Statistical characterization of the dynamics of Glioblastoma Multiforme subtypes through parameters estimation of the epigenetic landscape

Marcos Guilherme Vieira Junior[a,*], Flávia Raquel Gonçalves Carneiro[b,c,d], Adriano Maurício de Almeida Côrtes[e,f], Nicolas Carels[g], Fabrício Alves Barbosa da Silva[h,*]

[a]*Graduate Program in Computational and Systems Biology, Instituto Oswaldo Cruz (IOC), Oswaldo Cruz Foundation (FIOCRUZ), 20231050, Rio de Janeiro, Brazil*
[b]*Center of Technological Development in Health (CDTS), FIOCRUZ, 21040-361, Rio de Janeiro, Brazil*
[c]*Laboratório Interdisciplinar de Pesquisas Médicas, IOC, FIOCRUZ, 20231050, Rio de Janeiro, Brazil*
[d]*Program of Immunology and Tumor Biology, Brazilian National Cancer Institute(INCA), 20231050, Rio de Janeiro, Brazil*
[e]*Department of Applied Mathematics, Institute of Mathematics, Federal University of Rio de Janeiro, RJ, 21941-909, Brazil*
[f]*Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, RJ, 21941-972 Brazil*
[g]*Laboratory of Biological System Modeling, CDTS, FIOCRUZ, 21040-361, Rio de Janeiro, Brazil*
[h]*Scientific Computing Program, FIOCRUZ, 20231050, Rio de Janeiro, Brazil*

Abstract

This report aims to enhance the understanding of Glioblastoma Multiforme (GBM) subtypes' dynamics by exploring the association between ergodicity and GBM aggressiveness. We introduce new regulation functions to model the network dynamics and perform parameter estimation using centroids obtained from multiple clustering methods. Our findings demonstrate a high degree of compatibility between centroids of experimental and simulated data, and we thoroughly investigate the ergodicity of the simulations. The transitions between basins provide insights into cancer recurrence and subtype development, while the compatibility between sample and time averages supports the internal ergodicity assumption. This approach could help assess and monitor the evolution of malignant states, serving as a valuable tool for diagnosis and treatment and ultimately contributing to the development of targeted therapies.

Keywords: Glioblastoma Multiforme, epigenetic landscape, parameter sets estimation, ergodicity, gene regulatory network dynamics

1. Introduction

Despite all advances in its comprehension and treatment, cancer is still one of the major causes of death worldwide [1]. Among the different types of cancer, Glioblastoma Multiforme (GBM) is the most common and aggressive type of brain tumor, with an average survival time of around 15 months and only approximately 10% probability of 5-years overall survival [2]. Fortunately, the increasing availability and variety of omics data have enabled deeper investigations that may change this scenario. For instance, multiple studies have identified subtypes of GBM, a fundamental step for choosing the treatment to be applied [2–5]. However, little is known about each subtype's development and evolution, partly due to disregarding dynamic effects in classification studies.

In the last decades, scientific advances have been changing paradigms regarding cancer. Cancer can be understood as a complex series of events leading to a disease of state instead of a mere consequence of individual mutations [6, 7]. In this context, we call *state* the biochemical conditions of a cell such that 'state disease' is characterized as the progression of

*Corresponding authors.
Email addresses:
`marcosvieira.research@gmail.com` (Marcos Guilherme Vieira Junior), `fabricio.silva@fiocruz.br` (Fabrício Alves Barbosa da Silva)

states of a complex system. Instead of following the old paradigm, which relied on a linear progression by considering the sequence of mutations in a one-dimensional manner, the new paradigm embraces a statistical perspective that accounts for the complex interactions and dependencies among multiple factors in a 'multidimensional' trajectory. One possibility of formalizing this process is from the metaphor proposed by C. Waddington in 1957, in which cellular differentiation is described as a sphere rolling down an (epi)genetic[1] landscape composed of peaks and valleys, with the highest points related to the undifferentiated phenotype.

In the Waddington (epi)genetic landscape, a given differentiated state is reached through developmental paths that are influenced by external factors, the cell's historical events, such as lineage, gene expression patterns, and epigenetic modifications, as well as extrinsic noise at a larger scale [8]. According to this approach, cancer would be associated with one or more (defective or malignant) attractor states in the (epi)genetic landscape. These malignant attractors can either pre-exist, hidden within the landscape, or emerge due to genetic and epigenetic alterations. In both scenarios, they are undesirably reached due to the inherent stochastic fluctuations of biological systems at the biomolecular scale.

Different mathematical and computational approaches have demonstrated the possibility of modeling cancer using this concept of (epi)genetic landscape [9–14]. Thus, understanding the temporal dynamics of these models and their local and global stability is paramount to describing cancer evolution. However, such models usually depend on temporal data to estimate the parameters associated with the system's behavior, making their application difficult or even unfeasible. Thus, developing quantitative methods that do not require time series data is necessary.

In this sense, some methods have proposed the inference of temporality (the ordered succession of events) from the definition of the so-called pseudotime [15]. The inference of a pseudotime through single-cell RNA-seq data can be a way to estimate the transition between attractors. However, its use to infer the parameters and dynamics of cancerous states can be problematic. Events of mutations, cell cycle deregulation, and cellular heterogeneity reduce the confidence in the a priori knowledge, usually related to healthy cells, complicating the cancer dynamics characterization. Nevertheless, the loss of constraints that results from this uncertainty allows its stochastic modeling.

The difficulty in acquiring temporal measures is not unique to the system in question, but it is an old problem for which different theories and hypotheses have been produced [16]. For example, the concept of ergodicity allows applying mathematical and computational methods to obtain the parameters without the need for time series. Ergodicity is a property that characterizes the statistical behavior of the system by studying how it explores the phase space over time. Ergodicity can be present in different ways. Here we consider a weaker version and the concept of internal ergodicity [17]. We call weaker ergodicity the one in which the time averages converge to the ensemble averages in some statistical sense. Internal ergodicity refers to the presence of the weaker ergodicity in each component of the system (see S1.3 for more details) [18], with the amount of time spent in each of them proportional to the probability of reaching each configuration. With the weak ergodicity within each attractor, averaging time along a single trajectory provides a good approximation of its ensemble average, mutatis mutandis. Likewise, the ensemble average would provide a good approximation of averaging time along a single trajectory, with the necessary precautions.

In some cases, it is possible that exists regions of space that the system cannot access due to constraints or other limitations. In this report, we assume metric transitivity, which allows us to connect any two points in the phase space by a trajectory of minimum length (geodesic). In fact, in cases like cancer, noise enhances the exploration of the phase space and can lead to the breakdown of barriers between different basins of attraction, promoting ergodicity. In this sense, the GBM subtype's aggressiveness can be seen as a temporal progression. We expect there are dimensions of the gene space in

[1] Since the RNA sequencing data provide an estimate of expression after genetic and epigenetic regulations, and it is not possible to verify the contributions of each explicitly, we chose to use parentheses in (epi)genetics.

which the equivalence between the sample averages of several subtypes (aggressiveness scale) and time average (staging) is a good approximation. Based on these hypotheses, this work sought compatibility between a stochastic model of (epi)genetic landscapes and single-cell RNA sequencing (scRNA-seq) data of four subtypes of different aggressiveness levels based on the interactions of a gene regulatory network (GRN) of GBM.

Apart from the lack of time measurements, other challenges may limit practical applications, such as network construction and the number of parameters involved. The network must be constructed so that any simplification does not significantly impact its biological representativeness, especially when considering GRN dynamics. The parameter assumptions and interpretations should be as close as possible to what is known from biological data. This introduces considerable complexity to the problem in terms of parameter quantity and the methods used to estimate them.

This report proposes a method to estimate the parameter values corresponding to each network edge. We consider clusters of gene expression found within and across different scRNA-seq datasets as potential indicators of cell types. These clusters may reflect underlying stability in the system, suggesting that they could represent distinct cell states or phenotypes. While not all clusters necessarily correspond to stable states, we hypothesize that the parameter fitting process could differentiate between such states. By using the centroids of these clusters in the parameter fitting process, we propose an analysis of the low space dynamics of specific markers, as demonstrated in studies such as Sáez et al. [19]. We also explore variations of the deterministic dynamics with potential constraints on the temporal evolution of the system to address possible network limitations.

Variations in the deterministic dynamics were investigated as a basis for the implementation of an effective noise model that captures the behavior observed in experimental data. This was performed by considering the regulation functions as modified Hill functions since standard Hill functions pose limitations, such as the same regulation strength for all targets of the same gene and the monotonic behavior. These factors make it challenging to model biological mechanisms such as negative feedback loops. In addition, using data from multiple patients, the usual Hill function limits the inter-patient heterogeneity modeling because there is only a dependency on the regulating gene, not on the regulated one. Another positive side of the changes in the deterministic dynamics is that the model becomes constrained according to experimental data. This provides the model with more robust dynamics when compared with biological observations.

Biological data can be understood as a combination of deterministic dynamics and stochastic noise. However, multiple combinations of these deterministic and stochastic contributions may result in the same or similar observable experimental distributions, as shown by [20]. To better understand the underlying dynamics of the system and the sources of variability, we employed a computational method in conjunction with experimental data. We used statistical criteria to select our analysis's most representative deterministic and stochastic parameters.

In summary, this report provides important insights into quantifying the (epi)genetic landscape for the case of GBM and its subtypes, achieving a more accurate quantification than previous approaches. By proposing ways of modeling the dynamics of available scRNA-seq data, our work lays the ground for future investigations and potential practical applications. Furthermore, it contributes to developing studies on a biological system's possible long-term behavior and stability by considering its statistical properties.

2. Materials and Methods

2.1. Model background

The regulation of gene expression is a complex process that involves multiple layers and mechanisms. One possible measurement of gene expression is the number of mRNA molecules that effectively translate into proteins. The expression profile is a dynamic feature, changing in time according to cell types and characteristics. By considering a vector $\mathbf{X} = (X_1, X_2, ..., X_N)$ with N being the total number of variables and each vector component representing the quantification of mRNA molecules, the cell state can be modeled using system dynamics

theory. The basis of the gene regulation dynamics modeling is its associated deterministic differential equations system, an autonomous system of ordinary differential equations $\dot{\mathbf{X}} = \mathbf{F}(\mathbf{X})$, containing information about the temporal trajectory driven by the interaction forces between each of its components [21].

Although there are several possible functions to parameterize the interactions of a nonlinear model, it is common to use sigmoidal functions, among them the Hill function. It has many desired characteristics and experimental validation [22]. An example of the driving force F can be seen in Wang et al. [9], with a more general form described by:

$$F_i = -k_i X_i + \sum_{j \in \mathcal{A}_i} \frac{a_{ij} X_j^{n_{ij}}}{S_{ij}^{n_{ij}} + X_j^{n_{ij}}} + \sum_{j \in \mathcal{I}_i} \frac{b_{ij} S_{ij}^{n_{ij}}}{S_{ij}^{n_{ij}} + X_j^{n_{ij}}}, \quad (1)$$

where, for each gene i, represented by the component X_i, the index sets $\mathcal{A}_i$ and $\mathcal{I}_i$ represent the genes that interact with gene i through activation and inhibition, respectively. The value j represents the edge that bridges the regulation of transcription factors interacting with their target gene promoters. Note that in the case of self-activation or self-inhibition one has $i \in \mathcal{A}_i$ or $i \in \mathcal{I}_i$, respectively. The parameter S denotes the value where the Hill function reaches its maximum inclination, n represents the intensity of the transition, a is the activation coefficient, b is the inhibition coefficient, and k is the self-degradation constant. When a is a self-activation or b is a self-inhibition parameters, they will be denoted by sa and sb, respectively. The parameters k, a, and b have units of time^{-1}, while the remaining parameters are dimensionless.

As seen in the equation (1), in the most general form, the gene activation (a) and inhibition (b) parameters may vary for each interaction or even as a function of time (non-autonomous system). In addition, sigmoid coefficients may (i) be constant, (ii) vary according to interactions or some proposed functions, or (iii) present a time dependence. In this model, the gene inhibition is given by constraining its basal expression, as can be seen in the positive sign of the inhibition term, with the higher inhibitions obtained by lower values of b.

Although deterministic differential equations are adequate to study general behavior, biological systems are inherently stochastic. Due to thermal fluctuations in molecular positions, rotation movements, and varying conditions that affect the likelihood of interactions, these processes exhibit a probabilistic nature. Consequently, the number of molecules over time is not represented by a smooth curve but instead follows a fluctuating, noisy pattern. A common way to model this stochasticity is through the Chemical Master Equation (CME), a Markovian model that captures the probabilistic nature of molecular interactions. However, solving the CME can be computationally challenging, especially for large systems. An alternative is to use Langevin dynamics [23], which serves as an approximation of the CME, described by a deterministic term $F(x)$ and a stochastic term $\xi(t)$. In Langevin dynamics, the timescale of $F(x)$ is much larger than that of $\xi(t)$, allowing the latter to be treated as random fluctuations (without memory). The dynamics became:

$$\frac{d\mathbf{x}(t)}{dt} = F(x) + \xi(t) \,, \quad (2)$$

where $x(t)$ is the expression level as a function of time[2] relative to random variables of X, $F(x)$ is the deterministic term representing regulation due to network interactions, and $\xi(t)$ is the stochastic term with average $\langle \xi(t) \rangle = 0$ and amplitude given by its autocorrelation function $\langle \xi_i(t) \xi_j(t') \rangle = 2D_{ij}\delta_{ij}\delta(t - t')$, with D being the diffusion coefficient and representing a fluctuation scale factor.

With the presence of fluctuations, probability distributions model gene expression levels, and the temporal evolution $p(x, t)$ is described by the Fokker-Planck equation (eq S2). This equation provides a continuous approximation to the CME, capturing molecular diffusion kinetics across an epigenetic landscape. However, since the Fokker-Planck equation's driving force and noise components are unknown, an alternative approach to studying the system is to consider a stochastic differential equation (SDE) with possible deterministic and stochastic components. In this context, we consider a system described by [14]:

$$dX = \mu(X, t)\, dt + \sigma(X, t)\, dW \,, \quad (3)$$

[2] Implicit dependence.

with drift $\mu(X,t)$, noise parameter $\sigma(X,t)$ and Wiener standard process dW. We assumed $\mu(X,t) = F(X)$ by considering the drift due to the driven force. The noise $\sigma(X,t)$ can be divided into two major contributions: (i) intrinsic, which is related to the system's internal dynamics, and (ii) extrinsic, which is due to the effects of the environment/microenvironment. We considered a multiplicative noise $\sigma(X,t) = g(X,t)$, so that the fluctuations may be described by different time scales and constrained by the defined regulation function. The complete definition of the system is given both by the parameters of the deterministic and the stochastic components, with the specifics of the multiplicative noise and the regulation function to be detailed later in this paper.

2.2. Single cell RNA-seq data

In recent years, scRNA-seq has emerged as a powerful tool to study gene expression profiles at the individual cell level, enabling the investigation of cellular heterogeneity and the identification of distinct cell subpopulations. This technology has been particularly valuable for studying GRNs. It provides insights into the complex interactions between genes and the possible regulatory mechanisms that drive cell-type-specific gene expression patterns.

We used data generated and analyzed by Darmanis et al. [24], corresponding to RNA sequencing with single-cell resolution from patients with different subtypes of GBM. The study evaluated the tumor nucleus heterogeneity concerning the tumor mass periphery. The data refer to four patients with primary GBM and negative *IDH1* (without mutations in the *IDH* gene), comprising 3589 cells that passed the quality control. Samples were analyzed both from the tumor's center and the periphery, composed of neoplastic cells and related to different cell types of the central nervous system (vascular, immune, neuronal, and glial).

The analytic methodology of Darmanis et al. [24] identified cellular clusters from the dimensionality reduction with tSNE in the dissimilarity matrix, followed by clustering using the k-means method. These results defined the cell types by analyzing the most expressed genes in each cluster. Results were combined with data from healthy tissue samples used for classification. The remaining clusters were defined as neoplastic, for which the location was found to be mainly in the tumor nucleus, with high expression of *EGFR* and *SOX9*. Comparison to additional data from healthy brain tissue and GBM bulk RNA-seq validated results. It was observed that astrocytes were the only cell type absent in the tumor nucleus. In addition, an expression pattern for tumor cells in peripheral regions was verified for all patients analyzed [24].

2.3. GRN construction and implementation

The challenge proposed here is based on the mathematical and computational modeling of a GRN constructed from the regulatory interactions (edges) between the genes (vertices) of the system supposed to be representative of the case under analysis (GBM). This way, we established a biological criterion for defining the network edges and considered only molecular mechanisms that directly or indirectly affect the number of messenger RNA molecules. For this model, we considered *direct interactions* as those at the transcriptional level, given by the binding of transcription factors (TF) to the promoter of the target genes. These interactions directly affect the amount of mRNA and are represented by a direct connection between the transcription factor vertex and the vertex representing the targeted gene. On the other hand, *indirect interactions* are those that prevent or affect the binding of a TF to a gene promoter (e.g., ubiquitination of a TF leading to its degradation). Unlike direct interactions, the biological effects of indirect interactions are not immediate, and their consequences on the number of mRNAs need to be evaluated before they can be adequately represented within the model.

In this study, we first identified genes and markers associated with known subtypes of GBM based on the classification by Verhaak et al. We then utilized the MetaCore platform to search for interactions among these genes and markers [25]. Specifically, we employed the MetaCore transcription regulation network construction algorithm to build the network, identifying new vertices that bridge the initial gene list (provided in Supplementary Table rede_GenesList.xlsx). However, the output generated consisted of several disjoint subnetworks, with the initial genes scattered among them. We extracted the

initial and additional genes from these subnetworks to address this issue and created a new input list. This new list was then used to generate a connected network, ensuring that all the genes of interest were included in a single structure (Figure S1).

An algorithm in R [26] was developed to integrate scRNA-seq expression data into the network generated by MetaCore [27]. The data preprocessing was performed using the Seurat package [28], from which a *sctransform* normalization was applied to reduce technical bias and recover biologically significant distributions [29, 30]. Cell cycle effects were not removed, as such information may lose its accuracy in tumor cells [31]. The algorithm then performed the following operations: (i) selected interactions classified as *Transcription Regulation*, (ii) intersected the genes of the network with those present in the data, and (iii) removed the genes that were associated with null values, even if they were present in the network. As a result, the network may change with variations in data, whether due to the selected cell types or patient ID.

To specify our working network, we used the cell type classifications proposed in Darmanis et al. [24]. They identified various cell types present in GBM samples, such as astrocytes, oligodendrocytes, and neurons, among others, through clustering and other methods. Our study focused on neoplastic cells localized in the tumor core. We explored the expression value distributions for each patient in our dataset. To increase the probability of sampling over the entire phase space of GBM, we decided to obtain the landscape for all patients simultaneously instead of analyzing individual patients separately. This approach allows for better characterization of GBM by observing attractors related to the four subtypes of GBM, while still reflecting the specific characteristics of each individual. By filtering for neoplastic cells located in the tumor core, we avoided incorporating the different characteristics observed for neoplastic cells present in the periphery of the tumor.

To enable the automatic integration of the network with mathematical and computational models, we developed an algorithm, available in the provided code repository [32], that simplifies the process. The algorithm takes the network as input in a tabular format, representing the connectivity list between each node. It then processes the input and converts the tabular data structure into two directional graphs (digraphs), one for activation interactions and another for inhibition interactions, each represented by its adjacency matrix. With these transformed data structures, equation (1) can be written as:

$$\mathbf{F} = -\mathbf{k}\mathbf{X} + \text{rowsum}(\mathbf{M}^a \odot \mathbf{V}^a) + \text{rowsum}(\mathbf{M}^b \odot \mathbf{V}^b)\,, \tag{4}$$

with $\mathbf{k} = \text{diag}(k_1, \ldots, k_N)$ a diagonal matrix, $\mathbf{M}^a$ the activation matrix with entries $(\mathbf{M}^a)_{ij} = a_{ij}$, $\mathbf{M}^b$ the inhibition matrix with entries $(\mathbf{M}^b)_{ij} = b_{ij}$, $\mathbf{V}^a$ the activation Hill functions matrix with entries

$$(\mathbf{V}^a)_{ij} = \frac{X_j^{n_{ij}}}{S_{ij}^{n_{ij}} + X_j^{n_{ij}}}, \tag{5}$$

$\mathbf{V}^b$ the inhibition Hill functions matrix with entries

$$(\mathbf{V}^b)_{ij} = \frac{S_{ij}^{n_{ij}}}{S_{ij}^{n_{ij}} + X_j^{n_{ij}}}. \tag{6}$$

The $\odot$ denotes the Hadamard product (element-wise matrix product), and rowsum$(\cdot)$ returns the vector with the row-wise sums of the matrix.

Utilizing the symbolic capabilities of Wolfram Mathematica, the activation, self-activation, inhibition, and self-inhibition parameters (a, sa, b, and sb) are represented as arbitrary values. With this approach, during the parameter estimation process, these symbolic values can be replaced with numerical values (e.g., 1) using Mathematica's ReplaceAll function (e.g., /. a → 1). Once the parameter estimation is completed, the system behavior can be further explored by substituting the parameters with different arbitrary values (e.g., /. a → 0.7) to observe the effects of perturbations.

2.4. Ergodicity, clustering, and distributions

In the context of dynamical systems theory, we can regard different cell states as initial conditions and their trajectories representing the multidimensional evolution of gene expression over time. By sampling a range of these initial conditions, the system's trajectories evolve towards specific regions within the state space, known as *basins of attraction*. Within these regions, the system converges to absorbing

states called attractors. In deterministic scenarios, trajectories converge to equilibrium points, which can be stable or unstable. However, when the system is subject to noise within a particular amplitude range, the final states usually oscillate around stable states, eliminating concerns about unstable equilibrium points.

In practice, monitoring the temporal evolution of individual cells is extremely challenging due to experimental and technological limitations. As a result, the available data predominantly consists of static representations of the system's state. Although this data may seem to lack dynamic information necessary for a comprehensive understanding of the system's behavior, a sample of the system dynamics can still provide valuable insights when viewed as a collection of cell states obtained at different time points. For instance, in ergodic systems, the temporal and sample averages are equivalent, allowing the dispersion of trajectories to be inferred from the differences between each process repetition. Ergodicity implies that, on a given time scale, each trajectory in the system passes arbitrarily close to every point in the phase space, thus making it possible to describe the system using a statistical distribution.

In healthy cells, standard time scales and biological constraints limit the system such that assuming ergodicity may not be reasonable. However, in the case of GBM, the accumulation of mutations and rapid progression of the malignant state may relax the system's constraints, making it plausible to expect ergodic behavior. In this context, our goal is to extract meaningful information from the static data to better understand the behavior of GBM cells. To achieve this, we begin with Equation (1) and consider the centroids of data clusters (see Section 2.2) as approximations of the equilibrium states. Under this assumption, the driving forces (rates of change) of expression levels will be approximately zero, resulting in a system that satisfies:

$$F = \frac{dX}{dt} \approx 0 \,, \tag{7}$$

which enables a characterization of the system stability by the parameter estimation using experimental data.

In this approach, we regard the centroids as the means of Gaussian distributions characterizing the gene expression levels, enabling us to search for the attractors of GBM subtypes. We utilize the Waddington (epi)genetic landscape concept to study state transitions and the likelihood of finding specific gene expression distributions throughout the phase space.

However, experimental data inherently include noise originating from the experimental process, such as sample preparation and technical variability, which complicates the extraction of information specific to biological processes. Despite these challenges, we regard the correlation between experimental measurements and the underlying biological truth as adequate, given that experimental data are extensively utilized for diagnosis and treatment. We rely on data generated and analyzed in Darmanis et al. [24] (see Section 2.2). With the central limit theorem and the law of large numbers, we assume a Gaussian distribution for gene expression [9] as:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right] \,, \tag{8}$$

with $\bar{x}$ representing the data average and σ the standard deviation, both obtained through data analysis.

By considering Gaussian distributions, the observed experimental values would represent states traversing orbits around the mean value. While certain clusters could characterize attractors associated with different cell types and subtypes, it is important to note the possibility of spurious clusters that may not represent true attractors, as they could arise due to artifacts, noise, or other factors unrelated to the underlying biological processes. Additionally, some clusters might correspond to metastable states, which represent transient configurations in the phase space rather than stable attractors. This scenario is suitable for parameter estimation, as it allows for establishing a relationship between expression values and the variation rates in the attractor's centroids (null or close to zero).

From this assumption, we considered clustering methods to estimate the averages (centroids) and deviations as a good representation of probability distributions and statistical descriptions of the system dynamics. The observed variations (experimental and simulated data) would be deviations related to the states close to equilibrium and contained in the

respective basins of attraction for given fluctuations. We used dimensions relative to the markers of GBM subtypes to better characterize the clusters. They tend to have values varying at more restricted intervals and not ranging from zero to greater expression values, such as genes related to the cell cycle or others with high noise, making them a reasonable choice for clusterization and landscape visualization.

