaled 352
versions of the first four moments to classify Boolean gene distributions as unimodal, 353
bimodal, or zero-inflated. The scaled moments of Boolean distributions are fed to a 354
k-nearest-neighbours classifier that was trained on the scaled moments of reference 355
dataset, using their corresponding distribution types. Afterwards a by-category bijective 356
matching is performed in order to ensure that the synthetic scRNA-Seq distributions 357
correctly represent the underlying Boolean dynamics. 358

We applied this principle on three artificial Boolean models, exhibiting different type 359
of emerging dynamics. For each one of the models, Boolean trajectories representing the 360
dynamics of the network were obtained as described in the next paragraphs. Afterwards 361
multiple observations (corresponding to single cells) were sampled using SCBOOLSEQ 362
with a selected reference dataset (GSE81682). Then, we applied classical scRNA-Seq 363
dimensionality reduction methods to visualise the corresponding pseudocount 364
trajectories. Further details regarding the sampling procedure and projections can be 365
found in the supplementary materials. 366

The first artificial model is a star-like network (Fig. 6a) in which a single 367
Transcription Factor (TF) up-regulates the expression of a set of genes. This model was 368
simulated by performing one random walk with the fully asynchronous update mode 369
starting from the state where the node *tf* is active and all genes are inactive. The 370
resulting trajectory is a sequence of Boolean vectors where genes progressively activate, 371
in a random order. This gradual activation can be clearly distinguished in Figure 6b, 372
where cells with few active genes are coloured in dark blue and cells with all genes 373
active are coloured in light green. 374

The second manually-designed model is a bistable switch which represent a 375
simplified *cellular reprogramming* scenario (Fig. 6c) in which the cell finds itself in a 376
steady state (light blue, labelled *common*) characterised by the activation of *TF6* which 377
activates a small set of genes and inhibits a mutually exclusive switch. The activation of 378
TF7 node represents a perturbation which inhibits *TF6*, pushing the cell out of its 379
initial state and triggering a differentiation process. At the end, one of two different 380
stable states is reached. The third model (Fig. 6e) is a three-stable switch which has 381
been designed automatically from random scale-free topology and such that it exhibits a 382
two-level differentiation process: from an initial state three stable states are reachable, 383
with an intermediate branching state giving access to two of them. In both cases, we 384
generated Boolean trajectories covering the differentiation branches from the initial 385
states. These trajectories remain apparent in the projections of generated scRNA-Seq 386
data (Fig. 6d and Fig. 6f). 387

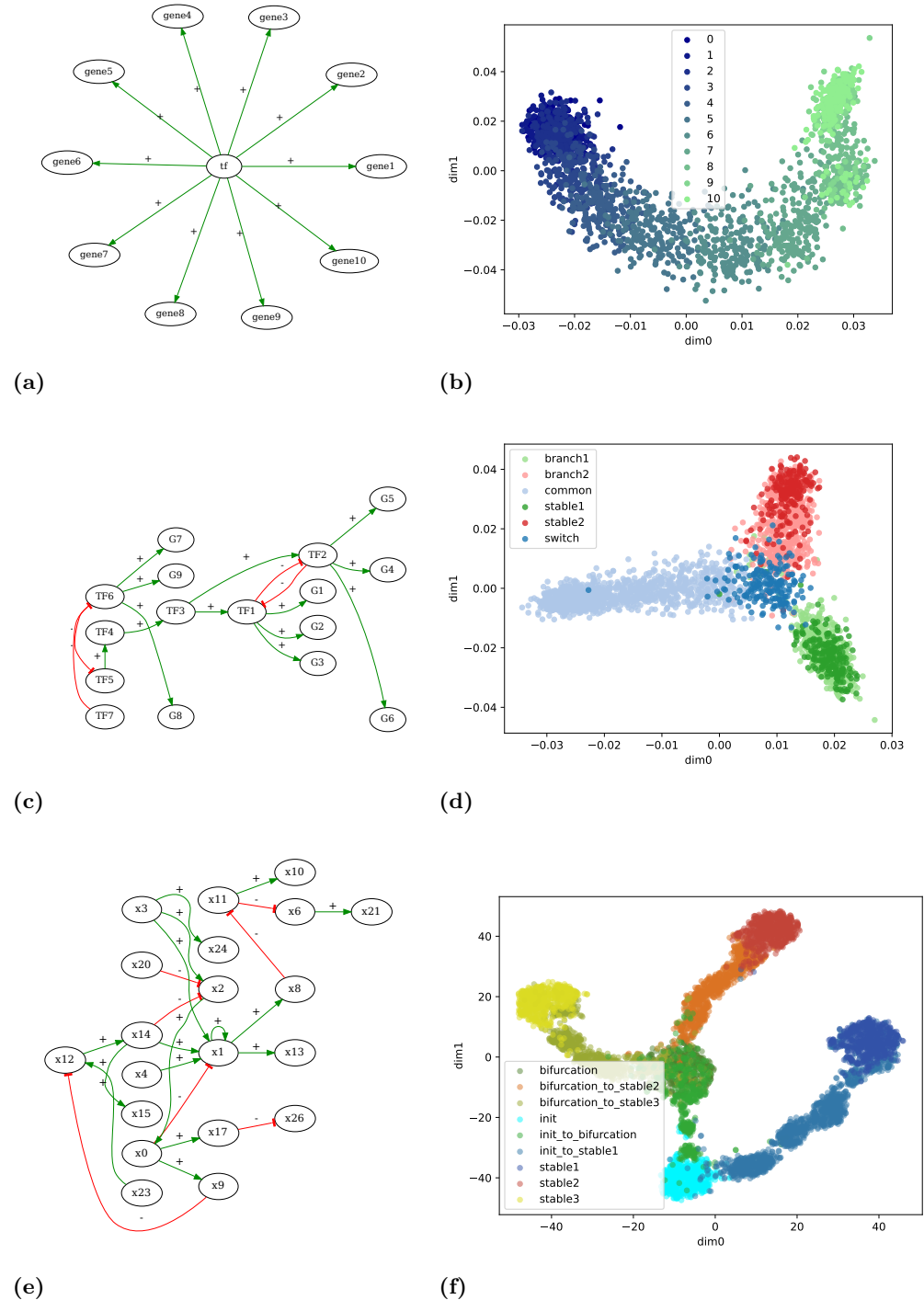


Fig 6. Artificial Boolean models and generated synthetic scRNA-Seq data. **Left:** Influence graphs of the Boolean models. See supplementary material for Boolean functions. **Right:** Two-dimensional projection of the synthetic scRNA-Seq data generated by applying SCBOOLSEQ to Boolean trajectories simulated from the models on the left. Dots are labelled with a description of the Boolean state they have been generated from: for (b) it is the number of active genes; for (d) and (f) they refer to the dynamical nature of the states in the 3-branches differentiation process.