In this step, the data were imported into Mathematica [33], which was used to analyze the gene expression histograms for each gene in the network and verify their distributions. Four marker genes, each associated with one of the GBM subtypes, were selected: *EGFR* (classic), *IDH1* (proneural), *NEFL* (neural), and *CD44* (mesenchymal). These marker genes serve as representative indicators of the respective subtypes and facilitate the identification of subtype-specific patterns.

The number of clusters was determined through dimensionality reduction (t-SNE), followed by applying various clustering methods, including k-means and Neighborhood Contraction (NbC), to identify the clusters. K-means was chosen because it is a widely-used and computationally efficient clustering algorithm that works well for partitioning data into spherical clusters. NbC was also selected as it is a density-based clustering method that can identify clusters of varying shapes and densities, providing a more flexible representation of the data [34, 35].

Cells classified into each cluster were visualized for each combination of the four marker genes to evaluate the clustering quality and data distribution in these dimensions. The standard deviations of each cluster were calculated using their respective averages, and these values were stored, along with the relative percentage of cells in each cluster. This approach enabled a comprehensive data assessment and facilitated the identification of gene expression patterns between GBM subtypes.

Within the context of gene interactions, changes in one gene can influence the behavior of others, leading to complex and interconnected dynamics. However, when the system is in a steady state, the statistical distribution already accounts for the interactions among genes, allowing for the expression of the system's behavior by considering each gene independently.

In this scenario, the mean-field approximation emerges as a powerful simplification technique. Focusing on the moments that describe the distributions resulting from the system's interactions as a whole, allows for a more tractable analysis while still capturing the essential behavior of the underlying dynamics. As the system is in the basins of attraction (described by a stationary distribution), the mean values and probability distributions can describe the dynamics in stochastic terms, despite the intricate interplay between genes.

Employing the mean-field approximation simplifies the complex network of individual interactions through an effective interaction related to a mean field, facilitating a more accessible understanding of the system's behavior. Consequently, the probability of each attractor α, initially dependent on all genes, can be estimated using this approximation [11]:

$$P_\alpha = \prod_{i=1}^{n} P_{\alpha,i} \ , \tag{9}$$

with i the index of genes and n the number of genes.

The probability of a cell being in some state will be given by the total probability P_{ss} of the steady state [11]:

$$P_{ss} = \sum_{\alpha=1}^{\alpha_{max}} w_\alpha P_\alpha \ , \tag{10}$$

where w_α is the percentage of cells in each attractor, and α_{max} is the maximum number of attractors (clusters) found by the method.

The total steady-state probability (equation (10)) can be estimated for both experimental and simulated data. For experimental data, the probability can be estimated by clustering cell types or subtypes under the assumption of ergodicity. For simulated data, the probability can be determined by allowing sufficient time for the system to evolve from sampled initial conditions towards steady states, and then estimating the probabilities at a given time point.

Studying the system's global behavior using (epi)genetic landscapes is based on the flow theory of nonequilibrium dynamical systems [12]. When the steady-state probability P_{ss} is obtained through sampling rather than an analytical solution of the Fokker-Planck equations, it is referred to as a population

landscape. This landscape is defined as a surface U derived from the negative logarithm of the total probability of the system:

$$U = -\ln P_{ss} \,. \tag{11}$$

In equilibrium systems where detailed balance is preserved, the Boltzmann relation holds, and the landscape corresponds to the equilibrium probability. The Boltzmann relation connects the probabilities of states in a system with their energies. Detailed balance refers to the condition where the rates of transitions between states balance each other, ensuring that the system remains in equilibrium. In contrast, in the non-equilibrium case considered here, the trajectories do not strictly follow the gradient due to the presence of the probability flux [9]. This deviation from the gradient complicates the study of state transitions and probabilities, as it introduces additional factors, such as stochastic fluctuations and transient states, that influence the system's dynamics. Consequently, predicting the system's behavior becomes more challenging, as transitions between states may not occur exclusively through the most probable pathways.

Despite these complexities, the (epi)genetic landscape can still provide valuable insights into the behavior of non-equilibrium systems. The landscape obtained in this study is a non-equilibrium landscape, as it represents the behavior of a non-equilibrium system derived from the probability distribution obtained through stochastic simulations. In non-equilibrium (epi)genetic landscapes, the gene states and their associated energies might not follow the Boltzmann relation exactly due to external forces and stochastic fluctuations. Additionally, the landscape may represent transient or metastable states, which persist for some time but are not as stable as true attractor states. Nevertheless, the landscape can still provide valuable insights into the probabilities of different gene states and their transitions, helping to identify these transient or metastable states and to understand the pathways through which the system undergoes these transitions between them.

By investigating the non-equilibrium (epi)genetic landscape, we expect to gain insights into the impact of stochastic fluctuations and transient states on system dynamics, which can help guide further investigations and develop more accurate models for the system's behavior. Using the dimensionless potential U, the landscape quantifies the probabilities of states and transitions. Higher probability states correspond to lower potentials (greater depths), while the barriers between the basins of attraction are related to the time spent in each state. In non-equilibrium systems, understanding the impact of probability flux and its influence on state transitions is crucial for accurately analyzing the system's behavior.

2.5. *Regulation functions*

Hill functions are widely used for modeling regulatory interactions in gene networks, accounting for the influence of activators and inhibitors on regulation strength. The steepness of the Hill function reflects the gene expression's sensitivity to variations in regulatory molecule concentrations. However, typically characterized by constant coefficients for each interaction, they result in a uniform regulation strength when applied to all target genes governed by a specific transcription factor. As monotonic functions, they display an increasing regulation strength in response to increasing transcription factor concentrations, posing challenges for cancer heterogeneity modeling. Additionally, they might face challenges in representing complex regulatory mechanisms when the network is incomplete.

To overcome these limitations, we introduce modified Hill functions, which provide a more flexible framework for modeling the intricate relationships between genes and their regulators, allowing experimental data incorporation to constrain the dynamics. We propose the modified Hill functions given by:

$$(\mathbf{V}^a)_{ij} = \left(\frac{X_j^n}{(h(X_i) + S_j)^n + X_j^n} \right) , \tag{12}$$

for activation, in place of (5), and by:

$$(\mathbf{V}^b)_{ij} = \left(\frac{S_j^n}{(h(X_i) + S_j)^n + X_j^n} \right) , \tag{13}$$

for inhibition, in place of (6). The index i is for the vertex related to the target, j for the vertex affecting the target, and $h(x_i)$ is a modifier function used to obtain shapes with desired properties. To determine the best model, we tested several options for $h(x_i)$,

including the original model with $h(x_i) = 0$. The proposed possibilities were:

$$h_1(x) = x\sqrt{\prod_{\alpha}(|x - \bar{x}_\alpha|)}\,, \tag{14}$$

$$h_2(x) = \left(\frac{1}{0.1x + 0.001}\right) + \sqrt{\prod_{\alpha}(|x - \bar{x}_\alpha|)}\,, \tag{15}$$

$$h_3(x) = \left(\frac{1}{0.1x + 0.001}\right) \cdot \sqrt{\prod_{\alpha}(|x - \bar{x}_\alpha|)}\,, \tag{16}$$

where $\bar{x}_\alpha$ is the average of gene i considering each attractor α. The first term of equations (15) and (16) was chosen to adjust the compatibility for values close to zero, and the constants in the denominator were empirically verified to optimize the results. The constant 0.001 was empirically estimated to adjust the results for $x = 0$.

In addition, we considered:

$$S_j = f \cdot (\max_{\alpha} [X_j^\alpha] + 1)\,, \tag{17}$$

where f is a proportionality constant and $\max_\alpha [X_j^\alpha]$ represents the transcription factor with the highest expression value among all attractors. Adding a unit to the base value of S ensures that the proportionality constant can adjust cases with only zeroes and avoids null denominators. This modification sets the value as a fraction of the maximum gene expression for all attractors found. This assumption is a biological simplification, proposing that regulation intensity is proportional to the highest equilibrium value of the transcription factor. This approach helps capture the effects of weaker transcription factor levels when using a single sigmoid function to represent regulation intensity. Utilizing lower levels could lead to a rapid increase in the sigmoid function from zero, impeding accurate modeling of regulation intensity across various transcription factor concentrations.

These modifications were proposed to impose constraints that are difficult to explicitly add to the system using a graph structure alone. By introducing these changes, we aim to make the model less sensitive to the incompleteness of the gene regulatory network graph and better integrate hidden information within the data to improve the biological description of the system dynamics. Additionally, these modifications allow us to consider different regulation intensities for each network interaction, which is a more realistic representation than assuming the same value for all interactions.

2.6. Parameter estimation

Dynamic model parameters can be categorized into deterministic and noise-related parameters. These parameters must be able to retrieve expression values compatible with the data and known biological behaviors. Since the expression values obtained from scRNA-seq data result from genetic and epigenetic regulation, using these data for parameter estimation requires implicitly considering all contributions, including those from the microenvironment.

It is important to note that the parameters are estimated for the system at values close to equilibrium. There is no guarantee that they will remain consistent in regions far from equilibrium. States far from equilibrium might exhibit genetic and epigenetic differences, deviating from the behavior predicted by the parameter tuning. Nevertheless, the dispersion of experimental data described by the model may sufficiently reflect the statistical dynamics of GBM for characterizing its subtypes.

To estimate the parameters, we used the regression through linear programming and L1-Norm Minimization [36] algorithm with the Simplex method in the Mathematica environment [33]. Consequently, equation (4) was rewritten in the form of the following equation:

$$\mathbf{F} = -k\mathbf{X} + \mathbf{M}_T\mathbf{V}_T\,, \tag{18}$$

where the subscript T indicates the juxtaposition of matrices and vectors. Regulation functions have been incorporated in the M_T matrix, with the parameters in the V_T vector.

We proposed different ways of addressing the challenges associated with estimating the deterministic parameters in our model. These challenges arise from a large number of parameters, the selection of an appropriate estimation method, and the need to ensure biological interpretability. One concern when dealing with a large number of parameters is overfitting; thus, it is important to be aware of this issue and to approach it cautiously.

Additionally, identifiability is a crucial aspect to consider, as it can be difficult to uniquely estimate parameters from the available data in models with many parameters. This challenge may lead to multiple sets of parameter values that produce similar model outputs, making it difficult to draw meaningful conclusions. Furthermore, estimating many parameters often requires significant computational resources and time due to the increasing search space for the parameter values. The complexity of models with many parameters can make them harder to interpret and understand, as the relationships between variables may be obscured.

Keeping these factors in mind, we explored various parameter estimation strategies that strike a balance between model complexity, computational efficiency, and biological interpretability. We considered the following scenarios: (i) with 2 parameters per equation (one for activation and one for inhibition), (ii) with one parameter related to each input node (divided between activation and inhibition), that is, $2n$ parameters per equation, and (iii) with a combination of "i" and "ii", i.e., $n \times n$ parameters (including null values). Equation (19) illustrates the first case:

$$kX_i = a_i \left(\sum_j V_{ij}^a \right) + b_i \left(\sum_j V_{ij}^b \right) , \tag{19}$$

where a possible biological interpretation is an activation and inhibition intensity proportional to the target gene, for example, due to epigenetic regulations. The equation (20) brings the second case,

$$kX_i = \left(\sum_j a_j V_{ij}^a \right) + \left(\sum_j b_j V_{ij}^b \right) , \tag{20}$$

where the consideration of $2n$ parameters would be due to intermediate factors affecting the interactions preceding the binding to the promoter region and the resulting gene transcription. The equation (21) brings the last case,

$$kX_i = a_i \left(\sum_j a_j V_{ij}^a \right) + b_i \left(\sum_j b_j V_{ij}^b \right) , \tag{21}$$

where the idea was (i) to obtain a different parameter for each edge and (ii) to capture as much information as possible. Each activation coefficient would be $a_{ij} = a_i a_j$, with an equivalent procedure for the inhibition coefficients b_{ij}.

The parameter estimation was conducted by simultaneously incorporating all centroids of all basins of attraction, meaning that the parameters were chosen to capture the contributions of all possible equilibrium states of the system. This approach was taken to avoid overfitting individual clusters, which could potentially hinder the representation of other attractors. By doing so and assuming $F \approx 0$, equation (18) could be employed to address a problem of the following type:

$$\mathbf{Mx} \geq \mathbf{b} , \tag{22}$$

$$\mathbf{x} \geq \mathbf{0} , \tag{23}$$

where $\mathbf{M} = \mathbf{M}_T$, $\mathbf{x} = \mathbf{V}_T$ and $\mathbf{b} = k\mathbf{X}$, with a L1-*norm Minimization* algorithm, that minimizes:

$$||\mathbf{Mx} - \mathbf{b}||_1 = \sum_i |\mathbf{Mx} - \mathbf{b}|_i , \tag{24}$$

suitable for the overdetermined system. Then we compute and store the maximum residual $R = \max |M\mathbf{x} - \mathbf{b}|$ and the total $T = \sum_i |M\mathbf{x} - \mathbf{b}|_i$ for each fit as a regression quality measurement.

The solutions were obtained through linear programming code for various combinations of n and S. The values of n ranged from 1 to 4 in increments of 1, while the proportionality constants f_a and f_b ranged from 0.1 to 1.3 in increments of 0.2. These constants were applied separately for activation and inhibition regulation functions. We used each of the three proposed regulation functions for each parameter combination. After estimating the parameters, the absolute residuals list, R, and T were saved. We performed another estimation for each set of parameters obtained using equation (20). This additional estimation allowed us to combine the newly estimated parameters with the initial set, as in equation (21).

We subsequently solved equation (3) numerically using the Euler-Maruyama method with an Itô interpretation, where noise is added before increasing expression levels. A Stratonovich interpretation is also possible, as discussed in [37]. We employed a multiplicative noise function $g(X, t)$, which accounts for the known mean and standard deviation dependence

after a logarithm transformation, as shown in equation (25):

$$g(x; c_0, c_1, p) = x\, c_0 \binom{x + c_1 - 1}{x - 1} p^x (1 - p)^{c_1} \,. \quad (25)$$

where c_0 is the noise amplitude that scales a negative binomial distribution, with c_1 and p empirically determined to fit the data best. p is any positive real number less than or equal to 1, and x represents the expression level as any positive real number.

For initial conditions such as experimental data points or attractor coordinates, the system is expected to evolve towards the centroid of the cluster they belong to or maintain trajectories around its average value. To investigate the system dynamics, we performed stochastic simulations with a time interval of 50 a.u. (time steps of 0.1) to ensure the system reaches equilibrium states. We varied the noise amplitude c_0 (3.5 and 7.0) and explored the impact of different activation and inhibition levels (a, sa, b, and sb) from 0.6 to 1.4 (0.1 by 0.1). These settings were determined through extensive preliminary simulations to understand the system dynamics better, reproduce the observed variability in the experimental data, and maintain computational feasibility.

We then stored the parameter configurations that exhibited an average of 15 or more genes remaining within two standard deviations from the respective attractors' values (as identified by clustering and fitted with the parameters to be stable points) as described in equation (26). In essence, given:

$$z = \frac{|\bar{x} - \mu|}{\sigma} \,, \quad (26)$$

with μ representing the cluster centroid of each gene, $\bar{x}$ denoting the time average within an interval from time 30 to 50 (allowing the system to fluctuate around the centroid), and σ as the standard deviation of the simulated attractor. If $z \leq 2$ for at least 15 genes out of the 40-gene vector, we stored the parameter set. This approach allowed us to test the hypothesis that the clusters found in the data are statistically significant compared to those obtained by the model with the determined set of parameters.

Cases meeting these criteria were individually processed, exploring their distributions, landscapes, and other features. For the landscapes, the initial condition was sampled with a uniform spatial distribution, expecting that the trajectories would be distributed proportionally to the weight/relative percentage of cells in each cluster over a long period. We then computed the landscape using equations (9) to (11). Finally, we assessed the compatibility between distributions by superimposing the histograms of the data with the simulated data and comparing their respective landscapes.

3. Results

The results section is organized as follows: (i) We present the gene regulatory network, which defines the nodes (genes/TFs) and edges (regulation interactions). (ii) We discuss the initial challenges of choosing a clustering method, which will subsequently influence parameter estimation and optimization. One of the outputs of this optimization process is the selection of regulation functions, which we present in the following. (iii) We compare experimental data with various simulated data obtained using different clustering methods. This comparison serves not only to assess the accuracy of the parameter estimates but also to provide a valuable perspective on the simulation outcomes. (iv) We examine a chosen case from the simulations to verify the hypothesis of internal ergodicity. The results may also serve as a preliminary study for analyzing experimental data as it becomes more available. Finally, we aim to comprehensively understand the relationships between data analysis, parameter estimation, regulation functions, and simulations by presenting the results in this order.

3.1. Glioblastoma GRN

The first step was to construct a GRN (figure S1), considering the scRNA-seq data from the four GBM patients as previously described. Figure 1 shows the network considered for the subsequent results. It contains 40 nodes and 242 edges, with 187 activations, 11 self-activations, 41 inhibitions, and 3 self-inhibitions. Supplementary Table rede_BT_ALL_seurat.xlsx presents the list of initial nodes and nodes added by the MetaCore algorithm.

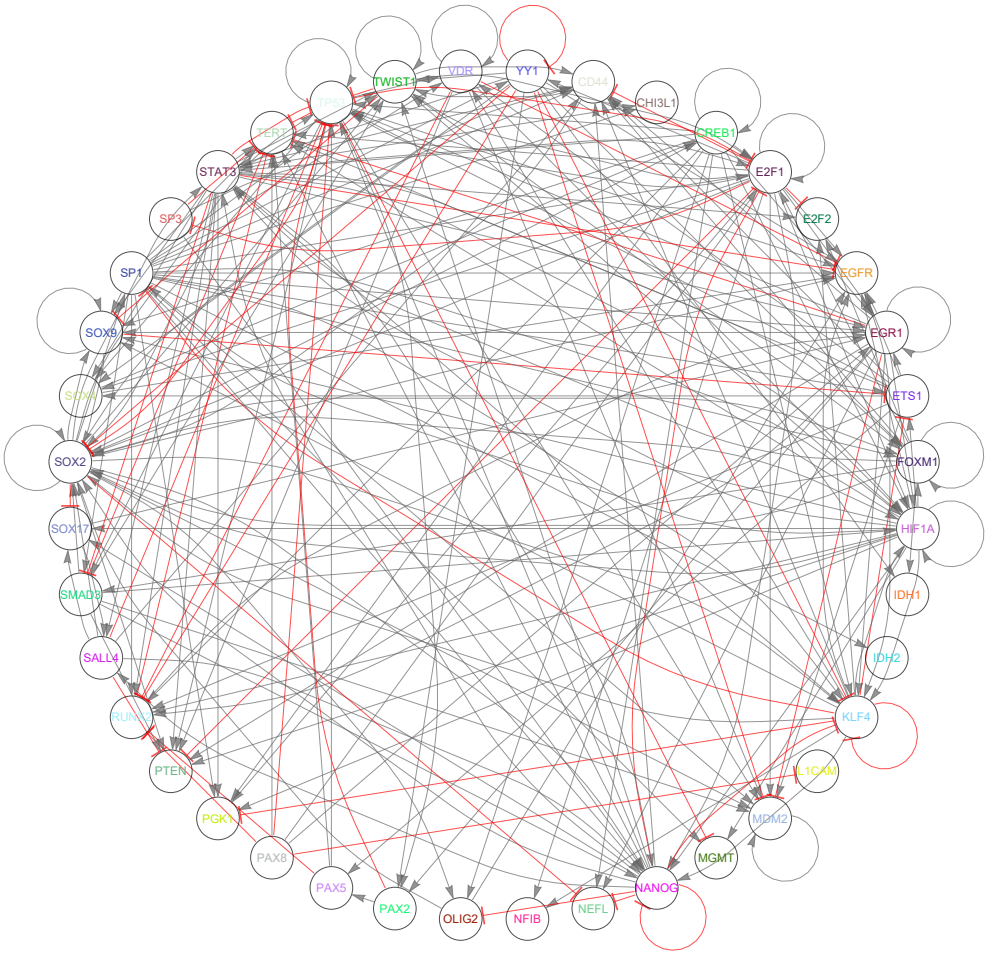


Figure 1: Gene regulatory network for single-cell RNA sequencing of the tumor core of 4 patients with GBM. Black lines with flat arrows represent activations, and red lines with arrowheads represent inhibitions. It contains 40 nodes and 242 edges, with 187 activations, 11 self-activations, 41 inhibitions, and 3 self-inhibitions.

3.2. Clustering methods and parameters estimation

Initial data analysis revealed a group of genes with apparently multimodal distributions. This observation led us to investigate the extent to which this pattern represented the presence of multiple clusters. The clustering procedure involved evaluating data distribution in two reduced dimensions derived from the four marker genes chosen to be characteristic of GBM subtypes (see section 2.4). We first reduced the four-dimensional data to two dimensions using the t-SNE implementation in Wolfram Mathematica, with a perplexity of 60, which we observed to provide a suitable visualization for identifying the number of potential clusters (see figure S2).

To perform the clustering analysis, we utilized two different clustering methods available in Mathematica: k-means and Neighborhood Contraction (NbC). These methods were chosen for their complementary strengths and weaknesses, providing a comprehensive understanding of the underlying data structure. K-means is a widely-used clustering technique that is particularly effective when dealing with clusters of similar size that are uniformly and closely distributed around their central points [34]. However, its performance can suffer when faced with clusters of different sizes or shapes, intersecting clusters, or the presence of outliers [34]. In addition, k-means necessitates specifying the number of clusters in advance and allocating each data point to the centroid closest to it. Conversely, NbC is a hierarchical clustering approach that gradually moves data points toward higher-density regions, facilitating the defini-

tion of clusters [35]. Furthermore, unlike k-means, NbC does not require a predefined number of clusters and is better equipped to handle clusters with varying sizes and shapes. Nevertheless, NbC may face challenges when dealing with clusters of different densities or when clusters are close to each other [35]. Therefore, we configured the built-in Mathematica functions for both clustering methods with PerformanceGoal set to "Quality", CriterionFunction set to "StandardDeviation", and DistanceFunction set to EuclideanDistance. By comparing the results of these two clustering methods, we aimed to understand the underlying data structure better and identify the optimal number of clusters for the given gene expression data.

Parameter estimation was conducted using the centroid coordinates of all genes within each cluster in an oversized fit, considering $k = 1$. An interval ranging from 0.01 to 10 was used for adjusting the parameters. These values were initially established through manual checks, ensuring that changes in the interval would not lead to substantial variations in the quality estimator values[3]. The first estimate was performed using equation (19), obtaining 784 sets of parameters. The distribution of values for each parameter and both clustering methods can be seen in figure S3. As shown in the figure, some parameters have distributions that vary within the interval limits, such as a[*EGR1*] and a[*SOX4*]. In contrast, others display values constrained to smaller intervals, such as a[*E2F1*] and a[*TP53*]. This pattern is evident in both clustering methods, signifying a strong relationship with the network structure. The larger variations in specific parameters can demonstrate how these parameters are susceptible to changes within the established network. Moreover, it may reveal how alterations in regulating some genes can impact both clustering methods. In the context of cancer, this observation may suggest that some changes in gene regulation do not necessitate significant compensation for the system to remain close to its previous behavior. In contrast, others may require considerable compensation, without which it could lead to undesirable changes. The second parameter estimation using equation (20) was performed on top of the first to adjust the values for each interaction. This step's distribution can be seen in figure S4. The distribution of most parameters is around the unitary value, suggesting a possible correction of the previous estimate. This observation implies that the second estimation step fine-tunes the parameters to enhance the model's accuracy.

Figures S5 and S6 show the residual for each gene for the two clustering methods superimposed, one for the first parameters estimation and the other for the second parameter estimation. Both figures demonstrate that the two consecutive parameter estimations did not significantly affect the residuals of the first parameter estimation and indicate that it did not result in overfitting. Additionally, the model showed good compatibility for some genes based on the centroids of the experimental data clusters. For instance, genes such as *HIF1A* and *SOX2* may have presented low residuals due to their close centroid values proximity. In contrast, genes like *CD44* and *EGFR* exhibited high residuals and variations due to their distant centroid values. While these residuals may indicate a possible need for adjustments in network interactions or the model itself, they might also reflect the influence of the clustering method. Therefore, analyzing the relationship between residuals, network structure, and the clustering method is crucial for drawing more accurate conclusions. Furthermore, this analysis could provide insights into the model's limitations and help identify potential improvement areas.

The final step of the parameter estimation process aimed to assess the compatibility between the stability of cluster centroids derived from experimental data and those obtained from parameter adjustments. To accomplish this, we selected the set of parameters with the smallest T and R values (as displayed in Table 1). Figure S7 (a) and (b) illustrate the activation (left) and inhibition (right) matrices, with the color gradient signifying the logarithm of the parameter values for both clustering methods. The 'o' and 'x' symbols denote the regulations and self-regulations present in the network, respectively. Parameter values are acquired by multiplying parameters from the first and second fits (columns and rows weighting). The simulation proceeds by multiplying each ma-

[3]Differences in the interval resulting in approximately 1% changes in the indicator values were disregarded.

Table 1: Table with optimized parameters. The first and second lines correspond to parameters after k-means and Neighborhood Contraction clustering, respectively. The left side parameters correspond to the optimization using deterministic dynamics, and the indices 0 and 1 in R and T inform if it is for the first or second estimation. The right side parameters correspond to the optimization considering stochastic dynamics, with G/C the mean gene number per cluster when considering the parameters optimization.

n	$h(x)$	f_a	f_b	R_0	R_1	T_0	T_1	c_0	a	sa	b	sb	G/C
1	2	0.1	1.3	1.62	1.15	24.85	20.89	3.5	1.4	0.7	0.9	1.3	20.40
1	2	0.1	1.1	1.72	1.69	35.43	28.48	5.6	1.4	0.6	1.3	1.2	16.71

trix element by either 0 or 1 (1 if the element corresponds to an 'x' or 'o' position, and 0 otherwise) to account for only the combinations present in the network. Figure S8 displays the ceiling function (ceil()) of all parameter values for k-means clustering and NbC clustering, respectively. The ceiling function, ceil(), rounds up a given number to the nearest integer. In this case, it helps visualize the order of magnitude of the parameter values. The left images correspond to activation parameters, while the right images depict inhibition parameters.

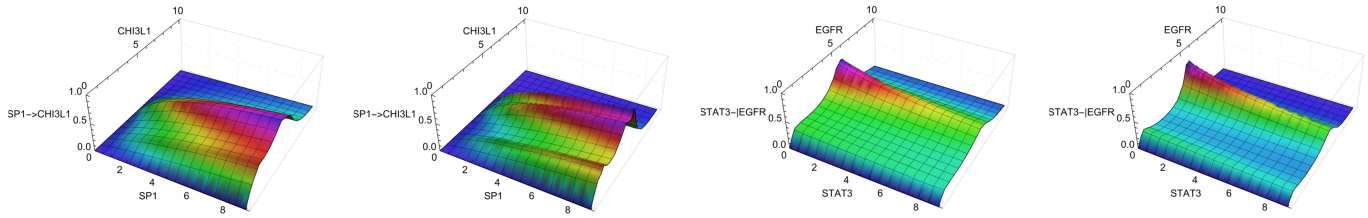
The global optimization process aimed to maximize the number of simulated gene distributions compatible with the experimental data by optimizing the global strength of activation, self-activation, inhibition, and self-inhibition ('x' and 'o' positions) using multiplicative factors. The factors were chosen to maximize the average number of genes that stay within two sigmas of their cluster centroids. In the first and second parameter estimation steps, 484 sets were generated for both k-means and Neighborhood Contraction clustering methods. The best sets, number 435 for k-means clustering and number 428 for Neighborhood Contraction clustering, were then used in the optimization process. Figure S7 (c) and (d) illustrate the distribution of genes compatible with each attractor during the stochastic parameter analysis for both clustering methods.

For the k-means clustering, the optimization process yielded a list of 191 values, with the best set being number 150 (out of 191). The experimental attractors were defined as A, B, C, D, and E. Varying the regulation weights in each stochastic simulation, attractors A and C had an average of about 15 genes within two sigmas concerning the observed data, attractor B had an average greater than 20, attractor D had an average close to 20, and attractor E had an average of only 5 genes compatible with the data. Moreover, the following compatibility with data values was observed: attractor A (15), B (24), C (30), D (29), and E (4) when using factors 1.4, 0.7, 0.9, and 1.3, respectively. In the case of Neighborhood Contraction clustering, the optimization process generated a list of 13 values. The best set was number 11 (out of 13), and the clustering ranged from A to G. The compatibility with data values for attractors A to G were 19, 20, 11, 20, 17, 17, and 13 when using factors 1.4, 0.6, 1.3, and 1.2, respectively. Table 1 presents the values of two parameter sets: one for k-means clustering (5 clusters) and another for Neighborhood Contraction (7 clusters).

When comparing the k-means and Neighborhood Contraction clustering methods, we observed differences in the number of genes compatible with each attractor and the quality of the estimated parameters. For k-means clustering, the total number of parameters that leads to genes compatible with attractors A, B, C, D, and E was 191, while 13 with attractors A to G for Neighborhood Contraction clustering. These results suggest that as the number of clusters increases, the set of parameters that matches the data for the established conditions decreases.

3.3. Regulation functions

One of the results of the first and second estimations was the selection of the regulation function. The modified regulation function depends on the transcription factor and the regulated genes. The total number of regulatory functions equals the number of interactions in the network (242 interactions). After the parameters optimization, the best compatibility between experimental and simulated data was obtained with the regulation function of the equation (15). For example, figure S9 shows the activation and inhibition regulation functions for some



(a) (b) (c) (d)

Figure 2: New regulation function V with $n = 1$ and $h_2(x)$ for different combinations of genes and/or transcription factors. The horizontal axis represents the transcription factor and gene quantification using the normalized amount of single-cell RNA sequencing of experimental data. The vertical axis represents the quantification of the interaction regulations. (a) and (b) Activation values using the k-mean and NbC clusters, respectively ($f_a = 0.1$); (c) and (d) Inhibitory interactions using the k-mean and NbC clusters, respectively ($f_b = 1.3$ and $f_b = 1.1$).

gene/transcription factor combinations. Each case presents a different relationship between the variables representing the amount of transcription factor mRNA and the activity of the gene promoter. Figures S9 (a) and (b) correspond to the regulation functions of the first clustering choice, while figures S9 (c) and (d) to the second choice. Figure 2 displays some of these cases, such as the activation of *CHI3L1* by *SP1* and the inhibition of *EGFR* by *STAT3*. All surfaces were obtained with $n = 1$ and $h_2(x)$, with (a) and (b) using $f_a = 0.1$ and $f_b = 1.3$ and (c) and (d) using $f_a = 0.1$ and $f_b = 1.1$.

The dependence of the regulation function on the transcription factor can be seen along its axis, as affected by the contribution of the parameter f (f_a for activation and f_b for inhibition) and the values ($\max_\alpha [X_j^\alpha]+1$). The dependence on the target gene is represented on the axis transverse to the transcription factor axis, where the maximum values for constant transcription factor concentration correspond to the centroid values observed in the experimental data. It is possible to observe how the peaks occur around the average values of *CD44*, *CHI3L1*, *EGFR* and *IDH1* for activation functions and *CD44*, *EGFR*, *MDM2*, and *PGK1* for inhibition.

When evaluated at constant target gene concentrations, the regulation functions collapse the surfaces into one-dimensional curves resembling standard Hill functions with $n = 1$. For high target values, the regulation functions tend to zero. This observation agrees with the experimental data used in this report, which also demonstrates a similar trend. From a biological perspective, this behavior aims to reproduce possible mechanisms not explicitly present in the network, such as the presence or absence of microenvironment signals[4] or missing interactions in the network. It is important to remember that lower inactivation values represent higher inhibitions. Therefore, a lower value of V^b means greater inhibition, leading to lower basal activation. Furthermore, the modified regulation function does not exhibit peaks where the target concentrations are zero. This modification was intended to eliminate the occurrence of activation peaks in the regulation functions, as this could influence the acquisition of the null values observed in the data. For points beyond these boundary cases, the system's evolution will be guided by the interplay of network interactions, parameter values, and noise.

3.4. Comparing experimental and simulated data

3.4.1. Noise and distributions compatibility

Data transformation often results in a dependence of the standard deviation on the mean [38]. Part of this dependency can be reduced depending on the normalization process. However, to achieve a better fit between the simulation outcomes and the sequencing data, the function for the multiplicative noise was defined as equation (25), with $p = 0.23$ and $c_1 = 8$, which were found empirically to fit

[4]Ex: negative *Feedback loop*

the data best. Figure S10 presents the graphic for these parameter values. The optimized set of parameters was evaluated by comparing the histograms of expression levels for data at different time points to assess whether the system had reached a steady state. Figure S11 compares experimental and simulated data. Figure S11 (a) shows the simulation outcome for time 50 (500 steps), (b) time 25 (250 steps), (c) time 5 (50 steps), and (d) the initial conditions. The means of the simulated distributions are mostly within two standard deviations of the means of the observed multimodal distributions in the experimental data, demonstrating good compatibility for most genes. However, mainly for higher expression values, the simulated outcomes exhibit smaller standard deviations than those observed experimentally.

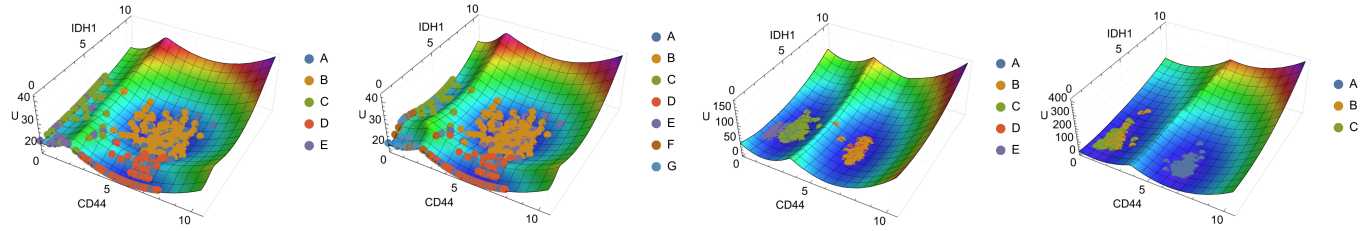
3.4.2. Clusters compatibility

To assess if the patterns in simulated data resemble experimental data, we clustered the simulated data using various algorithms, including k-means, Neighborhood Contraction (NbC), and Gaussian Mixture. Figure S12 shows clustering outcomes for each method applied to simulated data generated using parameters selected after k-means and NbC clustering. We also clustered experimental data using these methods for comparison. We obtained the clusters in figure S12 within two reduced dimensions instead of the four original marker gene dimensions. These plots allowed visual assessment of clustering quality, which would be challenging in a four-dimensional space. The analysis shows that NbC performed better in the reduced dimensions for simulations generated with parameters after both k-means and NbC clustering of experimental data.

To delve deeper into this comparison, we examined the centroids of simulated data obtained after each clustering method for all cases alongside the data distributions. Figure S13 and S14 display these comparisons for the k-means and NbC clustering results, respectively. The centroids for simulated data were somewhat similar, independently of the clustering method used, particularly when comparing NbC and Gaussian Mixture methods. Interestingly, it seems that the centroids derived from the simulated data capture the centers of the multi-modal distributions of the experimental data more accurately than by clustering the original data. This result could help identify or confirm biologically significant information within the data, especially if these multi-modal distributions are relevant to the underlying biological processes.

To facilitate the clustering comparison, we rearranged each cluster of the simulated data to match its closest experimental data cluster, using the experimental clusters as a reference: (i) We computed the centroids of all clustered data, representing the average points or centers of each cluster. (ii) We calculated the Euclidean distances between the centroids of each element of the simulated set of clusters and each cluster in the reference set, creating a distance matrix. The algorithm identifies the minimum distances between the centroids of the simulated set of clusters and the reference clusters sequentially. (iii) For each element in the second set, the code finds the reference cluster with the smallest distance, maps them together, and removes the elements corresponding to the selected reference clusters from the distance matrix. This ensures that each element in the simulated set is associated with a unique reference cluster. (iv) The input clusters are then reordered based on these minimum distances, minimizing the overall difference between the two sets of clusters. (v) The reordered set of clusters is returned, enabling comparison and analysis against the reference clusters. This process aids in identifying similarities and differences between the two sets and helps to understand the underlying structure of the data.

To gain more insight into the data, we compared the number of points within each cluster by considering the pie charts for both clustering of the experimental data and all cases of simulated data. Figures S15 and S16 display these pie charts and reveal that clusters A and B exhibit nearly the same proportions across all cases, except for the Gaussian Mixture for the NbC simulated data. The other clusters vary according to each method. This variation might be due to the presence of metastable clusters, which are not very stable or well-separated from one another. These metastable clusters can significantly impact clustering methods because they can cause misclassification or make identifying clear boundaries between clusters difficult. Additionally, different clustering methods may handle metastable clus-



(a) (b) (c) (d)

Figure 3: (Epi)genetic landscapes for experimental (a and b) and simulated (c and d) data, with experimental and simulated points overlaid for compatibility visualization. (a) Landscape for experimental data with k-means clusters; (b) Landscape for experimental data with NbC clusters; (c) Landscape for simulated data after parameter estimation with the k-means centroids and clustering with Gaussian Mixture; (d) Landscape for simulated data after parameter estimation with the NbC centroids and clustering with Gaussian Mixture. The horizontal axes show the expression values of each marker gene, while the vertical axis represents the values of the landscape.

ters differently, which can lead to discrepancies in their qualitative and quantitative identification. As a result, the choice of clustering method and the presence of metastable clusters can influence the observed results in cluster analysis.

To further investigate these findings, we examined the distribution within each cluster. Figures S17 and S18 revealed the standard deviation for each cluster in every case of simulated data was lower compared to the experimental data. This may result from the estimated parameters providing narrow local stability, with potential noise-induced jumps directing the trajectories to another small phase space area. We computed boxplots and histograms to compare the distribution of gene expression levels within each cluster. Figures S19 and S20 show the boxplots for both cases of experimental data alongside all instances of the respective simulated data. They demonstrate that the best matches for gene expression distributions within clusters were found in clusters B, C, and E for the after k-means clustering and A, B, and G for the after NbC clustering, respectively. Figures S21 and S22 show the distributions per data cluster. The *NEFL* values were not considered due to their fixed distribution close to zero. These results reinforce the previous observations regarding centroid compatibility and the difference in standard deviation distributions. Furthermore, we also observed the instability of the results obtained after parameters estimation using NbC, which exhibited significantly different simulated clustering outputs.

Upon analyzing graphics that could be more representative of possible trajectories, we examined the scatter plots of the combinations of the marker genes, excluding the *NEFL* gene. As illustrated in Figures S23 and S24, it appears to be some points in transition between the clusters, which could support the hypothesis of narrow stability. However, this would also disrupt the assumption of internal ergodicity if it occurs frequently. Furthermore, since these points correspond to different cells at the same time step, further investigation into single-cell trajectories over time would be necessary. Another observation is that, despite the differences in distribution dispersion, a qualitative analysis (low and high) of the clusters with better compatibility would yield similar results to those found by both clustering methods. This suggests that the clusters may still be comparable when focusing on their overall trends rather than the specific dispersion of data points.

3.4.3. Experimental and simulated landscapes

Considering the results obtained, we now examine the corresponding (epi)genetic landscapes. Assuming that each cluster's average and deviation characterize each cell state's distribution, visualizing a potential surface derived directly from the experimental data can help evaluate the clustering quality and serve as a reference for a qualitative comparison with the simulation results. Figures S25 and S26 present the landscapes for both experimental data clustering methods and all simulated data clustering methods.

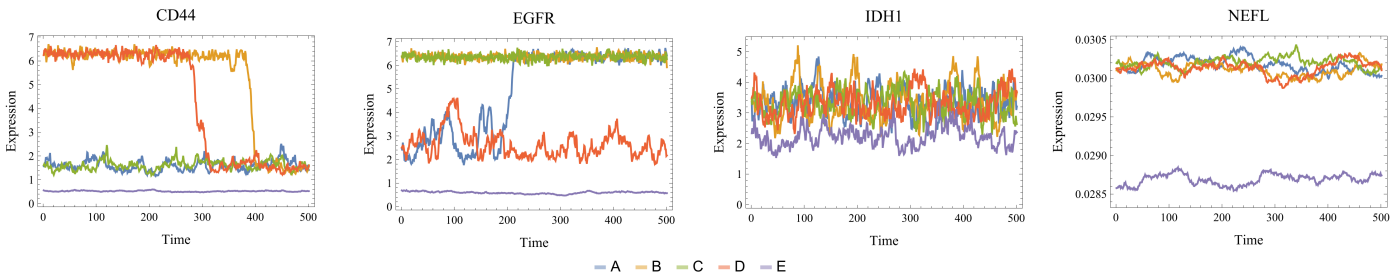


Figure 4: Trajectory plots for each basin (A, B, C, D, and E), showcasing the dynamics of the four marker genes (CD44, EGFR, IDH1, NEFL). The trajectories illustrate the time evolution from initial conditions as the centroid of the clusters found by Gaussian mixture after the parameters found using the k-means centroids. The horizontal axis represents time steps, and the vertical axis shows the expression level of each gene.

A constant of 0.01 was added due to the presence of null values (or close to zero) for the standard deviation of some genes. This value was sufficient to avoid numerical issues and significantly smaller than non-null values.

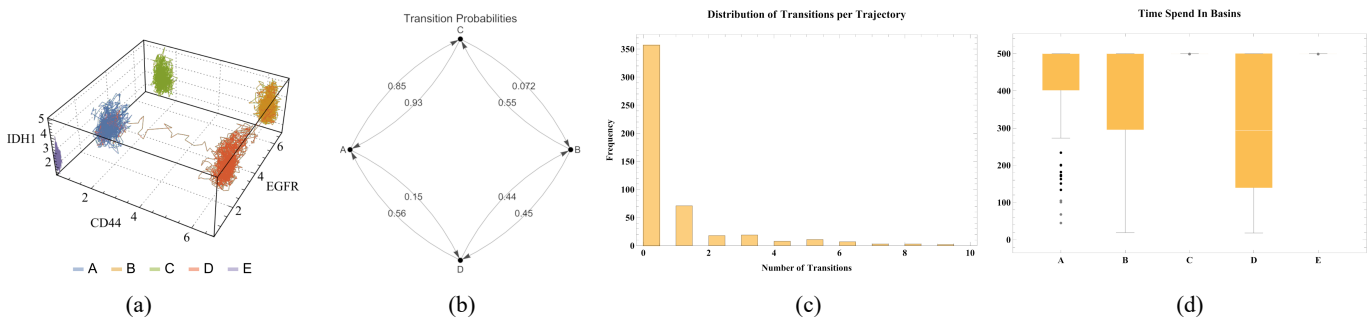
Qualitative changes can be observed due to the greater number of attractors in the NbC method compared to the k-means one, particularly in the dimensions *CD44×IDH1* (figure 3 (a) and (b)), *CD44×NEFL*, and *EGFR×IDH1*, with the presence of new basins. When comparing these experimental landscapes to the simulated ones (figure 3 (c) and (d)), as expected from previous findings, there are significant differences. The landscape is a precise structure that requires a high degree of correspondence between means, standard deviations, weights, and attractors. It is important to note that the experimental landscape does not consider dynamics and takes into account data points that may be in transient or metastable states. Regardless, the previous results are relevant and may indicate avenues for further improvements and studies on the system dynamics.

3.5. Ergodicity analysis for chosen simulated case

The results and analysis presented above lay the groundwork for further investigations into various topics related to the system's dynamics and properties. Here, we will explore results concerning the hypothesis of internal ergodicity. First, let's examine a collection of trajectories originating from the centroids of the clustered simulated data. Then, we selected one case to investigate the dynamic properties of the trajectories. This case corresponds to the Gaussian mixture after the k-means parameters optimization. Figure S27 displays the trajectories starting from each cluster centroid for each gene, using the parameters obtained after k-means clustering and the Gaussian mixture cluster classification.

For most clusters and genes, the trajectories oscillate around their respective centroids. However, for these single trajectories of each cluster, we can observe that specific genes, such as *CD44* and *EGFR*, exhibit transitions (figure 4). Under the assumption of internal ergodicity, such transitions may not be frequent and may occur on a longer time scale than the component dynamics. This suggests that a more detailed investigation is needed to better understand the nature and frequency of these transitions in the context of the system's ergodic properties.

To further investigate the transitions between components, we focused on two main features: the identification and quantification of the transitions among the defined basins of attraction and the time spent in each basin before the occurrence of a transition. This analysis was performed by sampling the time steps of each trajectory starting from the centroid of each cluster, measuring each point's Euclidean distances to all basins, and assigning it to the closest one. We also considered establishing a threshold to determine whether a point belongs to a basin. Still, it would lead to the challenges associated with complex dynamical systems and basins of attraction. In addition, determining the basins' exact boundaries and neighborhood sizes can be difficult, as the basins' shapes may be irregular and not easily described by simple measures like distances or standard deviations. Ul-



(a) (b) (c) (d)

Figure 5: (a) Three-dimensional visualization of full trajectories in three of the markers space. All axes display the expression values of each marker gene. Each line represents an entire time of the three considered trajectories. Each color/letter indicates its respective basin. (b) Transition graph illustrating the connections between different basins. Nodes represent basins, and edge weights represent the probabilities of transitions between basins. (c) Frequency of transitions per trajectory. Histogram showing the frequency of transitions between basins in each trajectory, highlighting that most trajectories do not present any transition, and those that do tend to have a small number of transitions. The vertical axis shows the frequency of each number of transitions per trajectory, while the horizontal axis the number of transitions per trajectory; (d) Analysis of time spent in basins before a transition. Box plots reflect the distribution of time spent in each basin across all trajectories before they present a transition. The vertical axis represents the time spent in the basin, while the horizontal axis the correspondent basins.

timately, we decided to focus on the closest basin assignment, acknowledging the limitations and complexities involved in this approach.

We sampled 100 trajectories for each cluster and tracked their paths. Figure S28 (a) and (b) show the quantification of the transitions between the basins. Absent basins indicate that there were no transitions to or from them. Figure S28 (a) displays the number of transitions between each cluster, which may be affected by multiple transitions within a single trajectory. The most frequent transitions were observed between clusters A and C, followed by D to A. Both of these transitions are visible in figure S27. Figure S28 (b) quantifies the probability of a jump from one cluster to another. For example, transitions from cluster A have a probability of 0.85 to go to C and the remaining probability of going to D, and so on.

Studying the time spent by a trajectory within each cluster can help us to understand the system's stability and the relative significance of each identified state. Long residence times within a cluster suggest a stable state, while shorter times could indicate a transient or metastable state. Figure S29 (a) displays the time spent in each cluster before it jumps to another. The absence of dispersion in clusters C and E suggests that the trajectories starting in these clusters remained there. This observation does not contradict the transition graph; it simply means that when jumping to cluster C and returning to their original cluster, we might consider that the transition was not fully completed, as the system may not have reached the narrow stability of cluster C. This could be due to significant fluctuations in other dimensions. In any case, we can observe that the time spent before transitioning is relatively high for all basins compared to the total observation time.

Additionally, figure S29 (b) shows the histogram for the number of transitions within each trajectory. In other words, considering the 500 trajectories (100 for each cluster), most did not exhibit any transition. The most frequent number of transitions was 1, followed by a decrease in frequency as the number of transitions increased. These results agree with the internal ergodicity hypothesis of low-frequency transitions. However, quantifying a time scale within each basin and for basin transitions is still required to verify the hypothesis and further understand the system's behavior.

Before assessing the time scale, let's consider figures S30 to S33, illustrating the dispersion of trajectory time points for each basin. We can see the well-defined localization and even the time points possibly representing transitions between clusters. The points are all time points for three trajectories of each cluster. Figure 5 summarize the results, with (a) one of the 3D trajectories, (b) the transition matrix, (c)

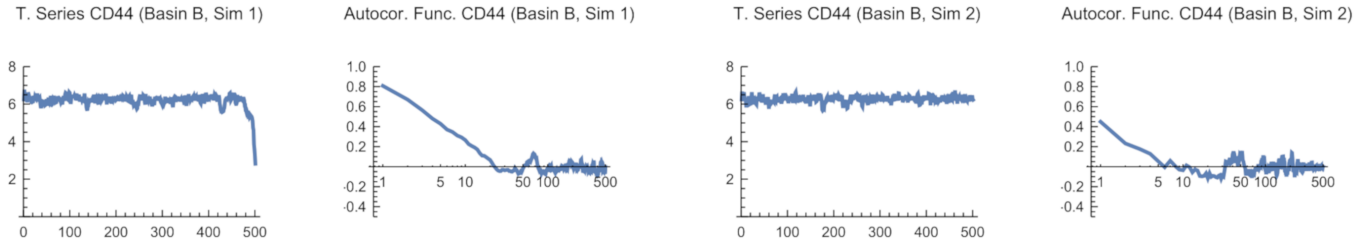


Figure 6: Autocorrelation analysis of time series data for CD44, basin B, and two simulations. Each pair of plots within includes a time series plot (left) with the horizontal axis representing time and the vertical axis representing expression values, and an autocorrelation plot (right) with the horizontal axis representing time lags and the vertical axis representing autocorrelation values.

the frequency of transitions per trajectory, and (d) the time spent before a transition. The 3D plots help visualize the complexity of each trajectory, and it is important to remember that they are a simplification of the 40-dimensional space. This analysis provides a qualitative intuition, given that the biological system is much more complex.

To better understand the system's dynamics and gain insights into the time scales involved, studying autocorrelations within the trajectories is essential. Autocorrelation analysis can reveal the system's temporal structure. By examining these structures, we can establish a time scale that characterizes each basin's internal dynamics and the time scales associated with transitions between basins. Furthermore, this approach allows us to connect the qualitative intuition provided by the trajectory plots with quantitative measures that can more accurately describe the system's behavior and its relation to internal ergodicity.