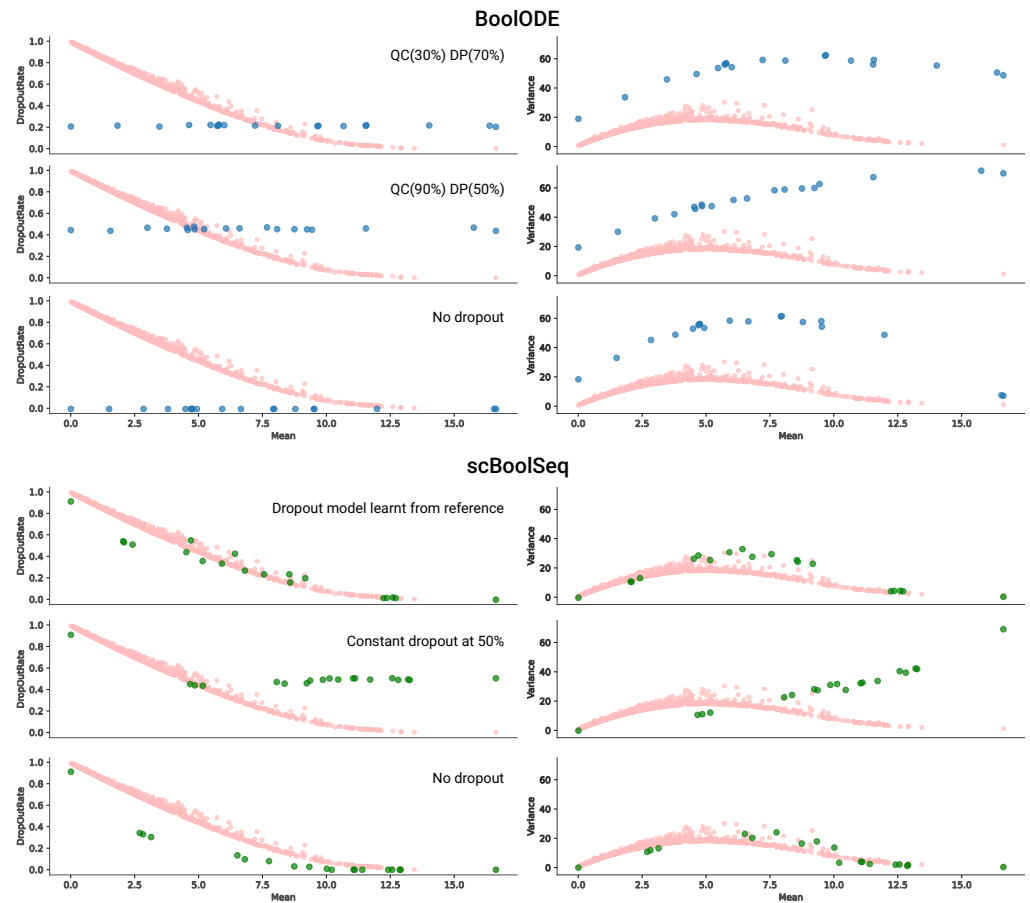


Fig 7. Comparison of the per-gene Mean-Variance and Mean-DropOutRate profiles of reference dataset GSE122466 (red), BOOLODE (blue), and scBOOLSEQ (green). QC represents the quantile below which BOOLODE simulates dropouts with a constant probability DP .

Comparison with BoolODE Given an artificial Boolean network, the tool BOOLODE [34] is capable of producing synthetic pseudocount datasets which exhibit clearly defined trajectories when applying dimensionality reduction techniques such as t -SNE. However, the generated dataset do not exhibit observed statistics of experimental scRNA-Seq dataset.

Fig. 7 provides comparisons between datasets generated by BOOLODE and scBOOLSEQ from one of the largest curated model of the benchmark of [34], a Boolean network of human gonadal sex determination (GSD) [55]. It has two main fixed point attractors of biological interest, namely Sertoli cells and granulosa cells which correspond to male and female phenotypes. We notably compared the mean-variance and mean-dropout profiles of generated data with different dropout models, as proposed by both tools. Besides the dropout rate being constant, the mean-variance relationship of BOOLODE appears to be at very different scale than typical scRNA-Seq data (Fig. 2). It should be noted that when enforcing a constant dropout rate with scBOOLSEQ, the resulting dropout-mean profile is not constant as 0 values can still be sampled from learnt pseudocount distributions: gaussian distribution can give non-zero probabilities to negative values, which are corrected as 0. This is not the case with BOOLODE because of the noise added to ODE-simulated values, which prevents generating values being exactly 0.

```
import pandas as pd
from scboolseq import scBoolSeq

# cells are rows and genes are columns
reference = pd.read_csv("reference_scRNA_highly_variable_genes_pseudocounts.csv")

scbool = scBoolSeq()
# compute criteria (statistics and per-gene category)
scbool.fit(reference)

# binarise the reference dataset (or other)
coarse_grained = scbool.binarize(reference)

# Simulate scRNA_Seq experiments from Boolean data
boolean_states = pd.read_csv("simulated_boolean_dynamics.csv")
synthetic_rna = scbool.sample_counts(boolean_trace)
```

Fig 8. Python code snippet showing basic usage of SCBOOLSEQ for binarisation and synthetic data generation from reference scRNA-Seq data and Boolean states

Implementation and usage

SCBOOLSEQ has been implemented in Python on top of `pandas` [56], `statsmodels` [57], and `scikit-learn` [58] libraries. Fig. 8 shows basic usage of `scBoolSeq` to perform binarisation and synthetic data generation. Future engineering work will focus on leveraging the `AnnData` [59] Python package for handling large datasets that cannot be fit in RAM. Furthermore, using `AnnData` within SCBOOLSEQ will allow its integration in the `scverse` [60] computational ecosystem for single-cell omics data analysis.

SCBOOLSEQ is distributed as a standard Python package, and is integrated in the CoLoMoTo Docker distribution [61], which facilitates the accessibility of tools related to Boolean and logical models, and the reproducibility of related computational analyses.

Discussion

We introduced SCBOOLSEQ, a novel method which provides a bidirectional link between scRNA-Seq data and Boolean Models. Our method builds on the classification of gene empirical pseudocount distributions into unimodal and bimodal distributions proposed by [42], that we extended with a probabilistic gene-dependent dropout model. We showed that the resulting characterization suffices to capture the main statistical features of real scRNA-Seq data. Then, SCBOOLSEQ offers both the ability to binarise scRNA-Seq datasets and the ability to generate synthetic pseudocounts from binary states of genes.

From pseudocounts to binary states We illustrated on a concrete application how the binarisation offered by SCBOOLSEQ can be employed to process scRNA-Seq data in view of performing inference of Boolean networks, which are logical models of gene activity dynamics. First, SCBOOLSEQ coarse-graining method allows identifying cellular types of interest by detecting the presence (i.e. activation) of known marker genes. In addition to this, combining the binarised gene activity with community detection techniques could help to find previously unknown marker genes (genes which are binarised as active only in certain clusters and are not found in the literature). Then, coupled with a prior GRN, the deduced set of Boolean functions constitute a set of hypotheses that can guide future wetlab experiments in order to unveil the core regulatory mechanism of early retinogenesis.

It should be stressed that the binarisation of scBOOLSEQ can result in undetermined state when there is not enough statistical evidence for a binary classification. We believe that the fact that not all genes (and cells) cannot be classified with binarisation is good sign that the method enables discriminating cells in extreme state from cell in transient state, for which a fully binary view may not be adequate.