Figure S34 shows the time series side by side with its corresponding autocorrelation functions for all genes and basins considering two sampled trajectories. The autocorrelation functions represent the autocorrelation for each time lag up to a specific maximum lag. The autocorrelation function can be a powerful tool for understanding the characteristic timescales of the system's dynamics. By analyzing the autocorrelation function, we can identify the timescales at which the system exhibits significant correlations, indicating the persistence of certain behaviors or states. In addition, we can see that each basin and gene may present different time scales, with consistency for both simulations. However, besides this visual inspection, we need to define another way to quantify this due to the complexity of the data. Therefore, we computed the time scale as the minimum time step to reach an autocorrelation value below e^{-1} for each basin, variable, and simulation.

Figures S35, S36, and S37 aim to understand the distribution of these time scales within each internal dynamic and search for possible transition behavior. Figure S35 shows that even with time scales varying between genes, they tend to present similar values for each gene. However, compared to other clusters' time scale patterns, some discrepant cases are observed for *CD44* and *EGFR* genes and for cluster E. We can see in figure S34 (a) that these cases presented transitions, leading to increased time scales. Figure 6 display an example of increased time scale for the *CD44* by comparing the autocorrelation with and without the presence of a transition. These figures highlight the presence of different characteristic time scales within the basins and inter-basins which is in accordance with the internal ergodicity hypothesis.

Furthermore, figures S36 and S37 emphasize the values in each basin. Figure S36 shows the significant variations of time scales for the genes within each basin. Still, given the previous discussion, we note that some genes present very narrow time scale distributions while others have wide error bars. This suggests that the latter may be related to transitions and could be potential variables for further analysis.

Finally, figure S37 shows that the average over all genes and trajectories yields very close results, with only cluster E displaying a different pattern. This may suggest that the time scales are interconnected, and the system may somehow compensate for them.

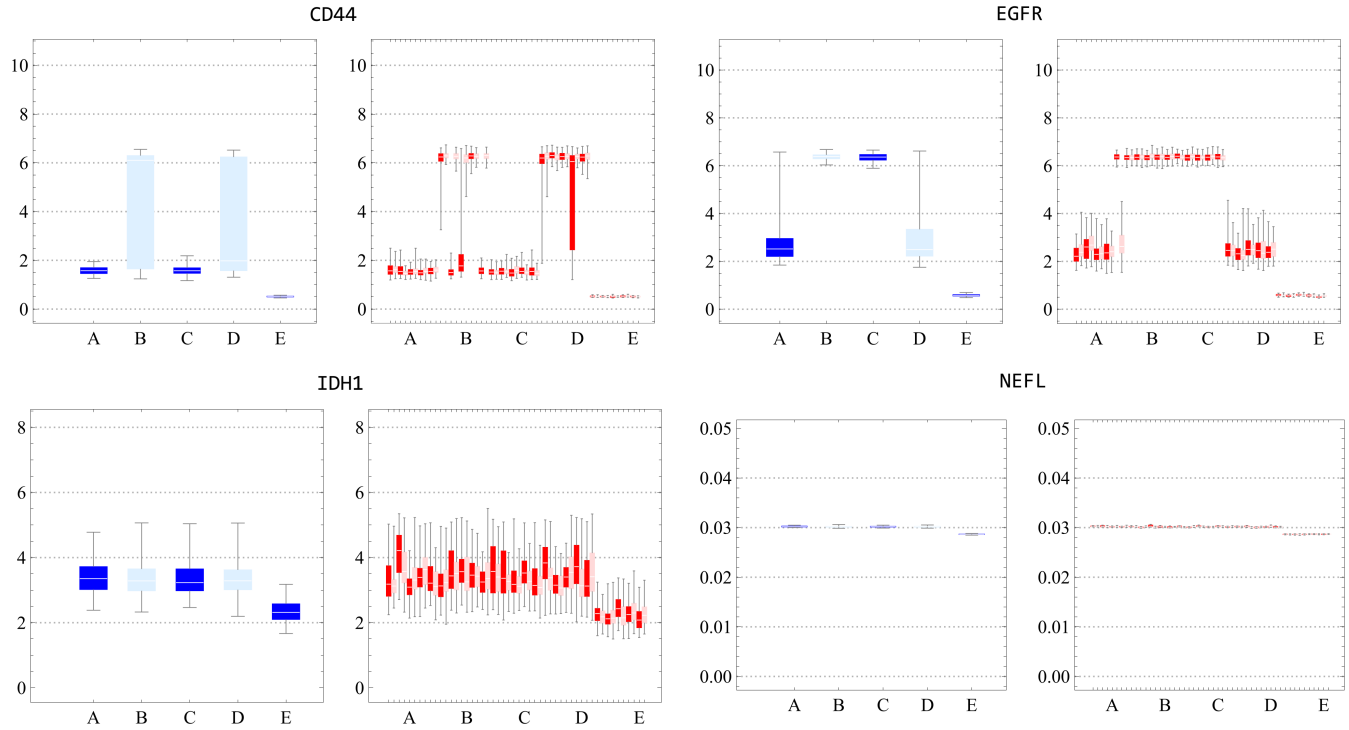


Figure 7: Boxplots for comparison of sample and time averages for all clusters and the four marker genes. The horizontal axis represents the cluster labels, and the vertical axis shows the expression values. The figure displays two sets of plots: the left set shows the sample average of 100 samples at the final time interval, while the right set represents the time average considering 10 trajectories from time 30 to 50 (steps 300 to 500).

The underlying biological processes at play within the system may lead to interconnectedness that allows for compensation across different time scales. This observation opens up opportunities for further investigation into the mechanisms behind this behavior and how it might relate to the system's overall function and stability.

After conducting all these analyses, we can finally assess the compatibility between time and sample averages. Figure S38 demonstrates this compatibility for nearly all clusters and genes. The left panels of figure S38 represent the sample average of 100 samples at the final time, while the right panels display the time average, considering 10 trajectories from time 30 to 50 (steps 300 to 500). Once again, the observed discrepancies, such as in *CD44* basin B and D, might be connected to transitions between clusters (figure 7). We computed these averages considering their departure states, and they may have jumped to another basin during the process.

In conclusion, these results support our initial assumption of internal ergodicity, even if only for simulations. The correspondence with the data suggests that ergodicity might also be present in the biological system. This finding encourages further investigation of these properties within biological samples. Our analyses can be compared and utilized for future applications, ultimately contributing to a better understanding of the system and developing more effective treatments.

4. Discussion

Despite the limitations of obtaining information on biological mechanisms and temporal trajectories, the ergodicity assumption helped to connect the available data with a theoretical background. Therefore, it was fundamental for developing the proposed methodology to provide a data statistical dynamics characterization. In this report, we proposed that, by analogy to the observations made by Mauro and Smedskjaer [17], cancer could initially take the system from a non-ergodic state to an ergodic one. Hav-

ing more freedom due to the mutations and not respecting the same biological constraints of healthy cells, the time scale to visit different phase space points tends to be shorter in cancer. This may be likely related to aggressiveness.

The methodology consisted of studying the deterministic dynamics with regulation functions as modified Hill functions, studying the stability of the attractors obtained by different clustering methods in the presence of noise, and comparing simulated results with experimental data. By integrating the average values of the clusters in the regulations, unknown factors that lead to the observed cellular states were indirectly implemented, whether related to (epi)genetic characteristics in the patient samples, different mutations, or even influences of the tumor microenvironment. These factors helped to constrain the dynamics according to experimental data, reducing the sensibility of the network structure.

Before the parameter estimation, clustering stable states related to subtypes of GBM was essential for further investigation. But, even for data clusters corresponding to stable states, the stability may depend on various factors, such as interactions between the system and its environment. In this way, supported by the constraints of deterministic dynamics, we studied the parameters combination whose noise could lead to a statistical behavior that matches data. Furthermore, we considered a noise that was neither too weak to not significantly impact the system dynamics nor too strong such that it would dominate the system behavior. Finally, by characterizing the dynamics this way, we considered that the individual trajectories did not correspond to a precise biological process, but instead, only a group of them could represent some statistical tendency.

The parameter estimation was carried out by simultaneously considering data from different patients to contribute to a more robust characterization of the GBM dynamics. Also, we considered all clusters together to avoid biases toward specific ones. This analysis can be seen as further analysis of the usual clustering for cell types and subtypes identification once the standard studies usually left out a dynamic characterization. Additionally, our analysis strengthened the evidence that few specific markers may contain enough information to distinguish cell types by showing that they can reflect different stable states.

The parameter modeling provided a powerful way to strengthen the biological description once many studies usually consider arbitrary parameters with fixed values. Using the residuals as the basis for the estimation quality indicators, we showed that they could also measure network and model quality, which needs further investigation. However, regardless of the set of parameters found, non-zero residuals were already expected since the characterization of an exact equilibrium would undoubtedly require several extremely complex variables and not even be realistic. It was observed that, even with non-negligible values of residuals, the results may still show compatibility with the experimental data.

By comparing the experimental and simulated data, we could evaluate the performance of our model and assess its predictive power. While there are still some differences between the two data sets, overall, the model captured the essential statistical dynamics features. For example, we retrieved most genes' experimental attractors and compatible centroid values. However, we still need to be more accurate in distribution shape for most of them. But it is essential to highlight that these discrepancies could be attributed to the various sources of noise and measurement errors that are not connected to underlying biological processes. Thus, the comparison with experimental data is more of a tool for searching the possible applications for the method than a binary verification of its validity.

The expression values are directly connected to the interactions defined by the GRN and the parameter values. These can change over time in the biological system due to mutations, epigenetic regulations, and other factors. Therefore, it is reasonable to approach them as constants for their characterization and quantification for a short time scale. However, their changing may also open possibilities to study the tumor progression by assuming that the landscape is still in a transient state.

In the final stage of the investigation, we dive deeper into the dynamics of each time trajectory to gain further insights into the system's behavior. This analysis involves examining the stability of each cluster, comparing time and sample averages, and

evaluating the time spent within each cluster, which can serve as a measure of stability. Furthermore, we assessed the frequency of transitions and timescales within each attractor to better understand the system's dynamics. These steps allowed us to quantify the internal ergodicity hypothesis for the simulated data to some extent. However, it would be valuable to study these aspects in more detail and to identify the factors influencing the system's properties. By doing so, we can determine what might cause the system to transit between ergodic and non-ergodic states.

Suppose cancer causes a system to shift from non-ergodic to internal ergodic and eventually to total ergodic states. In that case, studying these transitions could provide critical insights into cancer progression and potential interventions. In addition, these investigations may pave the way for a deeper understanding of the system's dynamics and the development of treatment options. Ultimately, this research could contribute significantly to advancing cancer therapy and patient care.

5. Conclusion

Our study has made substantial progress in understanding the dynamics of GBM subtypes through a novel modeling strategy that explores the association between ergodicity and GBM aggressiveness. By analyzing the stability of centroids obtained from multiple methods, we have gained valuable insights into the assumption of ergodicity. However, this concept has yet to be studied in this context. Additionally, introducing new dynamic regulation functions has expanded our understanding of potential approaches to model the network dynamics, mainly related to GBM subtype development.

Our findings demonstrate a high degree of compatibility between centroids of experimental and simulated data, with which we thoroughly investigated the ergodicity of simulations. In these investigations, we detected transitions between basins and, despite this, observed the compatibility between sample and time averages. While our study has limitations, these valuable insights contribute to the current understanding of GBM subtypes and their development. The transitions between the basins reveal a possible interplay between subtypes at the core of cancer recurrence. We have shown that these transitions may be rare and do not significantly affect the internal ergodicity assumption. However, transitioning from a non-ergodic to an internally ergodic and even an ergodic system may be essential to recurrences and subtype developments.

Future research can further explore the dynamics within the (epi)genetic landscape of GBM subtypes to better understand the pathways that lead the system before the differentiation of each subtype. Such studies could provide even more powerful insights into subtype development and contribute to developing targeted therapies for GBM patients. Additionally, this approach could be used to assess and monitor the evolution of malignant states, serving as an instrument for diagnosis and treatment.

Competing interests

No competing interest was declared.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) through the Social Demand Program (Programa de Demanda Social, DS) under File Number 88887.597339/2021-00 - Finance Code 001. We want to express our gratitude for their invaluable support.

References

[1] Who, Cancer overview, 2022. URL: https://www.who.int/health-topics/cancer#tab=tab_1, [Accessed on 3th August 2022].

[2] A. O. Sasmita, Y. P. Wong, A. P. K. Ling, Biomarkers and therapeutic advances in glioblastoma multiforme, Asia-Pacific Journal of Clinical Oncology 14 (2017) 40–51. doi:10.1111/ajco.12756.

[3] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes, Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by

abnormalities in PDGFRA, IDH1, EGFR, and NF1, Cancer Cell 17 (2010) 98–110. doi:10.1016/j.ccr.2009.12.020.

[4] H. S. Phillips, S. Kharbanda, R. Chen, W. F. Forrest, R. H. Soriano, T. D. Wu, A. Misra, J. M. Nigro, H. Colman, L. Soroceanu, P. M. Williams, Z. Modrusan, B. G. Feuerstein, K. Aldape, Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis, Cancer Cell 9 (2006) 157–173. doi:10.1016/j.ccr.2006.02.019.

[5] Y. Jiao, P. J. Killela, Z. J. Reitman, B. A. Rasheed, C. M. Heaphy, R. F. de Wilde, F. J. Rodriguez, S. Rosemberg, S. M. Oba-Shinjo, S. K. N. Marie, C. Bettegowda, N. Agrawal, E. Lipp, C. J. Pirozzi, G. Y. Lopez, Y. He, H. S. Friedman, A. H. Friedman, G. J. Riggins, M. Holdhoff, P. Burger, R. E. McLendon, D. D. Bigner, B. Vogelstein, A. K. Meeker, K. W. Kinzler, N. Papadopoulos, L. A. Diaz, H. Yan, Frequent atrx, cic, fubp1 and idh1 mutations refine the classification of malignant gliomas, Oncotarget 3 (2012) 709–722. doi:10.18632/oncotarget.588.

[6] M. Esteller, Epigenetics in cancer, New England Journal of Medicine 358 (2008) 1148–1159. doi:10.1056/nejmra072067.

[7] D. Hanahan, Hallmarks of cancer: New dimensions, Cancer Discovery 12 (2022) 31–46. doi:10.1158/2159-8290.cd-21-1059.

[8] C. Waddington, The strategy of the genes: a discussion of some aspects of theoretical biology, Allen & Unwin, 1957.

[9] J. Wang, L. Xu, E. Wang, S. Huang, The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation, Biophysical Journal 99 (2010) 29–39. doi:10.1016/j.bpj.2010.03.058.

[10] J. Wang, K. Zhang, L. Xu, E. Wang, Quantifying the waddington landscape and biological paths for development and differentiation, Proceedings of the National Academy of Sciences 108 (2011) 8257–8262. doi:10.1073/pnas.1017017108.

[11] C. Li, J. Wang, Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: Landscape and biological paths, PLoS Computational Biology 9 (2013) e1003165. doi:10.1371/journal.pcbi.1003165.

[12] J. Wang, Landscape and flux theory of non-equilibrium dynamical systems with application to biology, Advances in Physics 64 (2015) 1–137. doi:10.1080/00018732.2015.1037068.

[13] C. Li, J. Wang, Quantifying the landscape for development and cancer from a core cancer stem cell circuit, Cancer Research 75 (2015) 2607–2618. doi:10.1158/0008-5472.can-15-0079.

[14] Q. Li, A. Wennborg, E. Aurell, E. Dekel, J.-Z. Zou, Y. Xu, S. Huang, I. Ernberg, Dynamics inside the cancer cell attractor reveal cell heterogeneity, limits of stability, and escape, Proceedings of the National Academy of Sciences 113 (2016) 2672–2677. doi:10.1073/pnas.1519210113.

[15] W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods, Nature Biotechnology 37 (2019) 547–554. doi:10.1038/s41587-019-0071-9.

[16] C. C. Moore, Ergodic theorem, ergodic theory, and statistical mechanics, Proceedings of the National Academy of Sciences 112 (2015) 1907–1911. doi:10.1073/pnas.1421798112.

[17] J. C. Mauro, M. M. Smedskjaer, Statistical mechanics of glass, Journal of Non-Crystalline Solids 396-397 (2014) 41–53. doi:10.1016/j.jnoncrysol.2014.04.009.

[18] R. Palmer, Broken ergodicity, Advances in Physics 31 (1982) 669–735. doi:10.1080/00018738200101438.

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

[20] M. A. Coomer, L. Ham, M. P. Stumpf, Noise distorts the epigenetic landscape and shapes cell-fate decisions, Cell Systems 13 (2022) 83–102.e6. doi:10.1016/j.cels.2021.09.002.

[21] A. Meister, C. Du, Y. H. Li, W. H. Wong, Modeling stochastic noise in gene regulatory systems, Quantitative Biology 2 (2014) 1–29. doi:10.1007/s40484-014-0025-7.

[22] M. Santillán, On the use of the hill functions in mathematical models of gene regulatory networks, Mathematical Modelling of Natural Phenomena 3 (2008) 85–97. doi:10.1051/mmnp:2008056.

[23] D. T. Gillespie, The chemical langevin equation, The Journal of Chemical Physics 113 (2000) 297–306. doi:10.1063/1.481811.

[24] S. Darmanis, S. A. Sloan, D. Croote, M. Mignardi, S. Chernikova, P. Samghababi, Y. Zhang, N. Neff, M. Kowarsky, C. Caneda, G. Li, S. D. Chang, I. D. Connolly, Y. Li, B. A. Barres, M. H. Gephart, S. R. Quake, Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma, Cell Reports 21 (2017) 1399–1410. doi:10.1016/j.celrep.2017.10.030.

[25] C. Analytics, Metacore, 2019.

[26] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008. URL: http://www.R-project.org, ISBN 3-900051-07-0.

[27] M. Vieira, Gene expression network analysis, 2023. URL: https://github.com/marcosgvjunior/gene-expression-network-analysis, gitHub repository.

[28] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nature Biotechnology 33 (2015) 495–502. URL: https://doi.org/10.1038/nbt.3192. doi:10.1038/nbt.3192.

[29] C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized

negative binomial regression, Genome Biology 20 (2019). doi:10.1186/s13059-019-1874-1.

[30] S. Lab, Using sctransform in seurat, 2022. URL: https://satijalab.org/seurat/articles/sctransform_vignette.html, gitHub repository, [Accessed on 17th July 2022].

[31] A. K. Witkiewicz, V. Kumarasamy, I. Sanidas, E. S. Knudsen, Cancer cell cycle dystopia: heterogeneity, plasticity, and therapy, Trends in Cancer 8 (2022) 711–725. doi:10.1016/j.trecan.2022.04.006.

[32] M. Vieira, Graph matrix and combinatorics, 2023. URL: https://github.com/marcosgvjunior/graph-matrix-and-combinatorics, gitHub repository.

[33] W. R. Inc., Mathematica, Version 11.3, 2018. Champaign, IL, 2018.

[34] W. R. Inc., Kmeans, 2023. URL: https://reference.wolfram.com/language/ref/method/KMeans.html, [Accessed on 9th April 2023].

[35] W. R. Inc., Neighborhood contraction, 2023. URL: https://reference.wolfram.com/language/ref/method/NeighborhoodContraction.html, [Accessed on 9th April 2023].

[36] W. R. Inc., Constrained optimization, 2023. URL: https://library.wolfram.com/infocenter/Books/8506/ConstrainedOptimization.pdf, [Accessed on 12th July 2022].

[37] M. A. Coomer, L. Ham, M. P. H. Stumpf, Noise distorts the epigenetic landscape and shapes cell fate decisions, Proceedings of the National Academy of Sciences (2020). doi:10.1101/2020.12.21.423724.

[38] Bioconductor.org, Rna-seq workflow: gene-level exploratory analysis and differential expression, 2023. URL: https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html, [Accessed on 19th July 2022].

Supplementary material for *Statistical characterization of the dynamics of Glioblastoma Multiforme subtypes through parameters estimation of the epigenetic landscape*

Marcos Guilherme Vieira Junior[a,*], Flávia Raquel Gonçalves Carneiro[b,c,d], Adriano Maurício de Almeida Côrtes[e,f], Nicolas Carels[g], Fabrício Alves Barbosa da Silva[h,*]

[a]*Graduate Program in Computational and Systems Biology, Instituto Oswaldo Cruz (IOC), Oswaldo Cruz Foundation (FIOCRUZ), 20231050, Rio de Janeiro, Brazil*
[b]*Center of Technological Development in Health (CDTS), FIOCRUZ, 21040-361, Rio de Janeiro, Brazil*
[c]*Laboratório Interdisciplinar de Pesquisas Médicas, IOC, FIOCRUZ, 20231050, Rio de Janeiro, Brazil*
[d]*Program of Immunology and Tumor Biology, Brazilian National Cancer Institute(INCA), 20231050, Rio de Janeiro, Brazil*
[e]*Department of Applied Mathematics, Institute of Mathematics, Federal University of Rio de Janeiro, RJ, 21941-909, Brazil*
[f]*Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, RJ, 21941-972 Brazil*
[g]*Laboratory of Biological System Modeling, CDTS, FIOCRUZ, 21040-361, Rio de Janeiro, Brazil*
[h]*Scientific Computing Program, FIOCRUZ, 20231050, Rio de Janeiro, Brazil*

Contents

S1 Mathematical modeling **s2**
- S1.1 Deterministic dynamics s2
- S1.2 Stochastic dynamics s2
- S1.3 Ergodicity in experimental and simulated data s4
- S1.4 Distinguishing Waddington Landscape from Entropy s7

S2 Glioblastoma Multiforme **s7**

S3 Supplementary Sources **s8**

S4 Supplementary results **s9**

*Corresponding authors.
Email addresses:
`marcosvieira.research@gmail.com` (Marcos Guilherme Vieira Junior), `fabricio.silva@fiocruz.br` (Fabrício Alves Barbosa da Silva)

S1. Mathematical modeling

S1.1. Deterministic dynamics

Gene regulation can be modeled as a dynamical system, represented by a set of equations $\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^n$ representing the state vector and $\mathbf{F}(\mathbf{x})$ a vector field defining the dynamics within the gene regulatory network (GRN) model. In this work, we consider an attractor, denoted as $\mathbf{A}$, of the dynamical system as a set of points in the phase space that the system converges to under a specific set of initial conditions. We use the term attractor to represent a fixed point without discussing other types of attractors, such as limit cycles or strange attractors.

The basin of attraction, $\mathcal{B}(\mathbf{A})$, is associated with the attractor $\mathbf{A}$ and is defined as the set of points $\mathbf{x}_0 \in \mathbb{R}^n$ such that the trajectory of the system starting from $\mathbf{x}_0$ converges to $\mathbf{A}$ as time goes to infinity [S1–S6]. In this context, a basin of attraction represents a set of points in the phase space that lead to similar long-term behavior. We hypothesize that each basin of attraction corresponds to a group of genes related to specific conditions, such as genes involved in cancer or the development of cancer subtypes. A more formal definition could be:

$$\mathcal{B}(\mathbf{A}) = \mathbf{x}0 \in \mathbb{R}^n \mid \lim_{t\to\infty} \mathbf{x}(t, \mathbf{x}_0) = \mathbf{A}, \tag{S1}$$

where $\mathbf{x}(t, \mathbf{x}_0)$ denotes the trajectory of the system starting from the initial condition $\mathbf{x}_0$ at time $t = 0$.

The observed dispersion in cancer data raises the question of whether the pattern could be due to chaotic dynamics. This work addresses this question by assuming that the observed pattern could be due to stochastic noise, a chaotic attractor, or both. Regardless of the cause, the dynamics are assumed to occur within their respective basins of attraction.

We interpret gene expression dispersion as an effect of cancer progression. Genetic mutations, epigenetic changes, and altered cellular microenvironments likely increase the system's multiplicity, leading to a more disordered and complex expression distribution. In the context of cancer, higher multiplicity represents a greater number of ways cancerous cells can be organized or interact, often associated with increased complexity and heterogeneity of the tumor microenvironment.

Studies have been investigating the increase in entropy related to these facts, but this also indicates altered states and potentially altered energy levels. Since the differences in energy levels among various cell types are determined mainly by genetics and epigenetics, we seek to study the stability of cancer subtypes as basins of attraction of these altered states using the Waddington landscape. In conclusion, we propose using ergodic theory to study the system's behavior after reaching these configurations rather than focusing on how it reaches them.

S1.2. Stochastic dynamics

The presence of fluctuations in gene expression leads to probability distributions for possible expression values. The temporal evolution of these probability distributions, denoted as $p(\mathbf{x}, t)$, is described by the Fokker-Planck equation as given in [S7]:

$$\frac{\partial}{\partial t} p(\mathbf{x}, t) = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left[\mu_i(\mathbf{x}, t) p(\mathbf{x}, t)\right] + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial^2}{\partial x_i, \partial x_j} \left[D_{ij}(\mathbf{x}, t) p(\mathbf{x}, t)\right] , \tag{S2}$$

where the temporal change of the probability distribution $p(\mathbf{x}, t)$ is determined by two main components: the drift term, represented by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$, and the diffusion term, associated with the diffusion tensor $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$.

The drift term $\mu_i(\mathbf{x}, t)$ represents the deterministic forces acting on the system's variables x_i. These forces drive the system's state towards specific regions in the state space. When the drift term is multiplied by the probability distribution $p(\mathbf{x}, t)$, it essentially represents the average deterministic force experienced by the system at a particular state $\mathbf{x}$ and time t, weighted by the probability of being in that state. In other words, $\mu_i(\mathbf{x}, t)p(\mathbf{x}, t)$ describes the contribution of the deterministic forces to the evolution of the probability distribution, taking into account that the system is more likely to be influenced by these forces in certain regions of the state space.

Similarly, the diffusion term $D_{ij}(\mathbf{x}, t)$ accounts for the stochastic fluctuations in the system's variables x_i and x_j. When this term is multiplied by the probability distribution $p(\mathbf{x}, t)$, it represents the average effect of these stochastic fluctuations on the probability distribution at a particular state $\mathbf{x}$ and time t, weighted by the probability of being in that state. In summary, multiplying the drift and diffusion terms by the probability distribution in the Fokker-Planck equation allows for incorporating the effects of deterministic forces and stochastic fluctuations on the system's dynamics while considering the likelihood of the system being in a particular state.

The partial derivatives in the drift term, $\frac{\partial}{\partial x_i}$, represent the rate of change of the probability distribution concerning the variables x_i. This term captures how the deterministic forces in the system influence the evolution of the probability distribution. The second-order partial derivatives in the diffusion term, $\frac{\partial^2}{\partial x_i, \partial x_j}$, quantify the rate of change of the probability distribution due to stochastic fluctuations in the variables x_i and x_j. This term accounts for the effect of noise in the system.

The diffusion tensor $\mathbf{D}$ is defined as follows:

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2} \sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t)\sigma_{jk}(\mathbf{x}, t) \ . \tag{S3}$$

where $D_{ij}(\mathbf{x}, t)$ represents the diffusion tensors' elements, which quantify the strength of the system's stochastic fluctuations. The diffusion tensor is constructed from the product of the noise intensity matrix $\boldsymbol{\sigma}$ and its transpose $\boldsymbol{\sigma}^{\mathrm{T}}$. The elements of the noise intensity matrix $\sigma_{ik}(\mathbf{x}, t)$ describe the contribution of the k-th noise source to the i-th variable, capturing the magnitude and direction of the noise affecting the system's variables.

In summary, the Fokker-Planck equation (S2) describes the temporal evolution of probability distributions in a system subject to deterministic and stochastic dynamics, as represented by the drift and diffusion terms, respectively. This equation provides a framework for understanding how gene expression fluctuations influence the behavior of biological systems over time.

In the low noise regime, the solution to the Fokker-Planck equation can be shown to be a Gaussian distribution [21], characterized by its first and second moments (first and second coupled moments). Alternatively, one can find relationships between the moments and solve the corresponding system of differential equations. In the low noise regime, the first-order moment corresponds to the deterministic model [S8]. Disregarding cross-correlations, the coupled equation for moments is given by [11, 12]:

$$\begin{aligned} \dot{\mathbf{x}}(t) =& F[\bar{\mathbf{x}}(t)] \ , \\ \dot{\sigma}(t) =& \sigma(t)A^{\mathrm{T}}(t) + A(t)\sigma(t) + 2D[\bar{x}(t)] \ , \end{aligned} \tag{S4}$$

where $\bar{\mathbf{x}}(t)$ is the average value of each variable in time, $\sigma(t)$ is the covariance matrix, A is the jacobian matrix, A^T its transpose, and D is the diffusion parameter.

In this context, the system can be described by a Stochastic Differential Equation (SDE) [14]:

$$dX = \mu(X, t), dt + \sigma(X, t), dW \ , \tag{S5}$$

with drift $\mu(X, t)$, noise parameter $\sigma(X, t)$, and Wiener standard process.

S1.3. Ergodicity in experimental and simulated data

According to the stochastic system outlined above, ergodic theory can be applied to gene expression levels to analyze the long-term behavior of biological systems. By considering the influence of noise as described in the stochastic dynamics section, the basin of attraction definition would change to account for the influence of noise. But how noise shapes the basin, or what could be the limits of the basins such that the fluctuations don't drive the system out of it? To adjust the previous definition, first, we introduce a custom distance function as $d(\mathbf{x}, \mathbf{A})$ that can be chosen according to the specific characteristics of the system [S1]. For instance, it could be an Euclidean distance for symmetric deviations or other distance measures for non-symmetric deviations. The modified definition of the basin of attraction would then be:

$$\mathcal{B}(\mathbf{A}) = \mathbf{x}0 \in \mathbb{R}^n \mid \lim_{t\to\infty} \min_{\tau\geq 0} d(\mathbf{x}(t+\tau, \mathbf{x}_0), \mathbf{A}) \leq \epsilon \,. \tag{S6}$$

where $\mathbf{x}(t, \mathbf{x}_0)$ denotes the trajectory of the system starting from the initial condition $\mathbf{x}_0$ at time $t = 0$, $\epsilon > 0$ is a small positive number representing the tolerance level within which the system oscillates around the attractor $\mathbf{A}$, and $\tau \geq 0$ is a time shift that accounts for the oscillation around the attractor. This way, instead of converging to a fixed point, the noise drives the system to remain within the neighborhood of the attractor.

In practice, the choice of the distance function and the tolerance level depends on the characteristics of the system being studied and the specific dynamics around the attractor. For example, if the system's deviations are non-symmetric or have long-tailed distributions, alternative distance functions, such as the Mahalanobis distance or other appropriate measures, can better capture these properties. On the other hand, the tolerance level ϵ should accurately represent the fluctuations around the attractor, considering the inherent noise and the system's dynamical properties. In this work, by assuming Gaussian distributions, we choose Euclidean distances and ϵ empirically verified by trial and error of compatibility test with simulations.

In this way, by accounting for the stochastic nature of the system, the points representing the gene expression levels within a basin of attraction will be distributed according to a probability distribution. Here we proposed to identify the basins of attraction using clustering algorithms and ergodic theory to study the stability of these basins and their transition with each other. By doing so, we sought to understand the possible relations between gene regulation dynamics and expression patterns observed in experimental data. If a system is ergodic, it implies that its time average over a long period is equivalent to its ensemble average over all possible system states. Considering the quantification of gene expression for gene i at time t as $x_i(t)$, the time average $\langle x_i \rangle_t$ is given by [S9]:

$$\langle x_i \rangle_t = \lim_{T\to\infty} \frac{1}{T} \int_0^T x_i(t) dt, \tag{S7}$$

where T is the time interval over which $x_i(t)$ is observed. The ensemble average $\langle x_i \rangle_{ss}$ for the gene i of the gene expression quantification x_i over the stationary state, given the stationary probability distribution $P(x_i)$, can be calculated as [S9]:

$$\langle x_i \rangle_{ss} = \sum_j x_i^j P(x_i^j) \quad or \quad \langle x_i \rangle_{ens} = \frac{1}{N} \sum_{j=1}^{N} x_i^j, \tag{S8}$$

where x_i^j represents the expression quantification of gene i in the j-th realization of the system, N is the total number of realizations or observations, and the summation are taken over all possible quantification levels of gene i. The equivalence of the ensemble average with the time average is given by $\langle x_i \rangle_{ss} = \langle x_i \rangle_t$.

We supposed some conditions to be satisfied to verify the presence of ergodicity in experimental data or simulations. First, the system properties are statistically stationary over time, which does not necessarily

imply thermal equilibrium, but instead that the observation time be long enough to capture the system's dynamics and fluctuations. Second, the sampling must be sufficiently large to capture a wide range of possible states.

Instead of supposing ergodicity over all phase space, this work assumes ergodicity within the system's components (internal ergodicity [18]). Here we consider each basin of attraction as a component, thinking that the likelihood of jumping between them was low but possible. One possible approach to quantify the presence of internal ergodicity in gene expression data is to assume that the probability distribution observed in experimental data corresponds to the stationary distribution within each attractor.

This way, experimental validation of internal ergodicity can be done by analyzing gene expression data from different conditions and identifying clusters of genes that exhibit similar expression patterns. After that, numerical simulation can be optimized to generate synthetic data with attractors compatible with the experimental. Finally, once consistent, we could check if the numerical simulation that generated compatible distributions satisfies the necessary conditions.

In this sense, let x_i be the expression level of gene i, and T be a time interval over which x_i is observed. We say that x_i exhibits internal ergodicity if we observe:

- Ergodicity inside $\mathcal{B}(\mathbf{A})$
- Ergodicity outside $\mathcal{B}(\mathbf{A})$ (complement of $\mathcal{B}(\mathbf{A})$ - one or more basins of attraction)
- Time scale separation between $\mathcal{B}(\mathbf{A})$ and the outside system

However, we also considered a weaker type of ergodicity with which the approximation of the time average by the sample average improves as the length of a subinterval $T_{12} = t_2 - t_1$ increases and as the sample size N increases. In other words, the more data we have regarding the temporal interval and the number of samples, the better the approximation between the time average and the sample average for gene expression levels. With a time average $\langle x_i \rangle_{T_{12}}$ given by:

$$\langle x_i \rangle_{T_{12}} = \frac{1}{T_{12}} \int_{t_1}^{t_2} x_i(t')dt' \tag{S9}$$

with the time average of x_i in the subinterval $[t_1, t_2]$, and the sample average:

$$\langle x_i \rangle_{A_{ss}} = \frac{1}{N} \sum_{j=1}^{N} x_i^j, \tag{S10}$$

where N is the sample size in the basin of attraction $\mathcal{B}(\mathbf{A_{ss}})$, and x_i^j represents the j-th sample of the gene expression level x_i within the basin of attraction $\mathcal{B}(\mathbf{A_{ss}})$. The weaker type of ergodicity we propose is:

$$\lim_{T_{12}\to\infty, N\to\infty} \left| \langle x_i \rangle_{T_{12}} - \langle x_i \rangle_{A_{ss}} \right| = 0, \tag{S11}$$

whereas the length of the subinterval T_{12} and the sample size N approach infinity, the difference between the time average and the space average approaches zero. In other words, it is an equivalence in a statistical sense.

Finally, time scale separation helps assess whether the relaxation time of x_i is shorter within $\mathcal{B}(\mathbf{A})$ compared to outside of it. By examining the decay rates of the autocorrelation function for $x_i(t)$ both inside and outside $\mathcal{B}(\mathbf{A})$, the relaxation time can be estimated. The autocorrelation function, which measures the similarity between a variable and its lagged version, can provide this insight. A faster decay within $\mathcal{B}(\mathbf{A})$ indicates a shorter relaxation time, suggesting a greater tendency for the system to reach equilibrium. The autocorrelation function of x_i is given by:

$$r_{x_i x_i}(h) = \frac{\sum_{i=1}^{n-h}(x_i - \bar{x})(x_{i+h} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{S12}$$

where $r_{x_i x_i}(h)$ is the autocorrelation at lag h, x_i is the value of the time series at time step i, $\bar{x}$ is the mean of the time series, and n is the total number of data points in the time series. The numerator calculates the sum of the product of deviations between the original time series and its lagged version. In contrast, the denominator calculates the sum of squared differences from the mean, equivalent to the actual time series variance multiplied by the number of data points (n). The autocorrelation is the ratio of the sum of the product of deviations to the sum of squared differences from the mean, normalizing the measure between -1 and 1.

In a dynamic system with a specific driving force, the autocorrelation function at different lags can provide valuable insights into the system's behavior and relationship with the driving force. When the derivatives of the driving force are greater than zero, it implies that the system is being guided by 'a rule' due to the network interactions. In this case, the autocorrelation values at different lags may not oscillate around zero, as the time series exhibits a persistent structure caused by the driving force. As the lag increases, the autocorrelation values might decrease, but they could remain positive, suggesting that the influence of the driving force persists over time.

On the other hand, when the system reaches a steady state, the driving force's influence becomes constant or negligible, and the time series may primarily be affected by stochastic fluctuations. In this case, the autocorrelation values at different lags are more likely to oscillate around zero, suggesting a stationary or random process. Furthermore, the autocorrelation function will show little or no correlation between the time series values at different lags, indicating that the time series has become more unpredictable and less dependent on its past values. By using the clusters (i.e., basins centroids) as starting points, we expected to observe this behavior, with deviations possibly related to non-stability or transitions.

To proceed with the autocorrelation analysis, we defined the autocorrelation function as a vector, $r(h)$, where each element represents the autocorrelation at a specific lag h. The autocorrelation function $r(h)$ can be calculated as follows:

$$r(h) = [r_{x_i x_i}(1), r_{x_i x_i}(2), \ldots, r_{x_i x_i}(n-1)] \tag{S13}$$

where each autocorrelation value, $r_{x_i x_i}(h)$, is calculated using equation (S12). This list of autocorrelation values for different lags has a connection to the times t of the time series. By computing the autocorrelation values for increasing lags, we compare the time series values that are further apart in time. The connection to the time t of the time series is inherent in the fact that the autocorrelation measures the extent to which values at different time points in the time series are linearly related or correlated with each other. To quantify the relaxation time, we can define a threshold for the autocorrelation values as follows:

$$h^* = \min h \mid r(h) < e^{-1} \tag{S14}$$

where, h^* represents the smallest lag h at which the autocorrelation value falls below the threshold e^{-1}.

In conclusion, these methods provide a way to test the conditions, quantify the internal ergodicity of gene expression in the simulated dataset, and suggest its presence in the experimental data. This shows how internal ergodicity could be essential in studying gene expression of single-cell RNA-seq data.

S1.4. Distinguishing Waddington Landscape from Entropy

The Waddington landscape and entropy address different aspects of a system, although both involve probabilities and logarithms. The Waddington landscape represents a landscape guiding cellular development along different paths derived from the steady-state probability distribution of cells (P_{ss}). This landscape's mathematical form is analogous to Boltzmann's distribution in equilibrium statistical mechanics, but it does not represent actual energy, and the system is generally not in thermal equilibrium.

Entropy, on the other hand, quantifies the degree of disorder or randomness in a system. It involves calculating the sum of probabilities weighted by the logarithm of the probabilities themselves, which quantifies the uncertainty or randomness associated with each microstate, resulting in the total entropy of the system. Different types of entropy are relevant in specific contexts, such as intrinsic entropy and non-equilibrium entropy. For further details and discussion on these concepts, readers are referred to sources such as [12] and other relevant literature.

In summary, both the Waddington landscape and entropy concepts help us understand the behavior of biological systems, such as gene regulation and cellular development. However, they address different aspects of a system: the Waddington landscape focuses on landscapes that guide cellular development, while entropy measures the degree of disorder or randomness in a system.

S2. Glioblastoma Multiforme

Despite all the progress in its understanding and treatment, cancer is still one of the leading causes of death worldwide [1]. With all its biological diversity and complexity, efforts to beat it have let to understand its mechanisms of development and forms of treatment [7]. Among the different types of cancer, Glioblastoma Multiforme is the most common and aggressive type of brain tumor. With an average survival of around 15 months and only approximately 10% probability of 5-year overall survival, it has peculiarities due to its anatomical location and physiological characteristics. In general, without presenting symptoms before the malignant state, it has its highest incidence in individuals of more advanced ages, with approximately 50% cases for patients over 65 years old.

GBM can occur as a primary tumor - accounting for more than 90% of cases and with a worse prognosis - and as a secondary tumor - representing approximately 5% of cases, being rarer in younger individuals [2]. Both the primary and the secondary tumors have different characteristics. Primary tumors are observed with a greater expression of the epidermal growth factor receptor (*EGFR*) and mutations in the homologous phosphatase to tensin (*PTEN*) gene. In contrast, secondary tumors show mutations in the tumor suppressor gene *TP53*, not observed in primary tumors [2]. In addition to the division into primary and secondary tumors, GBM can still be classified into different subtypes according to various criteria and markers. The validation of markers is a difficult task. However, it is essential in the development of treatments. The main classification proposals are from Verhaak, Philips, and Jiao. Verhaak et al. proposed four subtypes of GBM: Classical, Mesenchymal, Neural, and Proneural. Philips included a proliferative, marker-enriched subtype of NSCs. Jiao focused on differences in mutations in Isocitrate Dehydrogenase (*IDH*) [2–5].

Verhaak et al., in addition to the classification of subtypes, observed similarities in the markers of each subtype and different cell types. For example, the Classic subtype is closer to neurons and astrocytes; the Neural presents markers typical of neurons; the Mesenchymal with both mesenchymal and astrocyte markers, and the Proneural with high expression of oligodendrocyte genes [3]. In general, the Classic subtype has high levels of expression for *EGFR*, the Mesenchymal has low expression of the *NF1* gene (Neurophymatosis type 1) and high *MET*, the Proneural subtype features altered Receptor Type Alpha for Platelet-Derived Growth Factor (*PDGFRA*) and mutations in *IDH1*, and the Neural subtype exhibits mutations in *TP53*, high expressions of *EGFR*, and deletions in inhibitor cyclin-dependent kinase 2A (*CDKN2A*). Additionally, a relationship was

found between the subventricular zone (SVZ) proximity and subtype characteristics: the Neural and Proneural types are closer to the SVZ, displaying faster progression, lower survival rates, and greater similarity with NSCs, whereas the Classic and Mesenchymal types are more distant, showing a diffuse aggregation pattern and a better response to more aggressive treatments. This association implies that the proximity to the SVZ is a prognostic factor related to pathogenesis [2].

Standard treatment consists of surgical approaches, radiotherapy, and chemotherapy treatments to minimize damage and avoid severe consequences to the nervous system. However, as a complicating factor, the blood-brain barrier still considerably limits the chemotherapy options, which requests for studies capable of proposing viable alternatives [S10].

S3. Supplementary Sources

[S1] L. Perko, Differential Equations and Dynamical Systems, Springer New York, 2001. doi:10.1007/978-1-4613-0003-8.

[S2] S. Wiggins, Introduction to Applied Nonlinear Dynamical Systems and Chaos, Texts in Applied Mathematics, Springer New York, 2003. URL: https://books.google.com.br/books?id=RSI4RGdwnU4C.

[S3] S. Strogatz, Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering, Second Edition, CRC Press, 2018. URL: https://books.google.com.br/books?id=wUBvDwAAQBAJ.

[S4] J. Guckenheimer, P. Holmes, Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields, Springer New York, 1983. doi:10.1007/978-1-4612-1140-2.

[S5] K. Alligood, T. Sauer, J. Yorke, Chaos: An Introduction to Dynamical Systems, Textbooks in Mathematical Sciences, Springer New York, 2000. URL: https://books.google.com.br/books?id=i633SeDqq-oC.

[S6] E. Pitzer, M. Affenzeller, A. Beham, A closer look down the basins of attraction, in: 2010 UK Workshop on Computational Intelligence (UKCI), IEEE, 2010. doi:10.1109/ukci.2010.5625595.

[S7] Kampen, Stochastic processes in physics and chemistry, Elsevier, Amsterdam Boston London, 2007.

[S8] D. Bover, Moment equation methods for nonlinear stochastic systems, Journal of Mathematical Analysis and Applications 65 (1978) 306–320. doi:10.1016/0022-247x(78)90182-8.

[S9] C. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences, Springer complexity, Springer, 2004. URL: https://books.google.com.br/books?id=wLm7QgAACAAJ.

[S10] O. Gallego, Nonsurgical treatment of recurrent glioblastoma, Current Oncology 22 (2015) 273–281. doi:10.3747/co.22.2436.

S4. Supplementary results

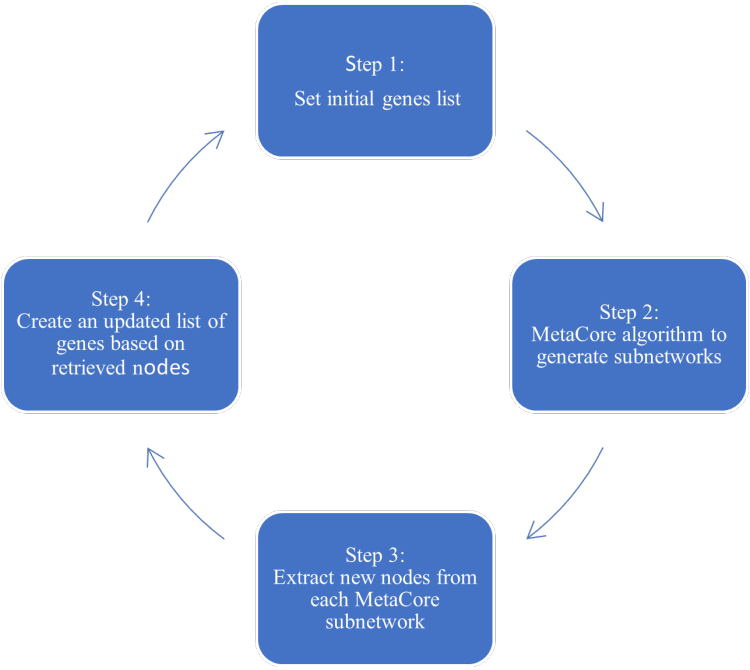


Figure S1: GRN building procedure: (i) Select a group of initial (marker) genes. (ii) Input the list into the MetaCore GRN generator. (iii) Retrieve additional nodes. (iv) Create a new list with the initial nodes and those added by the MetaCore algorithm. (v) Use the updated list as the initial group of genes. This cycle can be repeated until the network is interconnected or until no new nodes are added. The process was performed twice in this case, but the number of iterations can be adjusted according to specific requirements.

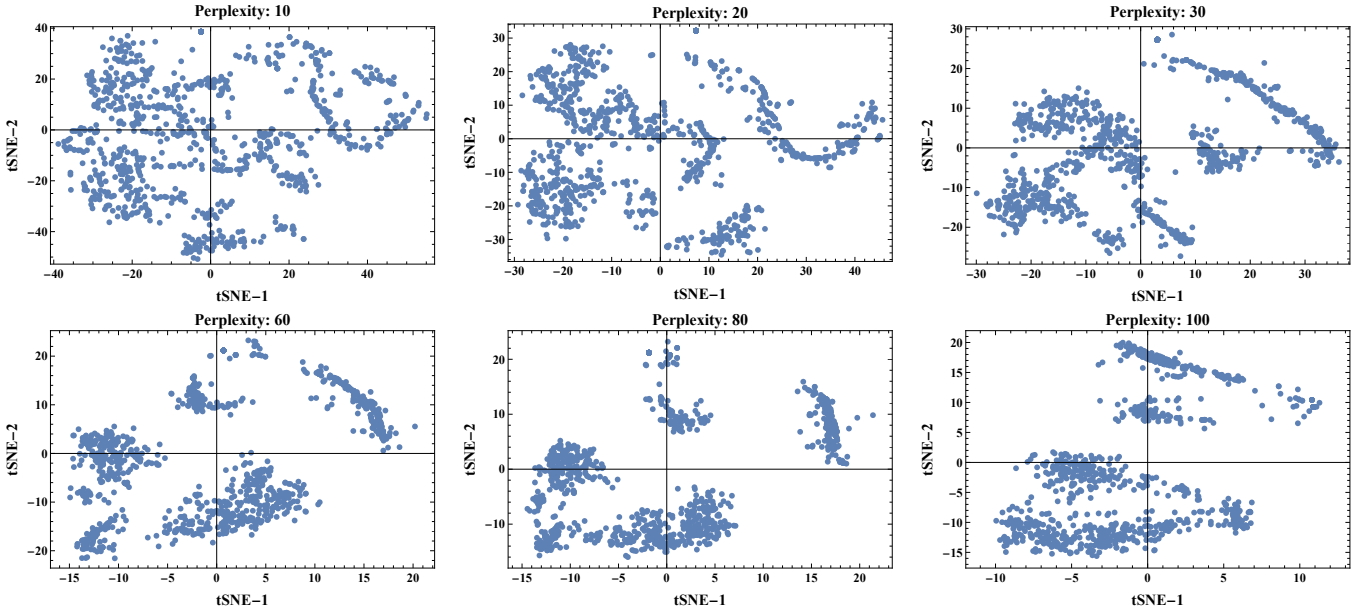


Figure S2: tSNE-based dimensional reduction visualizations with varying perplexity values (10, 20, 30, 60, 80, and 100). These plots demonstrate the effect of different perplexity values on the clustering and structure of the data, helping to identify the most suitable parameter for further analysis. The axes represent the two reduced dimensions obtained from tSNE.