One should note however that determining the activity of a gene based on its transcript level is a strong hypothesis. Methods such as VIPER [62] aim at adding information about each protein's regulon to better infer protein activity. Moreover, chromatin accessibility and other epigenetics information can also help to refine the binary classification.

From binary states to pseudocounts Another major contribution of scBOOLSEQ is its method for generating synthetic scRNA-Seq data from Boolean gene activation states by biased sampling from learned pseudocount distributions on a reference dataset. We showed that scBOOLSEQ provides a significant improvement over BOOLODE as it produces synthetic scRNA-Seq data whose statistical characteristics (mean-variance and mean-dropout profiles) closely resemble those of real data. In addition to this, scBOOLSEQ allows simulating any arbitrary distribution of gene-wise dropout rates. This represents an unprecedented contribution as it allows measuring the sensitivity of inference methods to the dropout rate distributions of scRNA-Seq datasets.

By offering the capability to generate synthetic scRNA-Seq datasets from ground-truth Boolean models with realistic statistical features, we believe that scBOOLSEQ is a clear asset for that generating benchmarks for the evaluation of various inference methods, such as GRN inference, trajectory reconstruction, and data-driven Boolean network inference.

Methods

Boolean networks and dynamics

A *Boolean network* on nodes $\{1, \dots, n\}$ is a function $f : \mathbb{B}^n \rightarrow \mathbb{B}^n$ mapping binary vectors of dimension n to themselves, where $\mathbb{B} = \{0, 1\}$ is the Boolean domain. For each node $i \in \{1, \dots, n\}$, we write $f_i : \mathbb{B}^n \rightarrow \mathbb{B}$ the i -th component of f , which is the Boolean function of node i . A Boolean vector $\mathbf{x} \in \mathbb{B}^n$ specifies a Boolean state for each component of the network, and is called a *configuration*.

The *influence graph* of a Boolean network f is a directed signed graph, noted $G(f)$, whose vertices are the nodes of the Boolean network. The influence graph captures the dependencies of Boolean functions, and corresponds to union of Jacobian matrices of f on configuration. Formally, there is a positive edge for node j to i ($j \xrightarrow{+} i \in G(f)$) in the influence graph if and only if there exists a configuration $\mathbf{x} \in \mathbb{B}^n$ such that

$$f_i(\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, 0, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n) < f_i(\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, 1, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n)$$

There is a negative edge for node j to i ($j \xrightarrow{-} i \in G(f)$) in the influence graph if and only if there exists a configuration $\mathbf{x} \in \mathbb{B}^n$ such that

$$f_i(\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, 0, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n) > f_i(\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, 1, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n)$$

Note that it is possible to have both edges $j \xrightarrow{+} i$ and $j \xrightarrow{-} i$ in a same influence graph. If it is the case for $G(f)$, then the Boolean network f is said to be non-monotone. Otherwise, f is *locally monotone*.

A *trajectory* of a Boolean network f is a sequence of configurations $\mathbf{x}^1, \dots, \mathbf{x}^k$ that can be computed according to a given *update mode*. For instance, the synchronous mode

computes trajectories such that any two successive configurations $\mathbf{x}^m, \mathbf{x}^{m+1}$ are such that $\mathbf{x}^{m+1} = f(\mathbf{x}^m)$; the fully asynchronous update mode computes trajectories such that any two successive configurations $\mathbf{x}^m, \mathbf{x}^{m+1}$ differ on only one node i , and verify that $\mathbf{x}_i^{m+1} = f_i(\mathbf{x}^m)$. The *most permissive* update mode [24] computes all the trajectories that are binarised from any asynchronous trajectory of multivalued and quantitative model compatible with the Boolean network. In general, it allows much more trajectories than synchronous and (general) asynchronous modes, which fail to capture trajectories of different class of quantitative systems, including incoherent feed forward loops [24].

A configuration $\mathbf{x} \in \mathbb{B}^n$ is a *stable state* if $f(\mathbf{x}) = \mathbf{x}$, i.e., it is a fixed point of f . A configuration $\mathbf{x} \in \mathbb{B}^n$ belong to an *attractor* of f under a given update mode whenever for any possible trajectory from \mathbf{x} to another configuration \mathbf{y} , there exists a trajectory going back to \mathbf{x} . Stable states are particular cases of attractors.

Inference of Boolean networks from influence graph and dynamical properties

From an influence graph \mathcal{G} and a set of dynamical properties, the tool BONESIS [51, 52], available at github.com/bnediction/bonesis, allows inferring all the locally-monotone Boolean networks f having their influence graph enclosed by \mathcal{G} , i.e., $G(f) \subseteq \mathcal{G}$, and that posses the input dynamical properties. The dynamical properties supported by BONESIS include the existence of most permissive trajectories between partially specified configurations, and stable state properties of (partially specified) configurations. A partially specified configuration specify a Boolean state for a subset of nodes. In that case, BONESIS is free to complete the unspecified nodes with any Boolean state. BONESIS also allows specifying optimization objectives to filter solutions, notably to enumerate only sparser models, i.e., with the smallest influence graphs.

We employed BONESIS to infer Boolean networks from scRNA-Seq scBOOLSEQ binarisation (see next section), and to generate artificial Boolean networks which possess multi-stability and branching behaviors from randomly generated scale-free influence graph (S1 Code).

Case Study: Early Born Retinal Neurons

We performed the analyses on the scRNA-Seq dataset of lane 1 of GSE122466. The main steps hereafter denoted in paragraphs refer to the analyses performed in their homonymous Jupyter Notebooks provided in S1 Code.

Highly Variable Gene Selection For this part we used the software STREAM [15]. We took the count matrix of the first replicate (Identified with the prefix `Lane_1` in their index). We performed standard quality control, with the same parameters as the analyses of the original article. Cells expressing less than 200 genes where discarded, as well as genes expressed in less than 3 cells. We selected the 1648 most highly variable genes and appended to them the two marker genes which were reported in the article but were not selected as being highly variable (*Rbp4*, *Pou4f2*).

Retinal Differentiation Clustering and Metadata In this part we took the aforementioned Highly Variable Genes (HVGs) and performed the scBOOLSEQ distribution learning with $\theta = 0.75$ to have a higher the amount of binarised observations on bimodal genes. We then used the instance to binarise the HVGs across all cells. We then identified cells matching the markers described in the original article. About 25% of all cells where labelled in this process. Subsequently, cells matching more

than one set of markers were discarded. The only pair of phenotypes which presented more than a couple ambiguous cells where Amacrine Cells (AC) and Horizontal Cells (HC) which had 23 cells matching both marker signatures. This was expected given that cellular types were defined with only two markers and one of them *Prox1* is shared. Having a larger (and preferably disjoint) set of markers could resolve this ambiguity. We used SCANPY [17] to perform louvain clustering on the log pseudocount HVGs, with the number of neighbours set to 15. With this analysis, 11 distinct clusters were found. A small cluster of cells (cluster 10 in the notebooks) was discarded as it was determined to be an unknown cluster of unknown Retinal Ganglion Cell-like U/RGC. Our Boolean analysis also found this isolated cluster to express signature genes of RGCs. Finally, clusters were labelled using the majority label of cells whose Boolean identity matched the markers. Most clusters had absolute majorities (85%, 98%) except for one (Cluster 3 had 53.84% of cells voting NB2, and 34.61% voting AC: It was labelled NB2). These labels were used as metadata in order to perform trajectory inference.

Trajectory Inference Using STREAM we performed trajectory inference, using the aforementioned cluster labels as metadata. We obtained a well-defined trifurcating trajectory which is distinguishable on two dimensions. We set the root (starting point) to be Retinal Progenitor Cells (RPCs) and the three final points to be the Cones, Retinal Ganglion Cells (RGC), and Amacrine Cells respectively. Cells associated with these terminal nodes of the inferred graph were taken to be representative of their corresponding phenotypes. For the two groups of neuroblasts (NB1 and NB2), cells within the two quartiles $Q(.25)$, $Q(.75)$ of the root node's pseudotime were chosen as representative of these transient phenotypes. This yields a total of 133 RGC, 79 NB1, 17 NB2, 109 AC, 78 RPC, and 69 Cones that were used to infer the Boolean model.

Binarisation of scRNA-Seq data We binarised all HVGs across all cells and employed the metadata obtained from the previous trajectory inference step to retrieve cell groups. We defined meta-observations by aggregating each group, using the mode as summary statistic. We further selected genes having non-null variance, which reduced the original 1650 genes to only 1426. We only retained binarised genes present in the mouse regulon database DOROTHEA [40], that is 1263.

Boolean Model Inference : Having our binarised observations and selected genes, we defined our GRN using DOROTHEA [40]. DOROTHEA gives a confidence score to each one of the interactions, based on the number of supporting evidence in different sources. In decreasing order, these levels are: *A,B,C,D,E*. We decided to exclude interactions with low supporting evidence, so we filtered out levels *D,E* and considered only levels *A,B,C*. With these filtered interactions, we extracted the core TF-TF network which we define to be the biggest strongly connected component of the departing graph. This core TF-TF network has 157 nodes. We then obtained the subgraph induced by these 157 core transcription factors and the binary genes comprising our observations. This yielded a GRN with 728 nodes. We tested and found that this GRN was not weakly connected. We extracted the biggest weakly connected component which contained 633 nodes. This weakly connected component was given to BONESIS as the domain of Boolean Networks to consider, and specified the desired trajectories and stable states using the specification given in S7 Fig.

Supporting information

S1 Fig. Example of distribution of rate parameters and dropout probabilities learnt by scBoolSeq. Left: Distribution of rate parameters λ

estimated on dataset GSE122466. Right: Dropout probabilities computed between the minimum and maximum values of a sample from the parametric distributions corresponding to the same dataset. Each line corresponds to an individual gene.

S2 Fig. Mean - Variance and Mean - DropOutRate relationships of HVGs using PROFILE parametric distributions for bimodal and unimodal genes on selected scRNA-Seq datasets. Each green point represents the average of 100 independent replicates with the same sample size as the reference dataset.

S3 Fig. Correlation between higher moments of real pseudocount data and from data generated from distributions and dropout model learnt by scBoolSeq on selected scRNA-Seq datasets

S4 Fig. Position of cells classified using scBoolSeq binarisation and prior-knowledge markers. t-SNE and UMAP projections trained on the top 25 principal components (log pseudocount matrix). Colours indicate cell identities determined by binary value of known markers (see Table 1).

S5 Fig. Result of trajectory reconstruction using STREAM on early-born retinal neurons scRNA-Seq data. UMAP projection of the first 25 principal components to 3 dimensions (only 2 are shown). The cluster labels are determined by the majority label of unambiguous cell types identified via SCBOOLSEQ binarisation.

S6 Fig. Influence graph of sparser Boolean networks learnt using BoNesis from qualitative dynamics of case study obtained with scBoolSeq binarisation. This graph comprises 184 nodes forming Boolean networks that can reproduce the Boolean dynamics of early-born retinal neurons differentiation process. This graph is a subgraph of the input DOROTHEA TF-TF interaction database. Green arrows indicate positive regulations, red arrows indicate negative regulations. Nodes without predecessors indicate nodes with constant function in the Boolean networks. Thus, the Boolean state of these nodes is identical in all stable states, and is in opposite state in the precursor state RPC.

S7 Fig. Python code snippet showing usage of BoNesis for the inference of Boolean networks for the retinal differentiation case study. See S1 Code for full pipeline.

S1 Code. Code and notebooks for reproducing binarisation case study and synthetic data generation SCBOOLSEQ source code is available at github.com/bnediction/scBoolSeq. The Python package can be installed using `conda` or `pip`; see link for instructions. Notebooks for demonstrating SCBOOLSEQ usage and reproducing the case studies presented in this paper can be visualised and downloaded at nbviewer.org/github/bnediction/scBoolSeq-supplementary.

Funding

Work of GML and LP was partly supported by the French Agence Nationale pour la Recherche (ANR) in the scope of the project “BNeDiction” (grant number ANR-20-CE45-0001). Work of GML was partly supported by the Talentos de Exportación - JuventudEsGto scholarship program of the Mexican State of Guanajuato. Work of LP was partly supported by the French government in the scope of France 2030

project “AI4scMED” operated by ANR (grand number ANR-22-PESN-0002). LC was party supported by ModICeD project from MIC ITMO 2020. AZ was supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