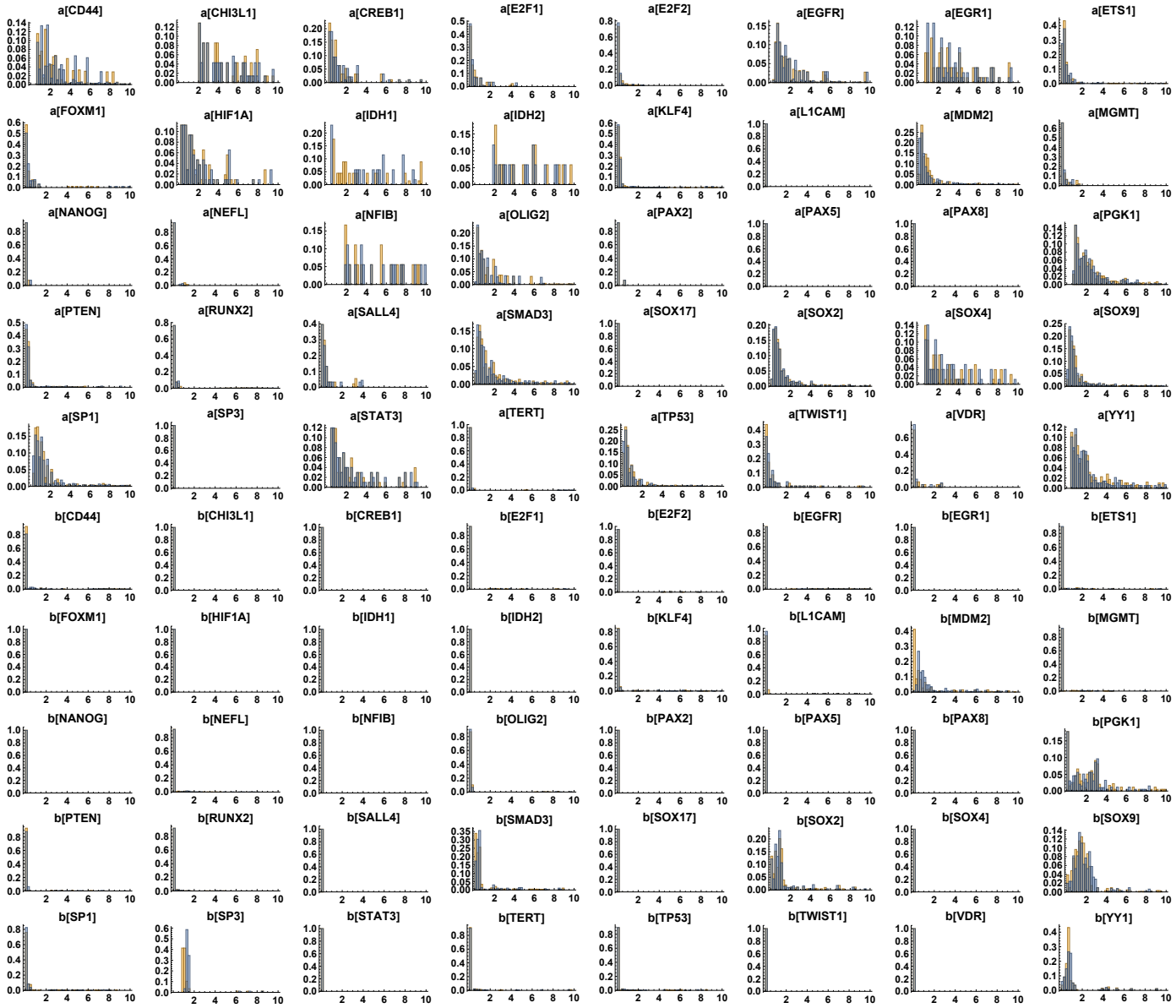


Figure S3: Distributions of the initial parameter estimation after clustering using k-means and NbC methods. The a[] and b[] notation refers to each gene's activation and inhibition parameters. The orange values represent k-means, and the blue values represent NbC. These parameters multiply all interactions in each gene equation, with the same values applied across all considered clusters. The horizontal axis shows the parameter values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

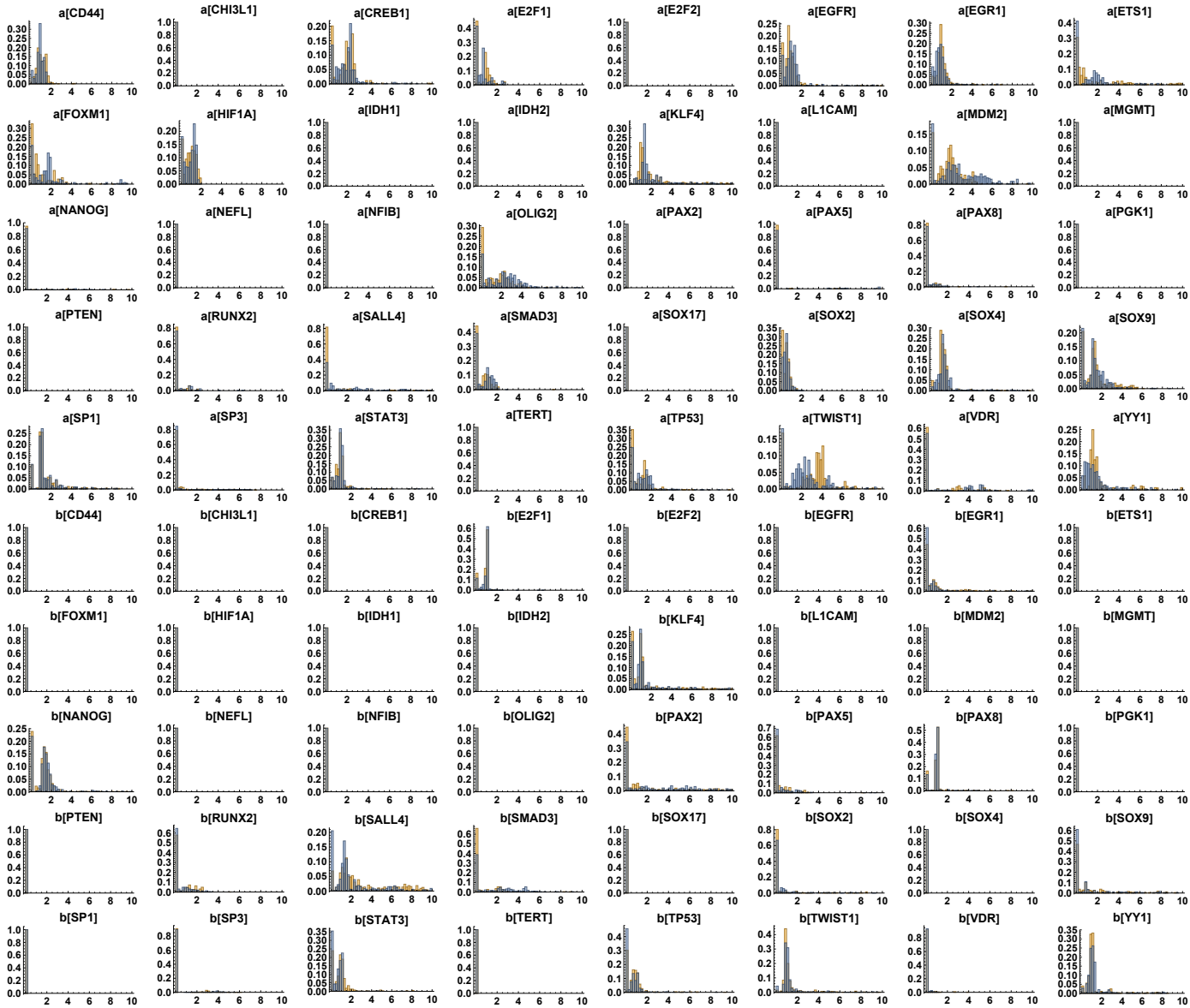


Figure S4: Distributions of the second parameter estimation after clustering using k-means and NbC methods. The a[] and b[] notation refers to each gene's activation and inhibition parameters. The orange values represent k-means, and the blue values represent NbC. These second parameters multiply all interactions in all gene equations regulated by these edges. It multiplies the activation or inhibition parameters previously computed to balance the dynamics better. The horizontal axis shows the parameter values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

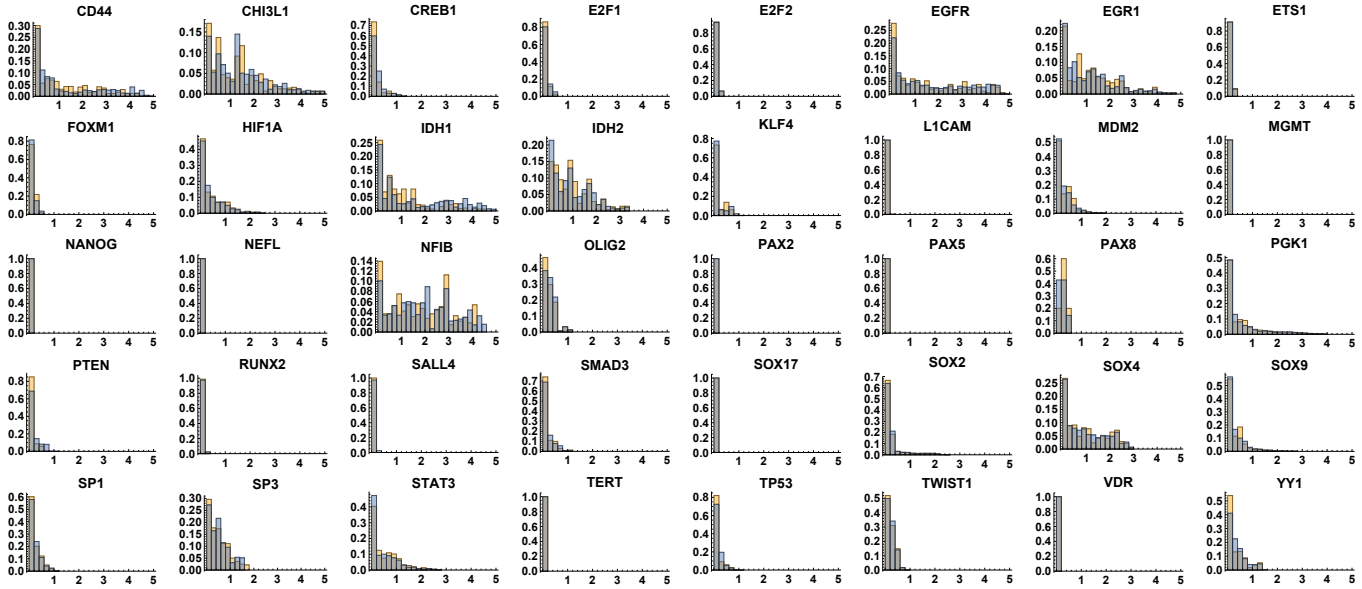


Figure S5: Superimposed residual distributions for the first parameters estimation after clustering using k-means and NbC methods. The orange values represent k-means, and the blue values represent NbC. For each estimation, the number of residuals for each gene was equal to the number of clusters (due to considering all clustering simultaneously). The horizontal axis shows the residual values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

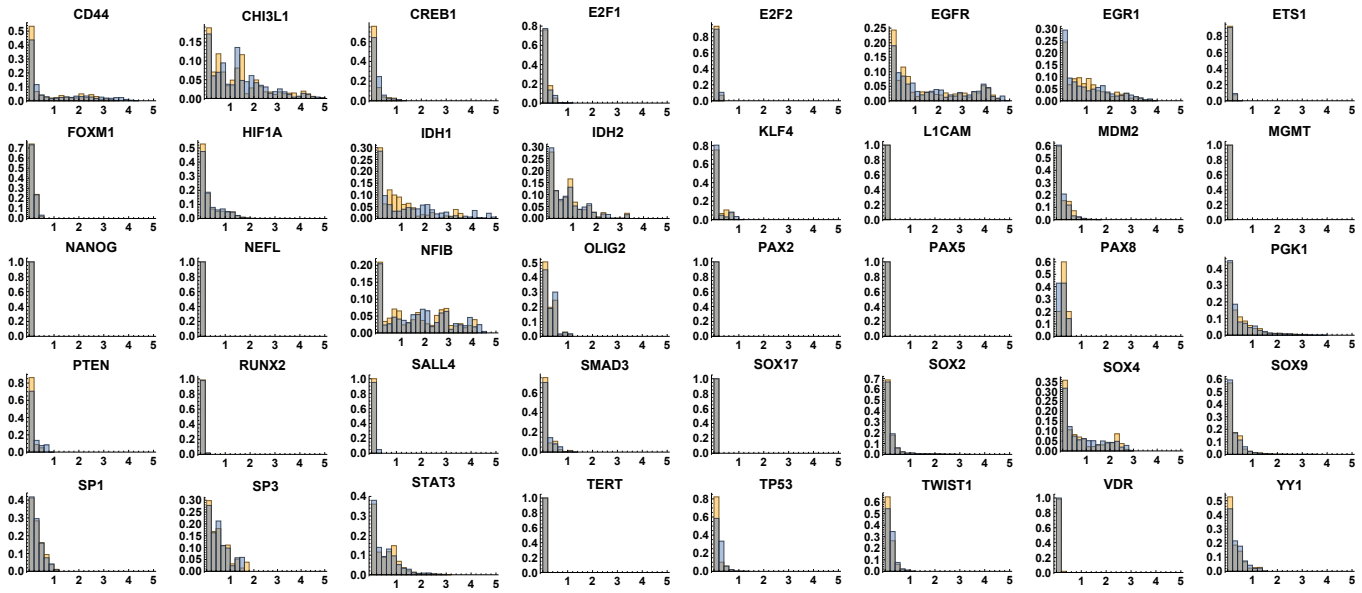
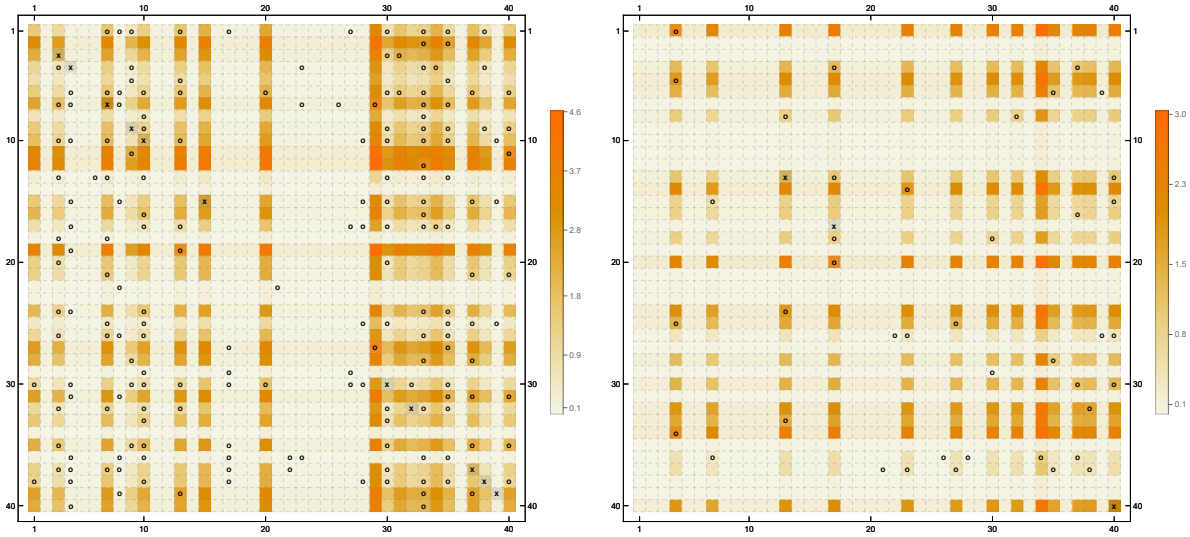
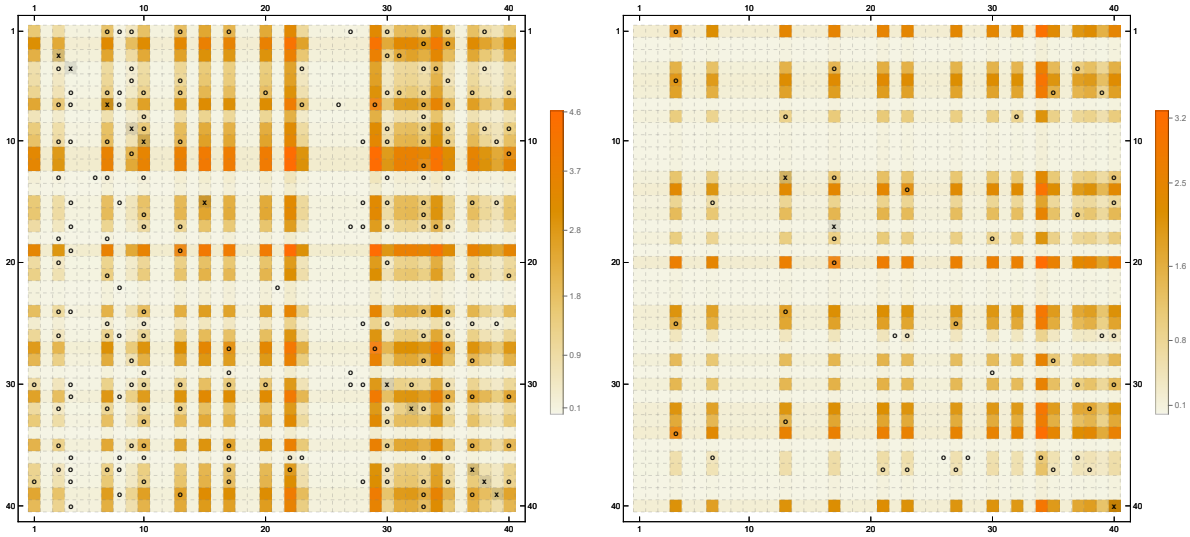


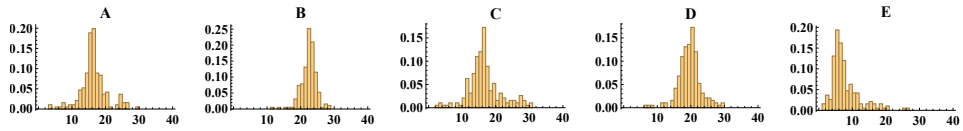
Figure S6: Superimposed residual distributions for the second parameters estimation after clustering using k-means and NbC methods. The orange values represent k-means, and the blue values represent NbC. For each estimation, we had the number of residuals for each gene equal to the number of clusters (due to considering all clustering simultaneously). The horizontal axis shows the residual values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.



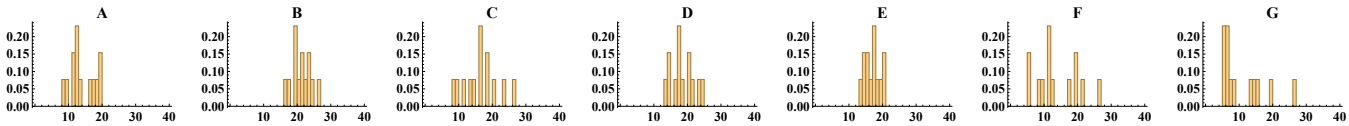
(a)



(b)

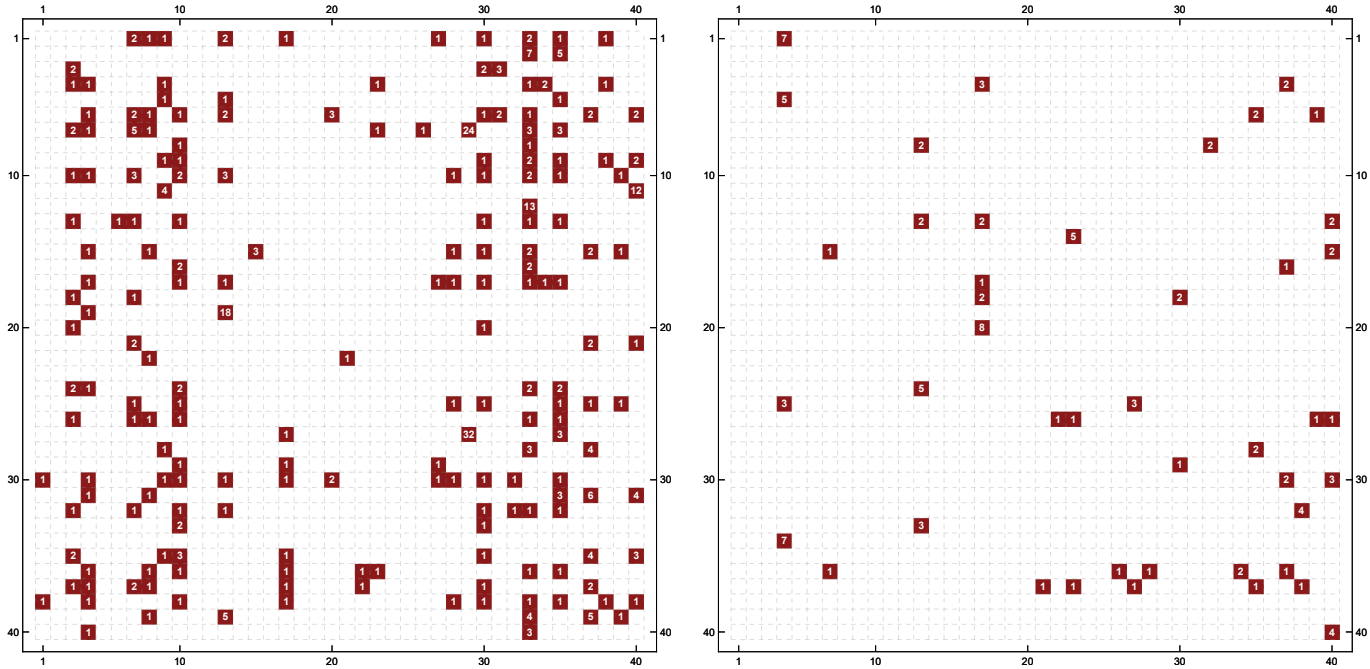


(c)

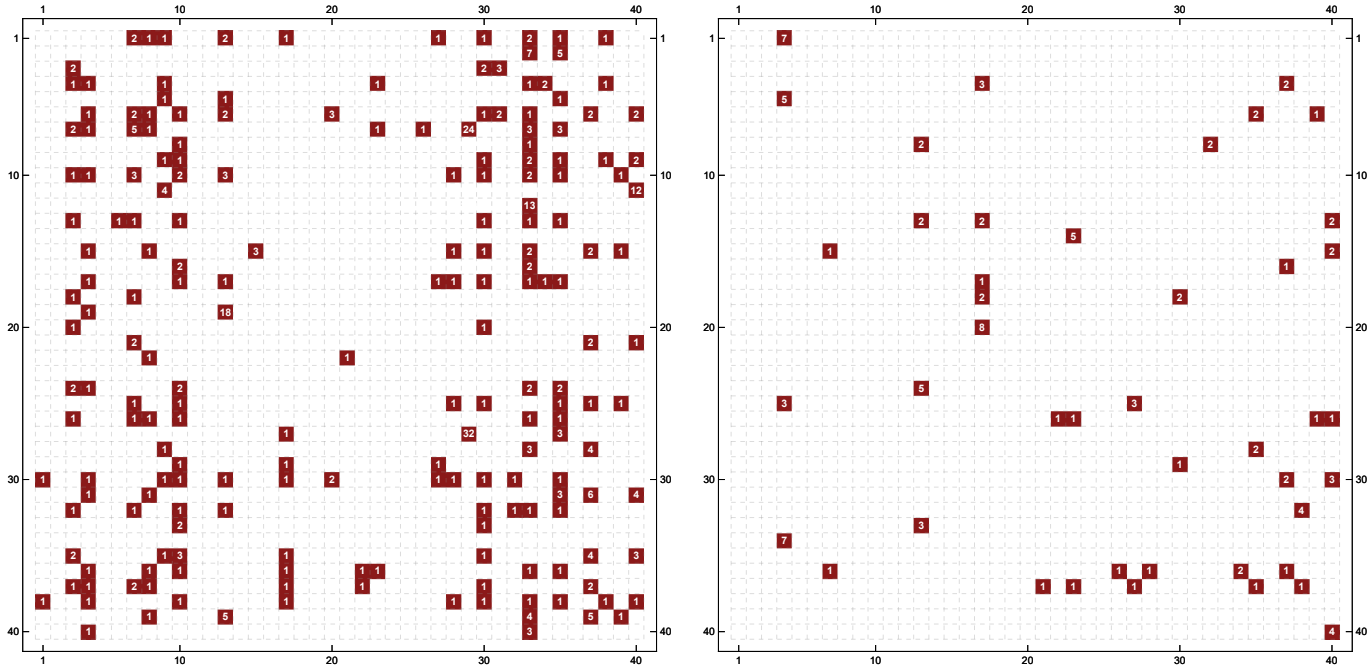


(d)

Figure S7: Figures (a) and (b) show the heatmap with the color gradient for the logarithm of the parameter values considering k-means and NbC, respectively. The 'o' represents the regulations, and the 'x' is the self-regulation edges in the network. Left: activation parameters. Right: inhibition parameters. The matrix indices in (a) and (b) represent each variable/node of the model. Figures (c) and (d) represent the distribution of the number of genes that match each cluster, considering all tested combinations of the multiplicative factors for *a*, *sa*, *b*, and *sb*. In (c) and (d), the horizontal axis represents the count of compatibility cases, and the vertical axis represents normalized counts or frequencies to indicate probabilities.