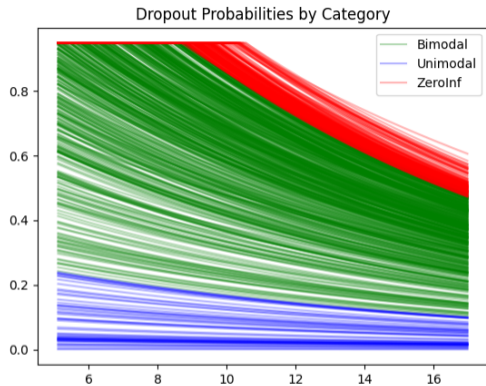
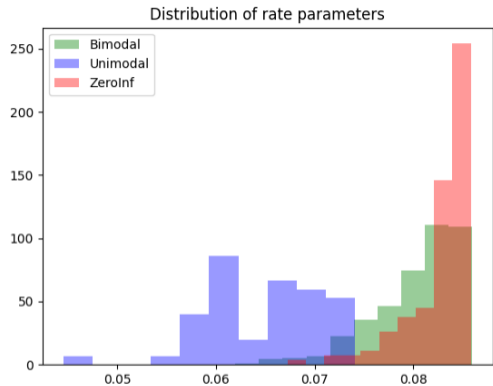
1. Kerkhofs J, Roberts S, Luyten F, Van Oosterwyck H, Geris L. A Boolean network approach to developmental engineering. In: TERMIS-EU 2011, Date: 2011/06/06-2011/06/10, Location: Granada; 2011.
2. Kerkhofs J, Roberts SJ, Luyten FP, van Oosterwyck H, Geris L. Relating the chondrocyte gene network to growth plate morphology: From genes to phenotype. *PLoS ONE*. 2012;7(4):1–11. doi:10.1371/journal.pone.0034729.
3. Lesage R, Kerkhofs J, Geris L. Computational modeling and reverse engineering to reveal dominant regulatory interactions controlling osteochondral differentiation: Potential for regenerative medicine. *Frontiers in Bioengineering and Biotechnology*. 2018;6(NOV):1–16. doi:10.3389/fbioe.2018.00165.
4. Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*. 2016;128(8):e20–e31. doi:10.1182/blood-2016-05-716480.
5. Hérault L, Poplineau M, Duprez E, Remy É. A novel Boolean network inference strategy to model early hematopoiesis aging. *Computational and Structural Biotechnology Journal*. 2023;21:21–33. doi:10.1016/j.csbj.2022.10.040.
6. Schwab JD, Ikonomi N, Werle SD, Weidner FM, Geiger H, Kestler HA. Reconstructing Boolean network ensembles from single-cell data for unraveling dynamics in the aging of human hematopoietic stem cells. *Computational and Structural Biotechnology Journal*. 2021;19:5321–5332. doi:10.1016/j.csbj.2021.09.012.
7. Krumsiek J, Marr C, Schroeder T, Theis FJ. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS ONE*. 2011;6(8). doi:10.1371/journal.pone.0022649.
8. Jerber J, Seaton DD, Cuomo ASE, Kumasaka N, Haldane J, Steer J, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics*. 2021;53(3):304–312. doi:10.1038/s41588-021-00801-6.
9. Giudice QL, Leleu M, Manno GL, Fabre PJ. Single-cell transcriptional logic of cell-fate specification and axon guidance in early-born retinal neurons. *Development (Cambridge)*. 2019;146(17). doi:10.1242/dev.178103.
10. Terfve C, Cokelaer T, Henriques D, MacNamara A, Goncalves E, Morris MK, et al. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC systems biology*. 2012;6(1):1–14.
11. Arshad OA, Datta A. Towards targeted combinatorial therapy design for the treatment of castration-resistant prostate cancer. *BMC bioinformatics*. 2017;18(4):5–15.

12. Gómez Tejeda Zañudo J, Scaltriti M, Albert R. A network modeling approach to elucidate drug resistance mechanisms and predict combinatorial drug treatments in breast cancer. *Cancer convergence*. 2017;1(1):1–25. 652–654
13. Gupta S, Silveira DA, Mombach JCM. ATM/miR-34a-5p axis regulates a p21-dependent senescence-apoptosis switch in non-small cell lung cancer: a Boolean model of G1/S checkpoint regulation. *FEBS letters*. 2020;594(2):227–239. 655–657
14. Wang Z, Ding H, Zou Q. Identifying cell types to interpret scRNA-seq data: How, why and more possibilities. *Briefings in Functional Genomics*. 2020;19(4):286–291. doi:10.1093/bfgp/elaa003. 658–660
15. Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nature Communications*. 2019;10(1). doi:10.1038/s41467-019-09670-4. 661–663
16. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*. 2017;14(10):979–982. doi:10.1038/nmeth.4402. 664–666
17. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*. 2018;19(1):15. doi:10.1186/s13059-017-1382-0. 667–668
18. Martínez-Sosa P, Mendoza L. The regulatory network that controls the differentiation of T lymphocytes. *Biosystems*. 2013;113(2):96–103. 669–670
19. Méndez A, Mendoza L. A network model to describe the terminal differentiation of B cells. *PLoS computational biology*. 2016;12(1):e1004696. 671–672
20. Offermann B, Knauer S, Singh A, Fernández-Cachón ML, Klose M, Kowar S, et al. Boolean modeling reveals the necessity of transcriptional regulation for bistability in PC12 cell differentiation. *Frontiers in genetics*. 2016; p. 44. 673–675
21. Abou-Jaoudé W, Monteiro PT, Naldi A, Grandclaude M, Soumelis V, Chaouiya C, et al. Model checking to assess T-helper cell plasticity. *Frontiers in bioengineering and biotechnology*. 2015;2:86. 676–678
22. Grieco L, Calzone L, Bernard-Pierrot I, Radvanyi F, Kahn-Perles B, Thieffry D. Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS computational biology*. 2013;9(10):e1003286. 679–681
23. Flobak Å, Baudot A, Remy E, Thommesen L, Thieffry D, Kuiper M, et al. Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS computational biology*. 2015;11(8):e1004426. 682–684
24. Paulevé L, Kolčák J, Chatain T, Haar S. Reconciling qualitative, abstract, and scalable modeling of biological networks. *Nature Communications*. 2020;11(1):1–7. doi:10.1038/s41467-020-18112-5. 685–687
25. Guttula PK, Monteiro PT, Gupta MK. A Boolean Logical model for Reprogramming of Testes-derived male Germline Stem Cells into Germline pluripotent stem cells. *Computer Methods and Programs in Biomedicine*. 2020;192:105473. 688–691
26. Cohen DPA, Martignetti L, Robine S, Barillot E, Zinovyev A, Calzone L. Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration. *PLoS Computational Biology*. 2015;11(11). doi:10.1371/journal.pcbi.1004571. 692–694