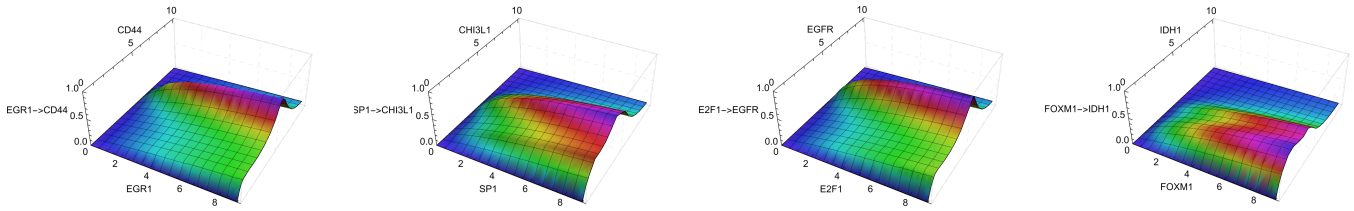


(a)

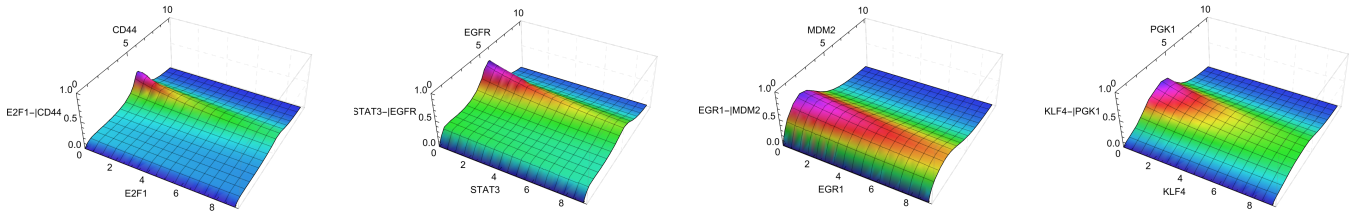


(b)

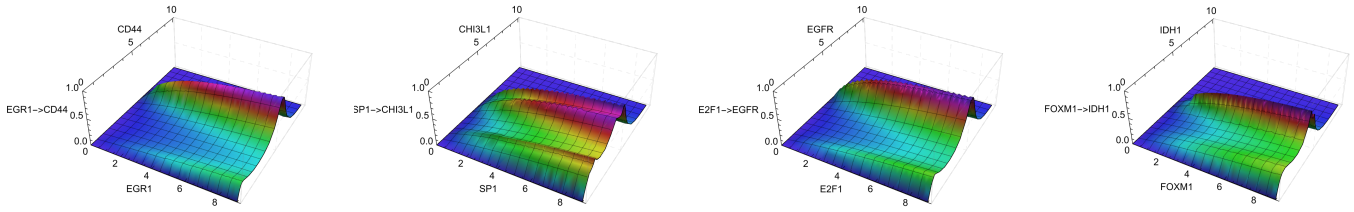
Figure S8: Figures (a) e (b) show the ceil() of the values of all parameters for the k-means clustering and NbC clustering, respectively. The left images correspond to activation parameters, and the right images to inhibition parameters. The ceil() was used to help the visualization of the values. The matrix indices in (a) and (b) represent each variable/node of the model.



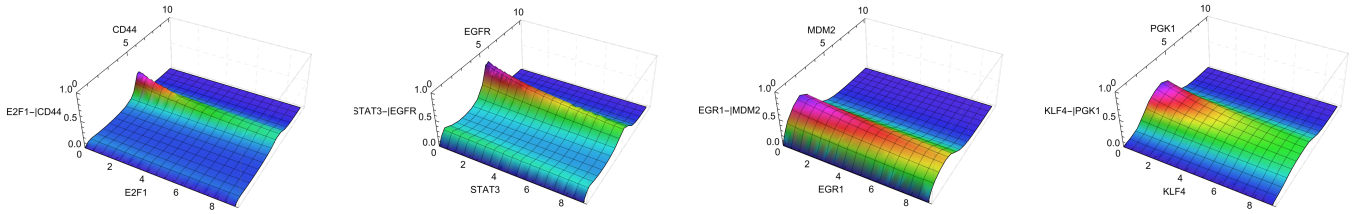
(a)



(b)



(c)



(d)

Figure S9: New regulation function V with $n = 1$ and $h_2(x)$ for different combinations of genes and/or transcription factors. The horizontal axis represents the transcription factor and gene quantification using the normalized amount of single-cell RNA sequencing of experimental data. The vertical axis represents the quantification of the interaction regulations. (a) Activation values using the five k-means clusters and $f_a = 0.1$; (b) Inhibitory interactions using the five k-means clusters and $f_b = 1.3$. (c) Activation values using the seven Neighborhood Contraction clusters and $f_a = 0.1$; (d) Inhibitory interactions using the seven Neighborhood Contraction clusters and $f_b = 1.1$.

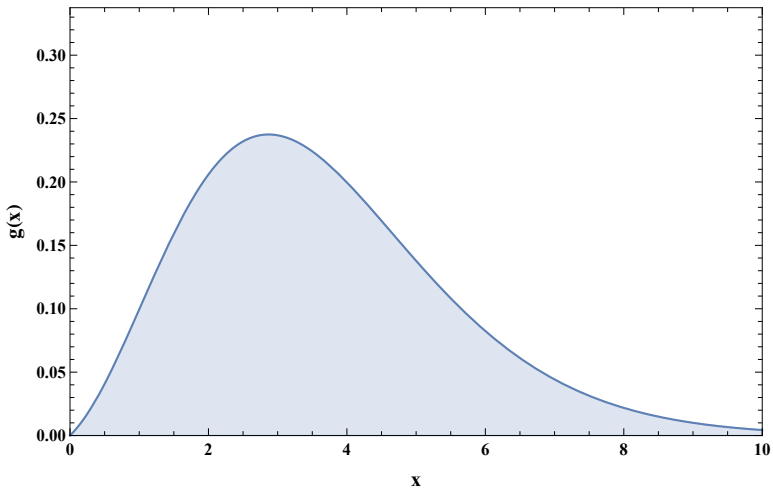


Figure S10: Function used for the multiplicative noise after manual verification of compatibility between the experimental data and those obtained by the simulation The horizontal axis represents the expression level quantification, and the vertical axis represents the noise amplitude before multiplication by the amplitude that was defined in the parameter estimation.

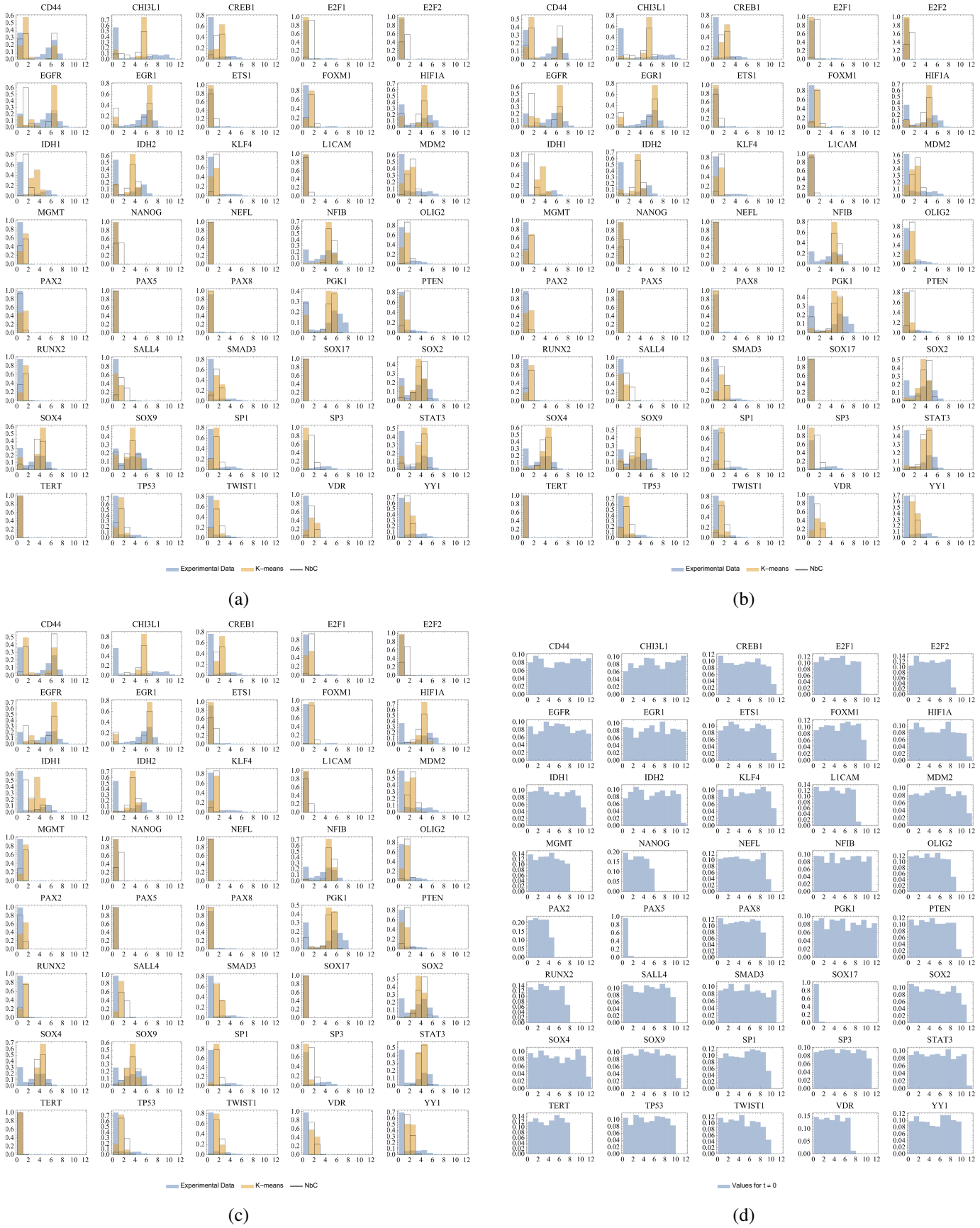


Figure S11: (a) to (c): Superposition of experimental (blue) and simulated (orange: K-means; black line: NbC) data for each gene of the GRN; (a) t = 50 (500 time steps); (b) t = 25 (250 time steps); (c) t = 5 (50 time steps); (d) Distribution of simulated initial conditions. The expression values are on the x-axis, and the relative frequency in the y-axis. The interval goes from 0 to 12 in unity steps. The different times were used to analyze the probability distribution evolution over time. The horizontal axis shows the expression values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

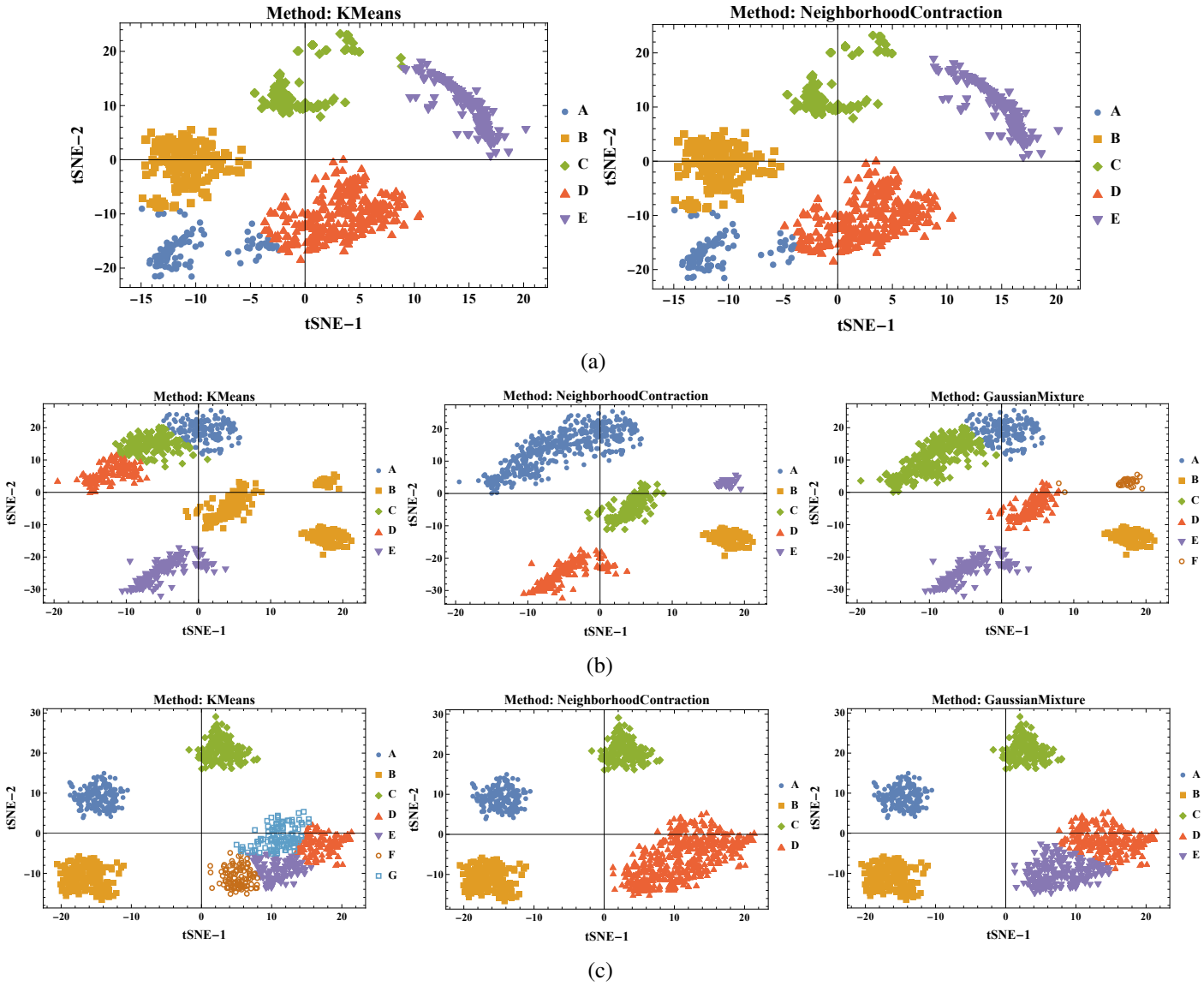


Figure S12: Clustering results using different methods. The axes represent the two reduced dimensions obtained for each tSNE. (a) Experimental data; (b) and (c) simulated data. The letters represent the identified clusters. These clusters were defined in this 2-dimensional space and were used solely to assess the clustering quality. The actual clustering was performed in the 4-markers dimensional space. All methods used the built-in Mathematica functions with PerformanceGoal set to ”Quality”, CriterionFunction set to ”StandardDeviation”, and DistanceFunction set to EuclideanDistance.

Figure S13: Cluster centroids superimposed on experimental data distributions for each gene. (a) Centroids found using k-means with experimental data; (b) Centroids found using k-means with simulated data after parameter estimation using k-means clusters; (c) Centroids found using NbC; (d) Centroids found using Gaussian Mixture. The horizontal axis shows the expression values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

Figure S14: Cluster centroids superimposed on experimental data distributions for each gene. (a) Centroids found using NbC with experimental data; (b) Centroids found using k-means with simulated data after parameter estimation using NbC clusters; (c) Centroids found using NbC; (d) Centroids found using Gaussian Mixture. The horizontal axis shows the expression values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

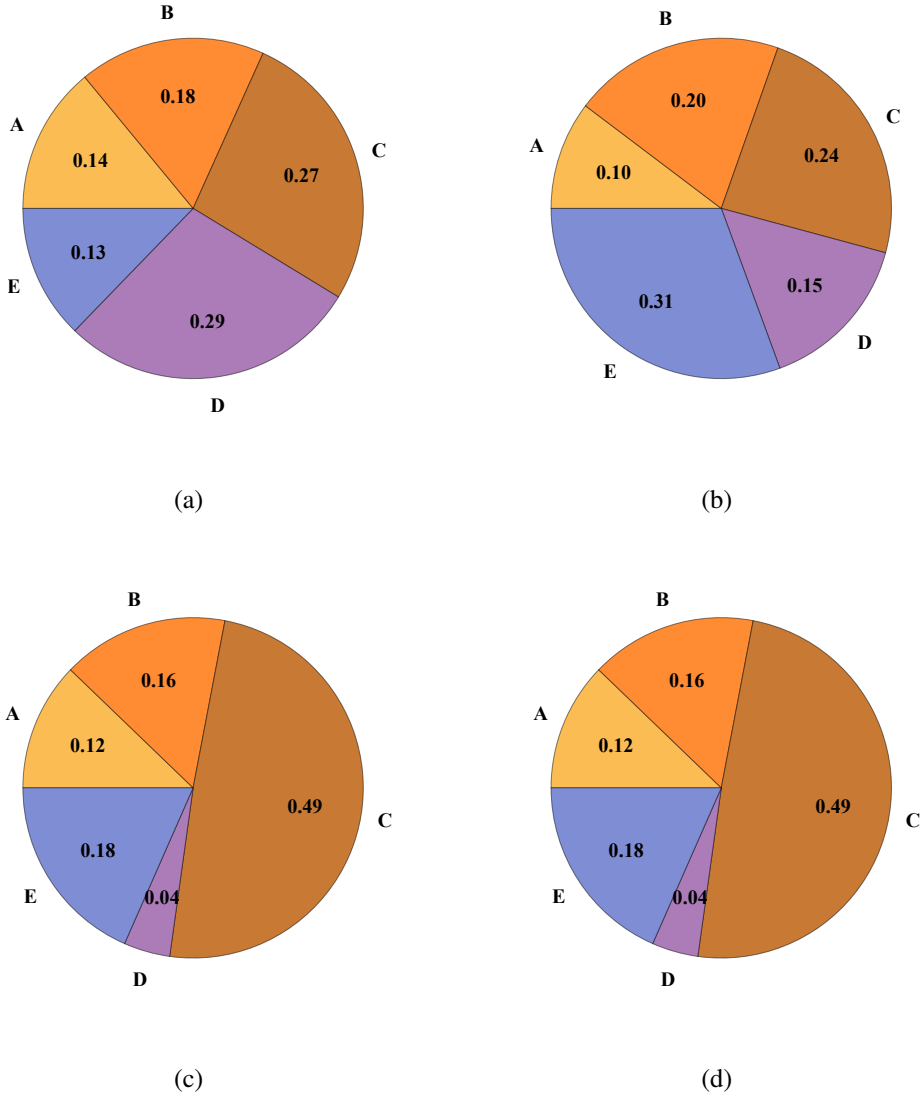


Figure S15: Pie charts representing the number of elements within each cluster. (a) Experimental data clusters using k-means; (b) to (d) Clusters of simulated data after parameter estimation using k-means clusters. (b) k-means; (c) NbC; (d) Gaussian Mixture.

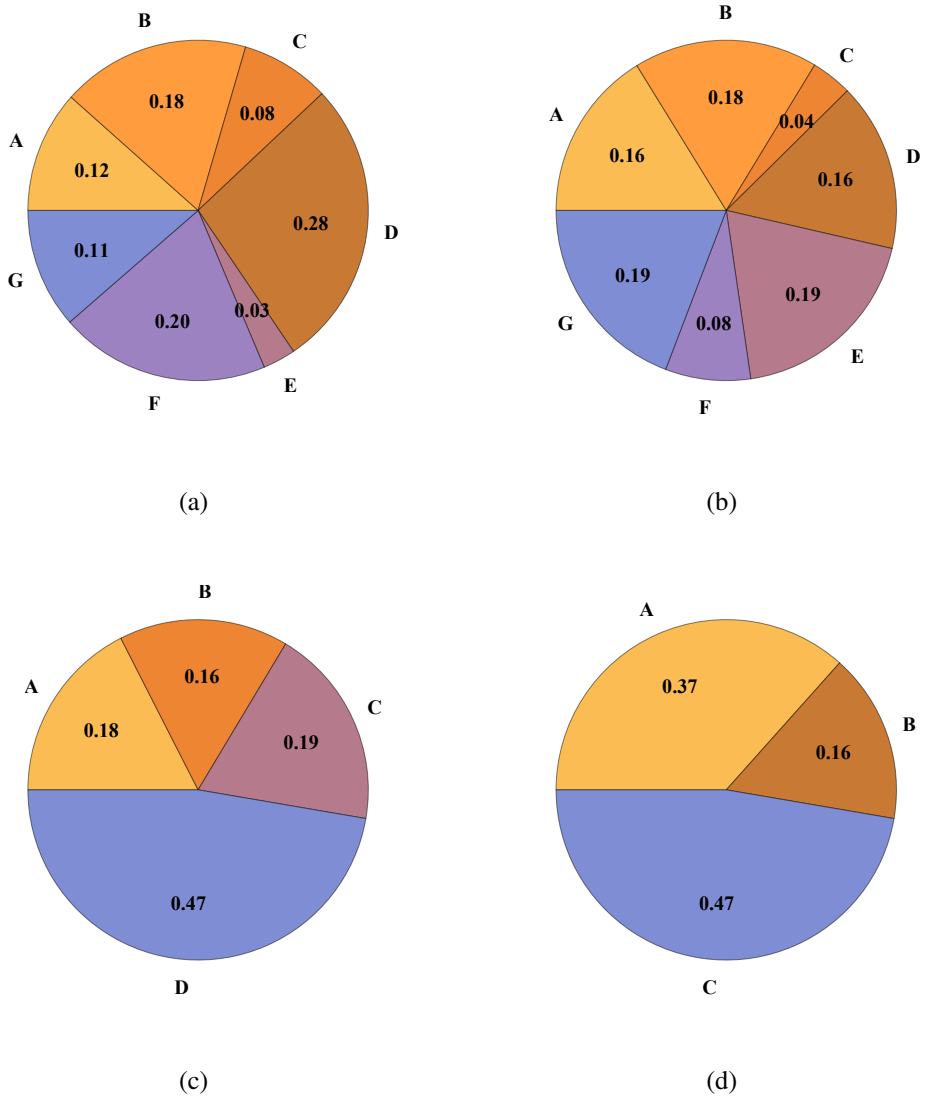


Figure S16: Pie charts representing the number of elements within each cluster. (a) Experimental data clusters using NbC; (b) to (d) Clusters of simulated data after parameter estimation using NbC clusters. (b) k-means; (c) NbC; (d) Gaussian Mixture.