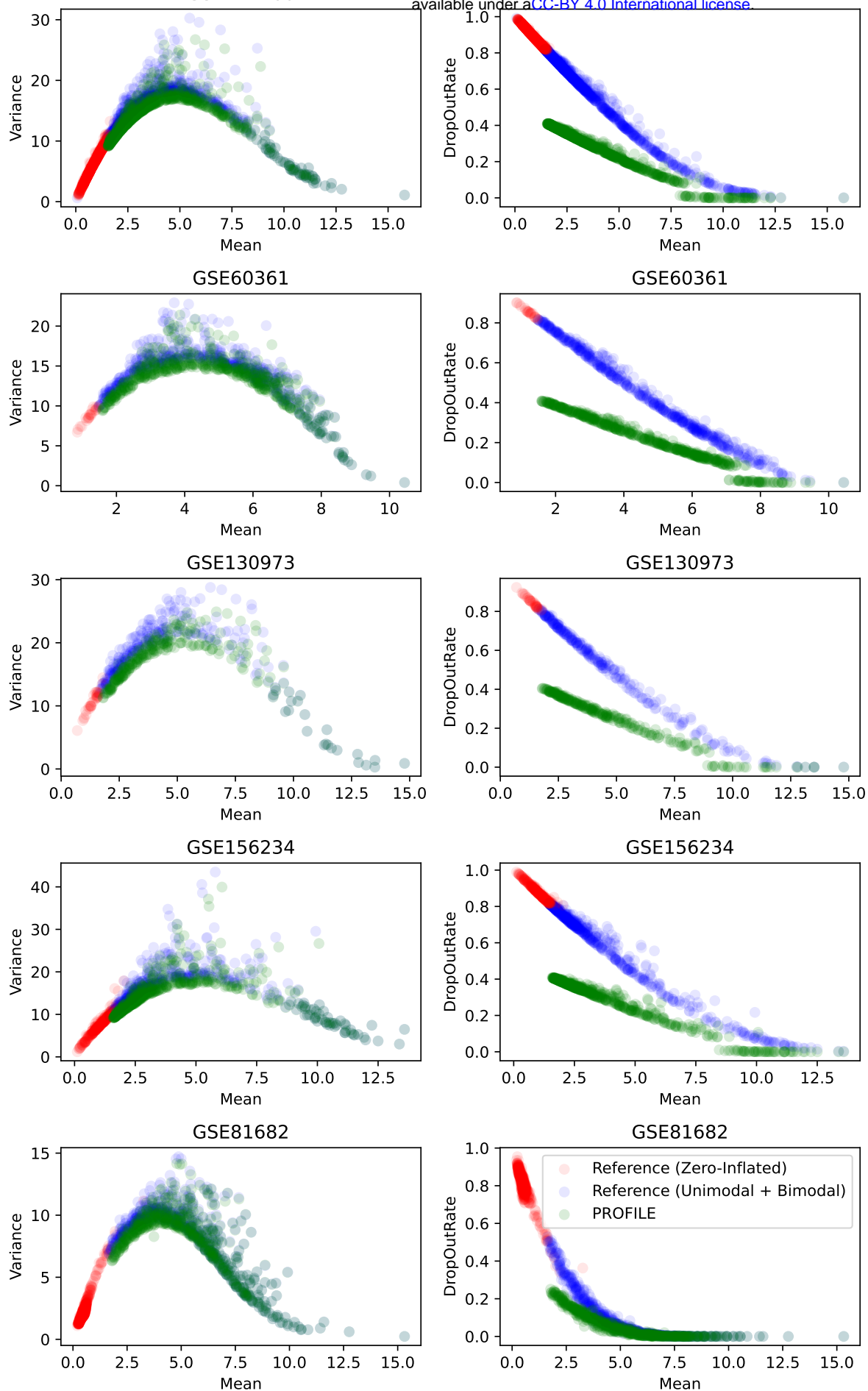
27. Li Y, Jann T, Vera-Licona P. Benchmarking time-series data discretization on inference methods. *Bioinformatics*. 2019;35(17):3102–3109. doi:10.1093/bioinformatics/btz036. 696
697
698
28. Jung S, Hartmann A, Del Sol A. RefBool: A reference-based algorithm for discretizing gene expression data. *Bioinformatics*. 2017;33(13):1953–1962. doi:10.1093/bioinformatics/btx111. 699
700
701
29. Zappia L, Phipson B, Oshlack A. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*. 2017;18(1):1–15. doi:10.1186/s13059-017-1305-0. 702
703
704
30. Assefa AT, Vandesompele J, Thas O. SPsimSeq: Semi-parametric simulation of bulk and single-cell RNA-sequencing data. *Bioinformatics*. 2020;36(10):3276–3278. doi:10.1093/bioinformatics/btaa105. 705
706
707
31. Baruzzo G, Patuzzi I, Di Camillo B. SPARSim single cell: A count data simulator for scRNA-seq data. *Bioinformatics*. 2020;36(5):1468–1475. doi:10.1093/bioinformatics/btz752. 708
709
710
32. Dibaenia P, Sinha S. SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *Cell Systems*. 2020;11(3):252–271.e11. doi:10.1016/j.cels.2020.08.003. 711
712
713
33. Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ. Transforming Boolean models to continuous models: Methodology and application to T-cell receptor signaling. *BMC Systems Biology*. 2009;3:98. doi:10.1186/1752-0509-3-98. 714
715
716
717
34. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*. 2020;17(2):147–154. doi:10.1038/s41592-019-0690-6. 718
719
720
35. Seçilmiş D, Hillerton T, Sonnhammer ELL. GRNbenchmark - a web server for benchmarking directed gene regulatory network inference methods. *Nucleic Acids Research*. 2022;50(W1):W398–W404. doi:10.1093/nar/gkac377. 721
722
723
36. Kim H, Gelenbe E. Stochastic gene expression modeling with hill function for switch-like gene responses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2012;9(4):973–979. doi:10.1109/TCBB.2011.153. 724
725
726
37. Gonze D, Abou-Jaoudé W. The Goodwin Model: Behind the Hill Function. *PLoS ONE*. 2013;8(8). doi:10.1371/journal.pone.0069573. 727
728
38. Bottani S, Veitia RA. Hill function-based models of transcriptional switches: Impact of specific, nonspecific, functional and nonfunctional binding. *Biological Reviews*. 2017;92(2):953–963. doi:10.1111/brv.12262. 729
730
731
39. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008;2008(10). doi:10.1088/1742-5468/2008/10/P10008. 732
733
734
40. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*. 2019;29(8):1363–1375. doi:10.1101/gr.240663.118. 735
736
737
738

41. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*. 2019;10(1). doi:10.1038/s41467-019-09234-6. 739-741
42. Beal J, Montagud A, Traynard P, Barillot E, Calzone L. Personalization of logical models with multi-omics data allows clinical stratification of patients. *Frontiers in Physiology*. 2019;10(JAN). doi:10.3389/fphys.2018.01965. 742-744
43. Hartigan JA, Hartigan PM. The Dip Test of Unimodality. *The Annals of Statistics*. 1985;13(1):70–84. 745-746
44. Wang J, Wen S, Fraser Symmans W, Pusztai L, Coombes KR. The bimodality index: A criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Informatics*. 2009;7:199–216. doi:10.4137/cin.s2846. 747-750
45. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*. 2020;11(1):1–9. doi:10.1038/s41467-020-14976-9. 751-752
46. Ye C, Speed TP, Salim A. DECENT: Differential expression with capture efficiency adjustment for single-cell RNA-seq data. *Bioinformatics*. 2019;35(24):5155–5162. doi:10.1093/bioinformatics/btz453. 753-755
47. Kharchenko PV. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*. 2021;18(7):723–732. doi:10.1038/s41592-021-01171-x. 756-758
48. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, Marguerat S, et al. BayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics*. 2020;36(4):1174–1181. doi:10.1093/bioinformatics/btz726. 759-762
49. Vellaisamy P. On the number of successes in independent trials; 1996. 763
50. Tukey JW, et al. *Exploratory data analysis*. vol. 2. Reading, MA; 1977. 764
51. Chevalier S, Froidevaux C, Pauleve L, Zinovyev A. Synthesis of boolean networks from biological dynamical constraints using answer-set programming. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*. 2019;2019-Novem:34–41. doi:10.1109/ICTAI.2019.00014. 765-768
52. Chevalier S, Noël V, Calzone L, Zinovyev A, Paulevé L. Synthesis and Simulation of Ensembles of Boolean Networks for Cell Fate Decision. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2020;12314 LNBI:193–209. doi:10.1007/978-3-030-60327-4_11. 769-773
53. Stoll G, Viara E, Barillot E, Calzone L. Continuous time boolean modeling for biological signaling: application of Gillespie algorithm. *BMC Systems Biology*. 2012;6:1–18. doi:10.1186/1752-0509-6-116. 774-776
54. Müssel C, Hopfensitz M, Kestler HA. BoolNet-an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*. 2010;26(10):1378–1380. doi:10.1093/bioinformatics/btq124. 777-779
55. Ríos O, Frias S, Rodríguez A, Kofman S, Merchant H, Torres L, et al. A Boolean network model of human gonadal sex determination. *Theoretical Biology and Medical Modelling*. 2015;12(1):1–18. doi:10.1186/s12976-015-0023-0. 780-782

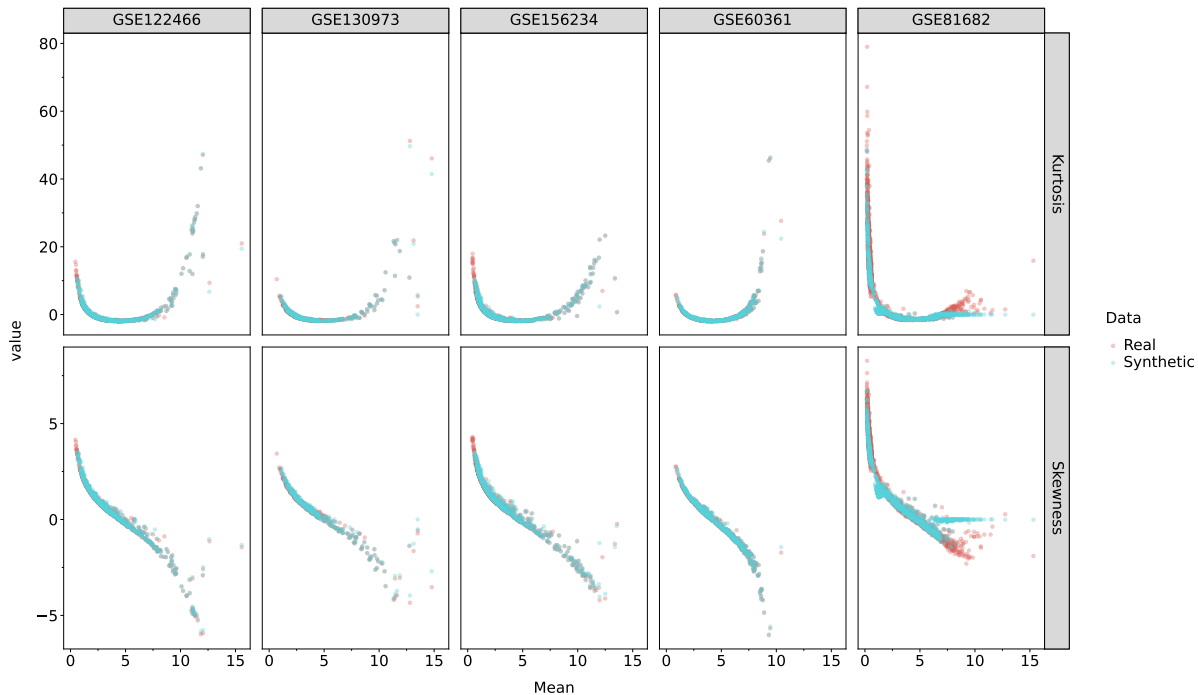
56. Wes McKinney. Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman, editors. Proceedings of the 9th Python in Science Conference; 2010. p. 56 – 61. 783
784
785
57. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference; 2010. 786
787
58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830. 788
789
790
59. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Annotated data. bioRxiv. 2021; p. 2021.12.16.473007. 791
792
60. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nature Biotechnology. 2023;41(5):604–606. 793
794
795
796
doi:10.1038/s41587-023-01733-8.
61. Naldi A, Hernandez C, Levy N, Stoll G, Monteiro PT, Chaouiya C, et al. The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks. Frontiers in Physiology. 2018;9:680. 797
798
799
800
doi:10.3389/fphys.2018.00680.
62. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Hilda Ye B, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nature Genetics. 2016;48(8):838–847. 801
802
803
804
doi:10.1038/ng.3593.



S1 Fig. Example of distribution of rate parameters and dropout probabilities learnt by scBoolSeq. Left: Distribution of rate parameters λ estimated on dataset GSE122466. **Right:** Dropout probabilities computed between the minimum and maximum values of a sample from the parametric distributions corresponding to the same dataset. Each line corresponds to an individual gene.

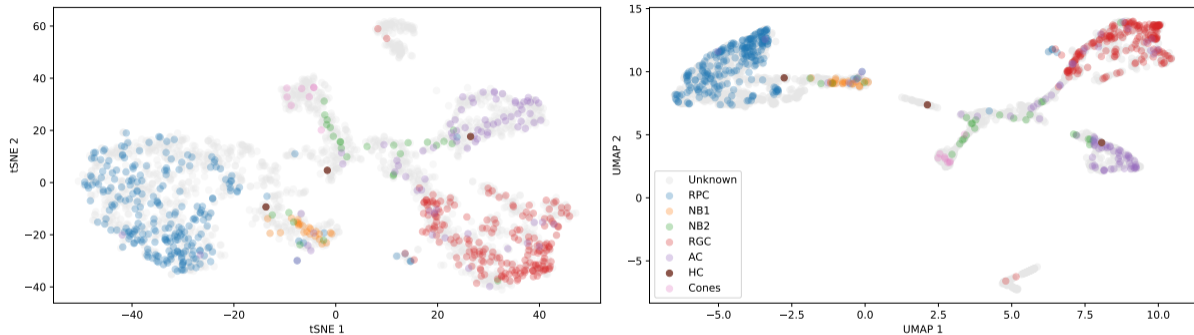


S2 Fig. Mean - Variance and Mean - DropOutRate relationships of HVGs using PROFILE parametric distributions for bimodal and unimodal genes on selected scRNA-Seq datasets. Each green point represents the average of 100 independent replicates with the same sample size as the reference dataset.

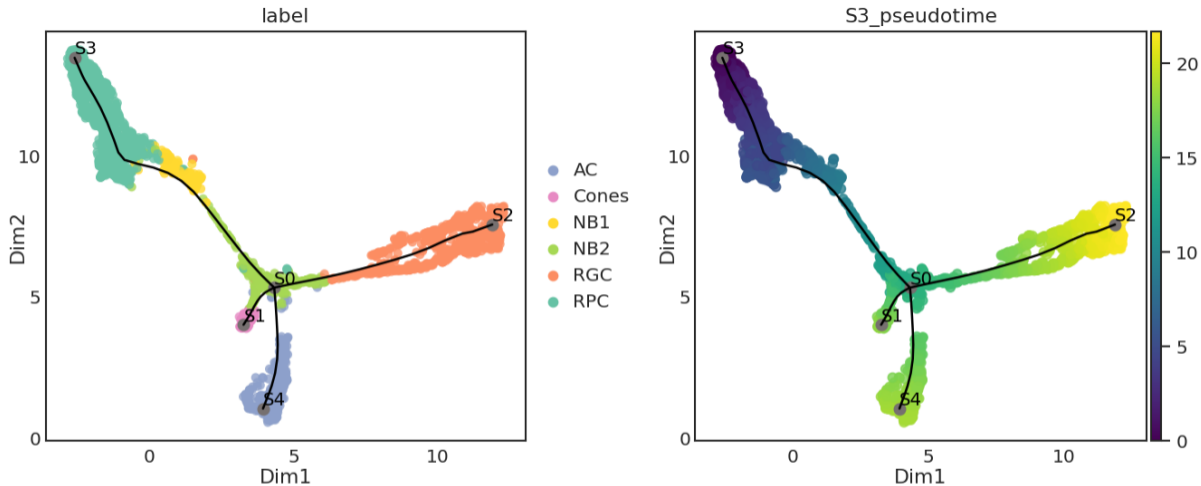


S3 Fig. Correlation between higher moments of real pseudocount data and from data generated from distributions and dropout model learnt by scBooSeq on selected scRNA-Seq datasets

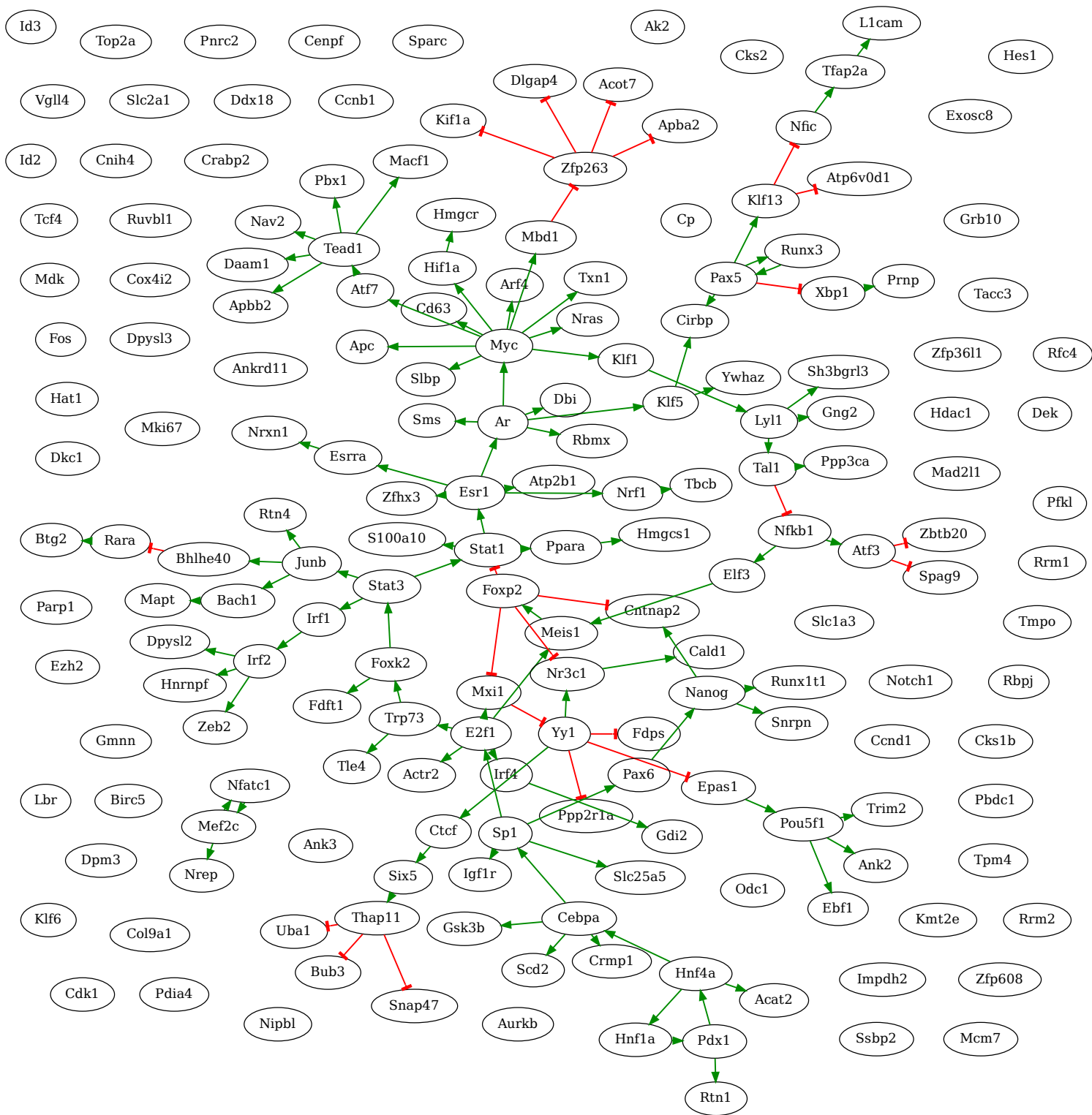
Cell labels determined by Boolean marker gene signatures



S4 Fig. Position of cells classified using scBoolSeq binarisation and prior-knowledge markers. t-SNE and UMAP projections trained on the top 25 principal components (log pseudocount matrix). Colours indicate cell identities determined by binary value of known markers (see Table 1 of main text).



S5 Fig. Result of trajectory reconstruction using STREAM on early-born retinal neurons scRNA-Seq data. UMAP projection of the first 25 principal components to 3 dimensions (only 2 are shown). The cluster labels are determined by the majority label of unambiguous cell types identified via SCBOOLSEQ binarisation.



S6 Fig. Influence graph of sparser Boolean networks learnt using BoNesis from qualitative dynamics of case study obtained with scBoolSeq binarisation. This graph comprises 184 nodes forming Boolean networks that can reproduce the Boolean dynamics of early-born retinal neurons differentiation process. This graph is a subgraph of the input DOROTHEA TF-TF interaction database. Green arrows indicate positive regulations, red arrows indicate negative regulations. Nodes without predecessors indicate nodes with constant function in the Boolean networks. Thus, the Boolean state of these nodes is identical in all stable states, and is in opposite state in the precursor state RPC.

```
# Domain of possible Boolean networks from input GRN
pkn = bonesis.domains.InfluenceGraph(
    pkn_biggest_weakly_connected_component, maxclause=8, allow_skipping_nodes=True, canonic=False
)
# BoNesis setup
bo = bonesis.BoNesis(pkn, binarised_data)
# trajectory from RPC to NB1 and then to NB2
~bo.obs("RPC") >= ~bo.obs("NB1") >= ~bo.obs("NB2")
# trajectory from NB2 to stable state Cones
~bo.obs("NB2") >= bo.fixed(~bo.obs("Cones"))
# trajectory from NB2 to stable state RGC
~bo.obs("NB2") >= bo.fixed(~bo.obs("RGC"))
# trajectory from NB2 to stable state AC
~bo.obs("NB2") >= bo.fixed(~bo.obs("AC"))

# explain dynamics of as much as genes as possible
bo.maximize_nodes()
# rely on the state changes of as less as genes as possible
bo.maximize_strong_constants()
# give access to the genes whose dynamics is necessary to obtain the dynamical properties
view = bonesis.NonStrongConstantNodesView(bo, mode="optN")
```

S7 Fig. Python code snippet showing usage of BoNesis for the inference of Boolean networks for the retinal differentiation case study. See S1 Code/3. - Retinal Differentiation BN Inference for full pipeline.