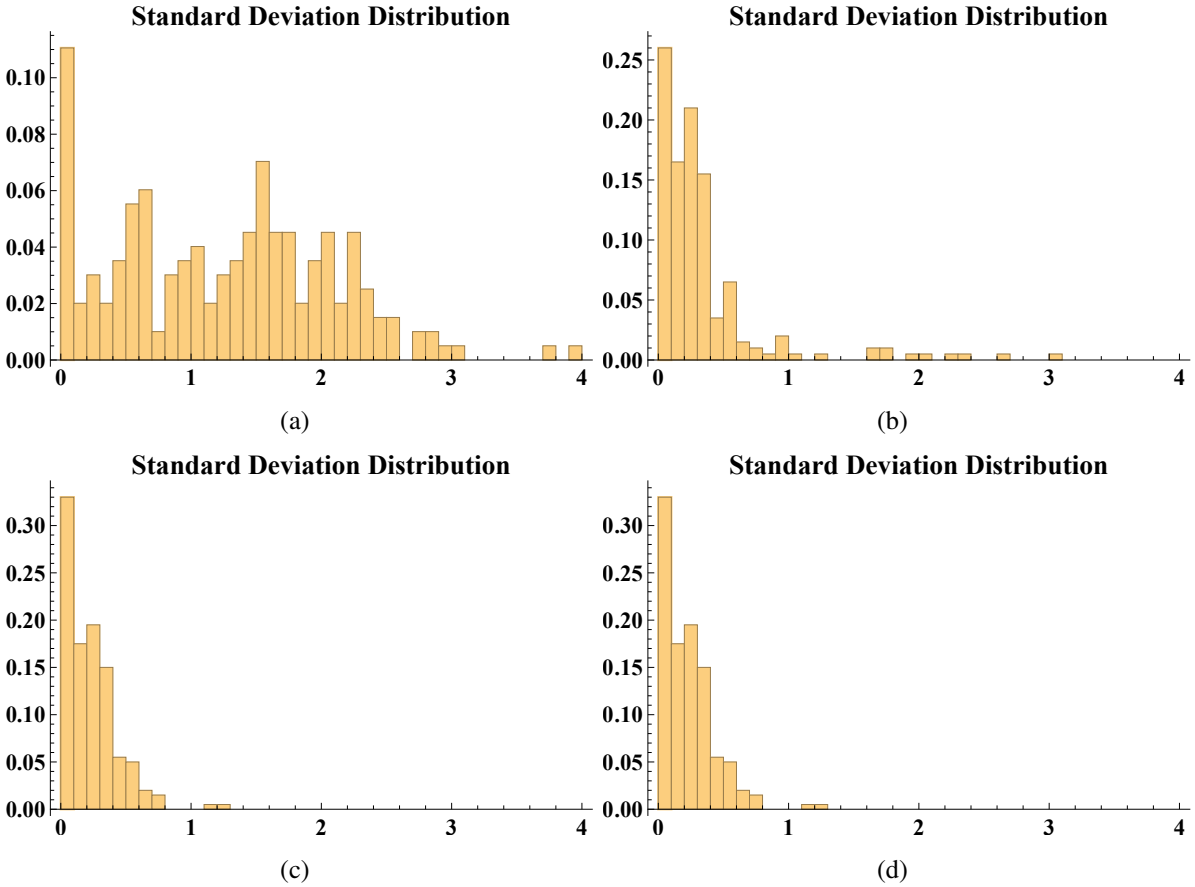


Figure S17: Standard deviation distribution of all genes within all clusters for each clustering method. (a) k-means applied to experimental data; (b) to (d) Clustering of simulated data obtained after parameter estimation using k-means clusters. (b) k-means; (c) NbC; (d) Gaussian Mixture. The horizontal axis shows the standard deviation values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

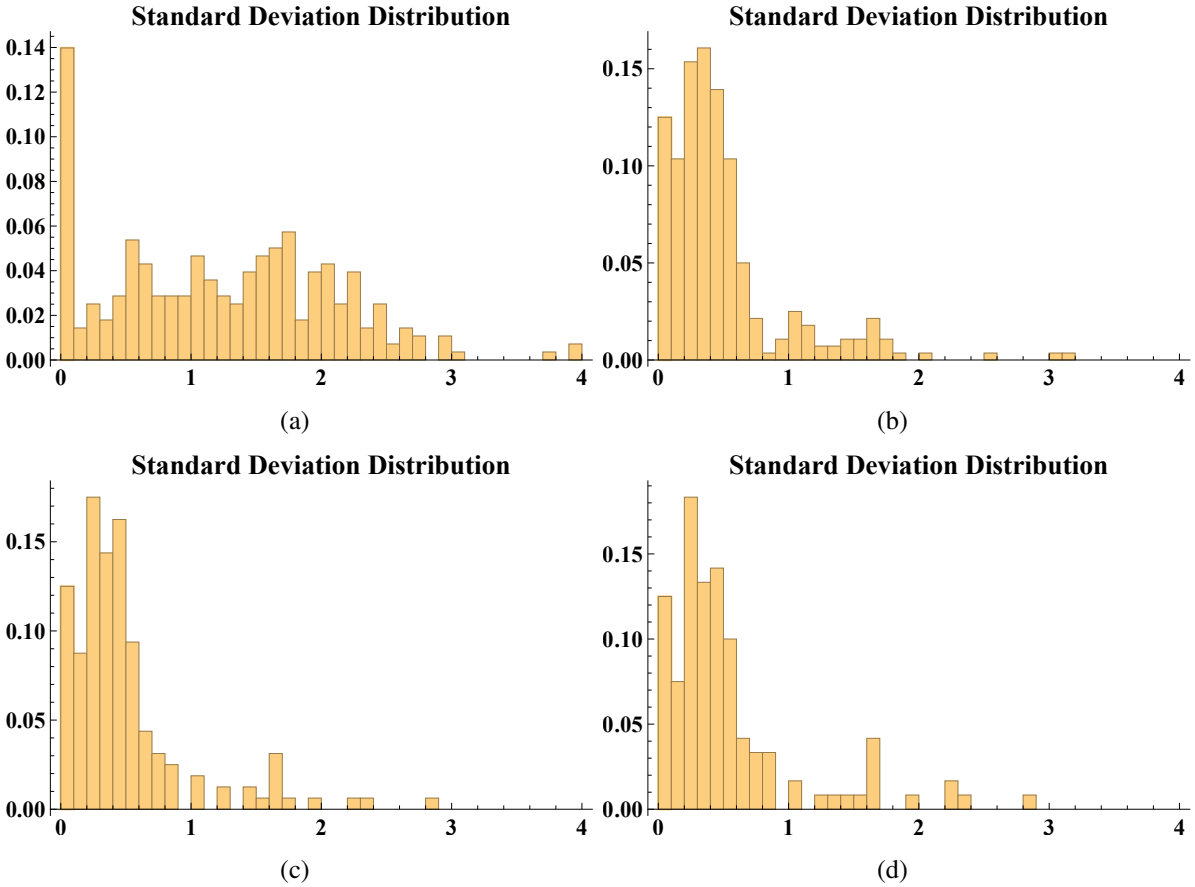


Figure S18: Standard deviation distribution of all genes within all clusters for each clustering method. (a) NbC applied to experimental data; (b) to (d) Clustering of simulated data obtained after parameter estimation using NbC clusters. (b) k-means; (c) NbC; (d) Gaussian Mixture. The horizontal axis shows the standard deviation values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

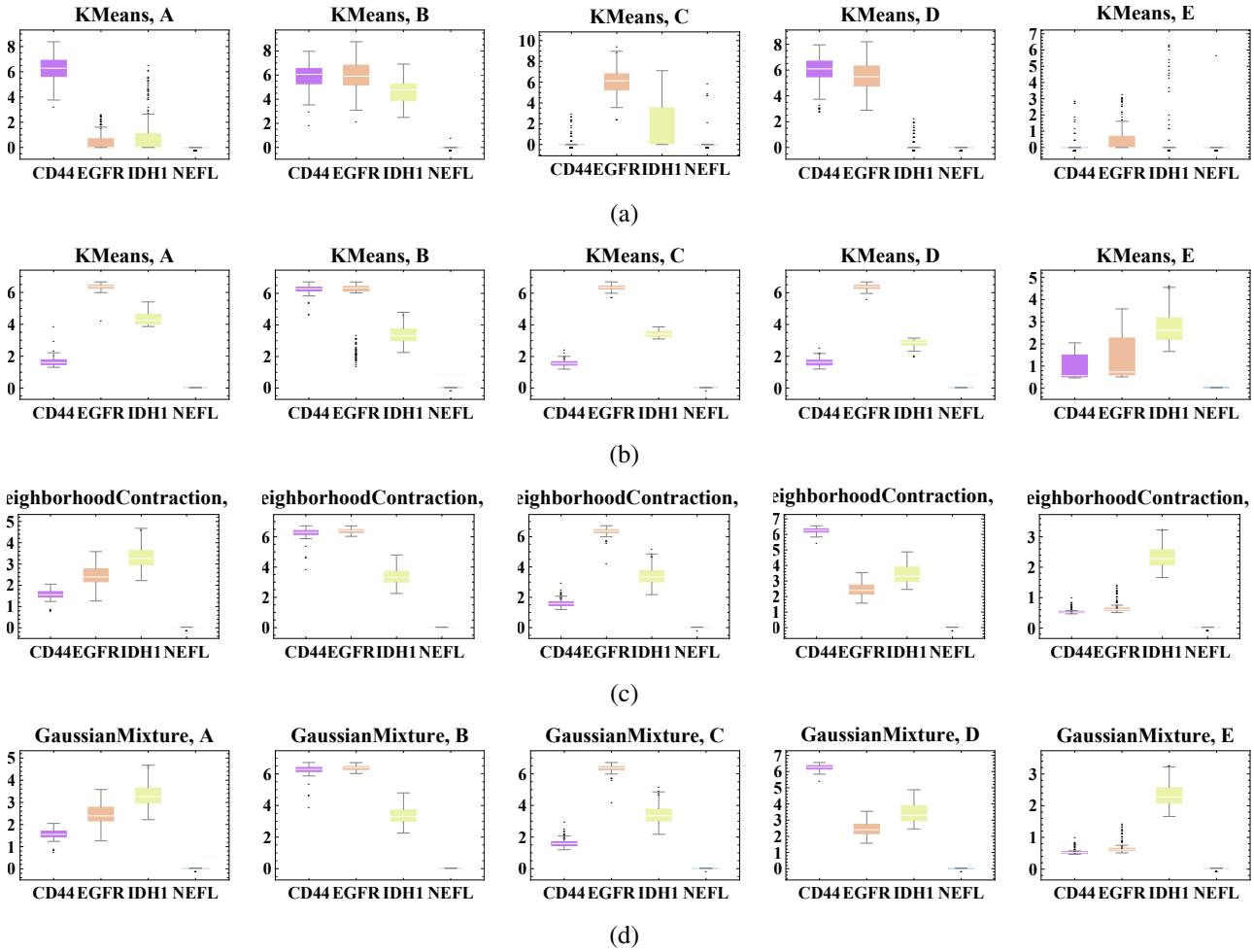


Figure S19: Boxplots of gene expression distributions for different clustering methods. (a) k-means applied to experimental data; (b) k-means applied to simulated data after parameter estimation using k-means clusters; (c) NbC applied to simulated data after parameter estimation using k-means clusters; (d) Gaussian Mixture applied to simulated data after parameter estimation using k-means clusters. The horizontal axis displays each considered marker gene, while the vertical axis represents the respective expression values.

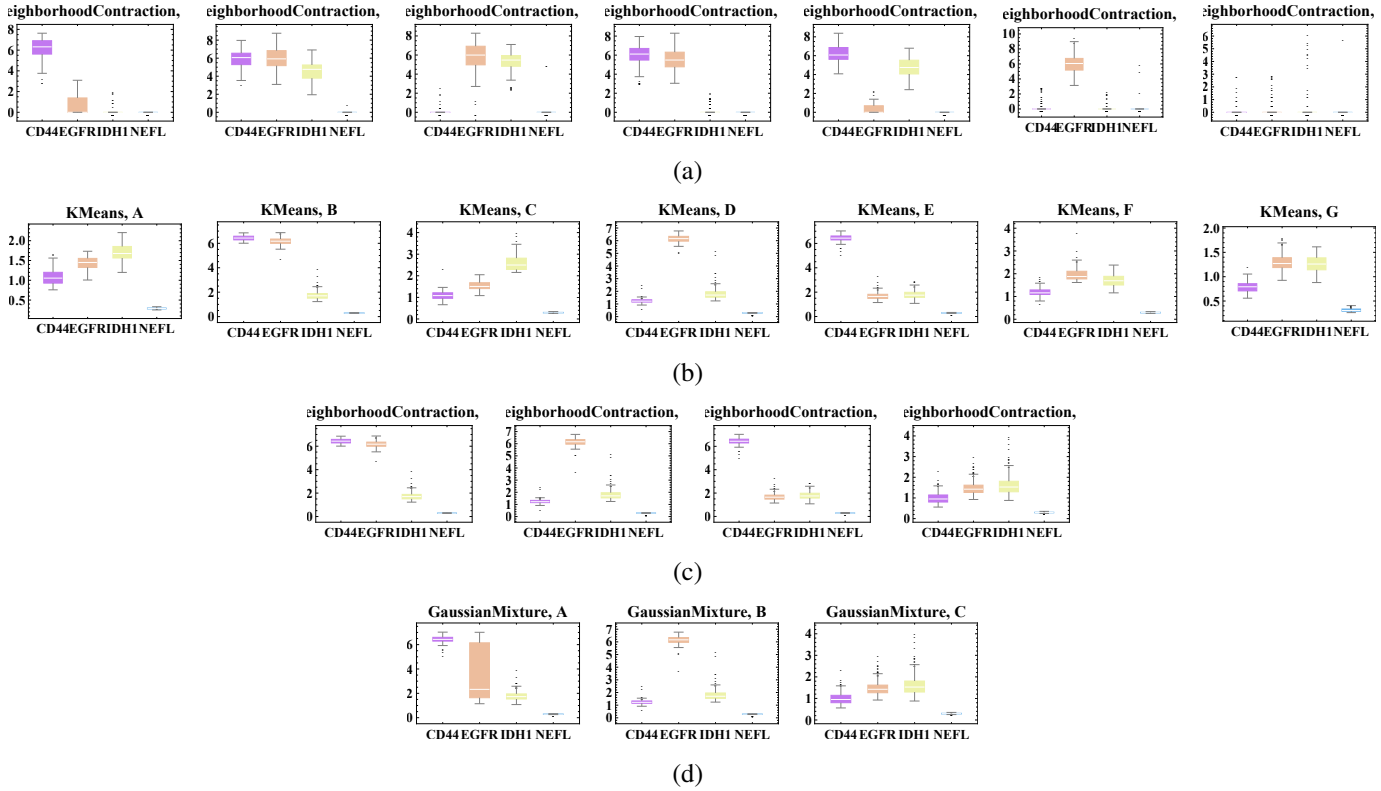


Figure S20: Boxplots of gene expression distributions for different clustering methods. (a) NbC applied to experimental data; (b) k-means applied to simulated data after parameter estimation using NbC clusters; (c) NbC applied to simulated data after parameter estimation using NbC clusters; (d) Gaussian Mixture applied to simulated data after parameter estimation using NbC clusters. The horizontal axis displays each considered marker gene, while the vertical axis represents the respective expression values.

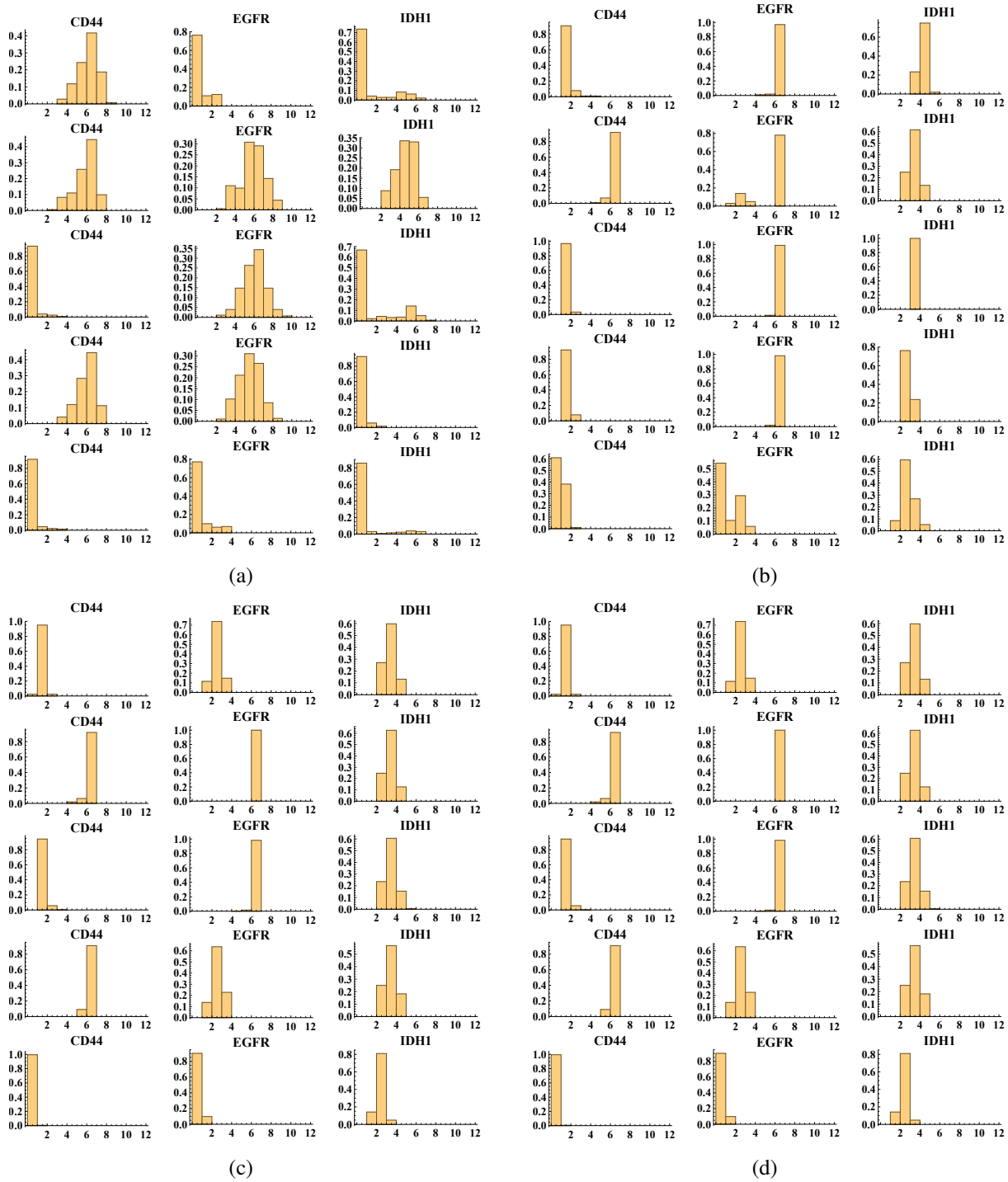


Figure S21: Gene expression distribution of markers for each cluster in different clustering methods. (a) k-means applied to experimental data; (b) k-means applied to simulated data after parameter estimation using k-means clusters; (c) NbC applied to simulated data after parameter estimation using k-means clusters; (d) Gaussian Mixture applied to simulated data after parameter estimation using k-means clusters. Each horizontal line shows the marker distribution for each respective cluster. For each case, the cluster identification ranges from A to E. The horizontal axis shows the expression values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

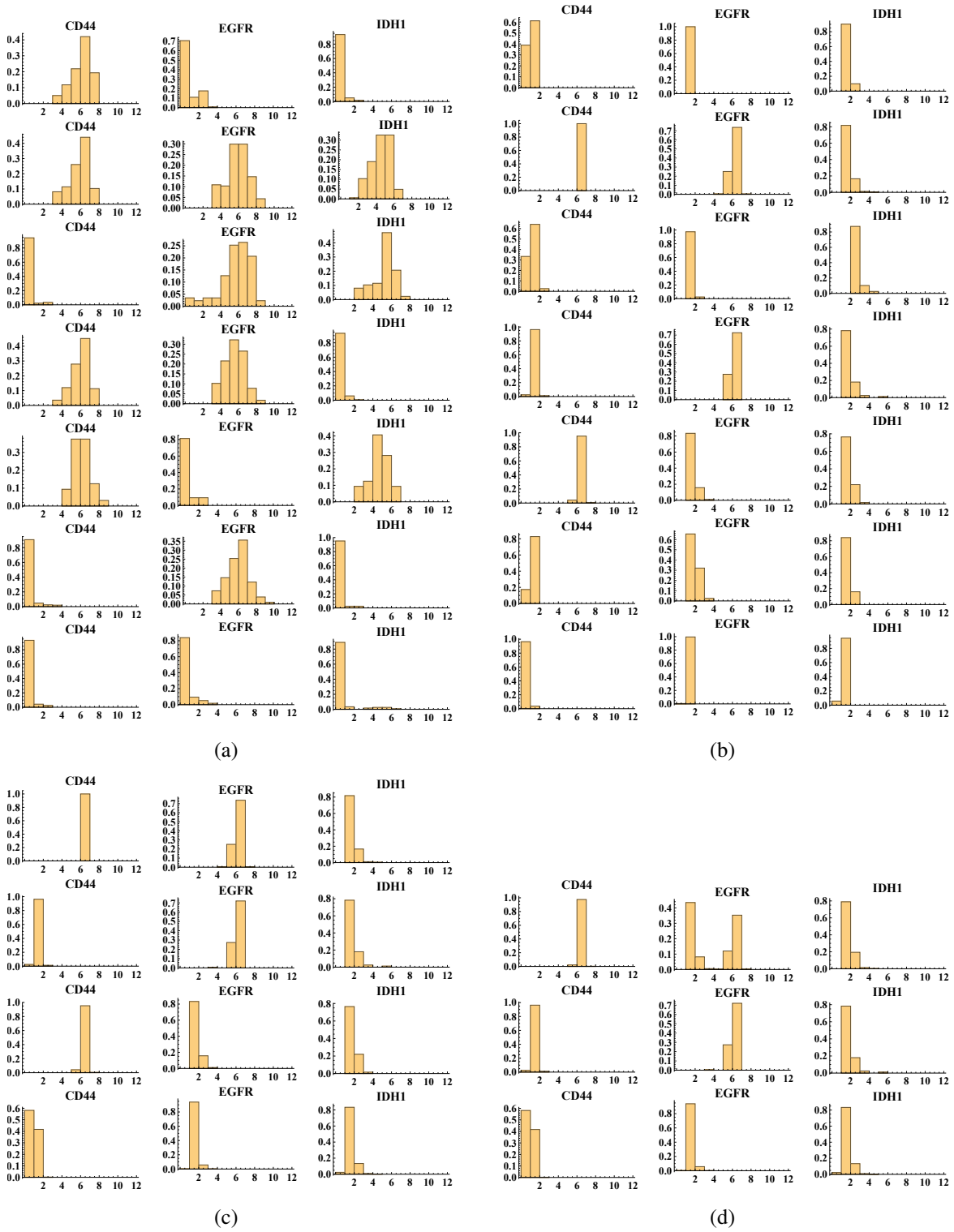


Figure S22: Gene expression distribution of markers for each cluster in different clustering methods. (a) NbC applied to experimental data; (b) k-means applied to simulated data after parameter estimation using NbC clusters; (c) NbC applied to simulated data after parameter estimation using NbC clusters; (d) Gaussian Mixture applied to simulated data after parameter estimation using NbC clusters. Each horizontal line shows the marker distribution for each respective cluster. For each case, the identification of the cluster ranges from A to its respective letter. The horizontal axis shows the expression values, while the vertical axis represents normalized counts or frequencies to indicate probabilities.

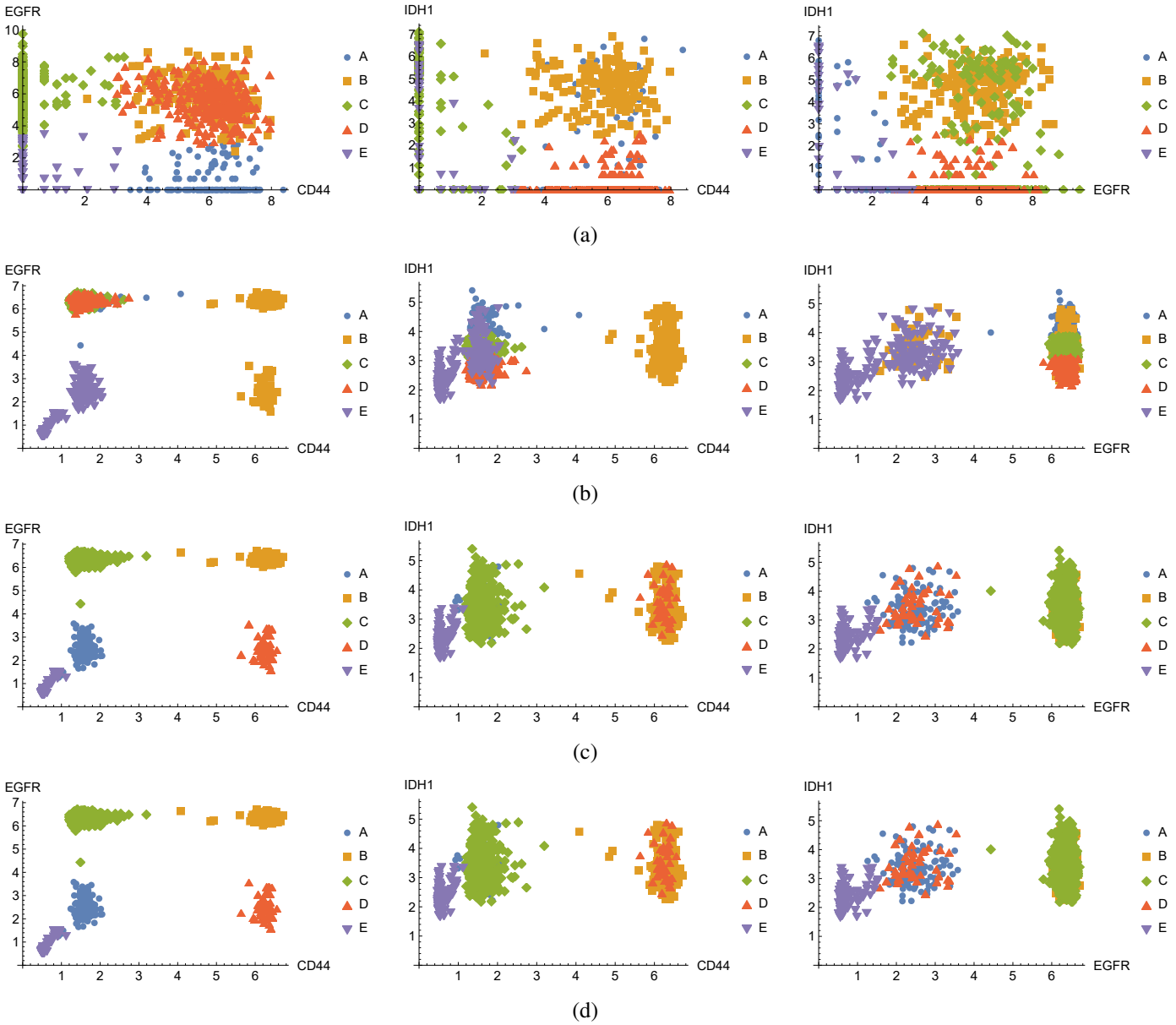


Figure S23: Scatter plots for combinations of marker gene expressions, excluding NEFL due to predominantly zero values. (a) Experimental data with five k-means clusters; (b) to (d) Simulated data after parameter estimation using k-means clusters. (b) Clustering with k-means; (c) Clustering with NbC; (d) Clustering with Gaussian Mixture. Each scatter plot represents the relationship between the expressions of different marker genes in each clustering method. The horizontal and the vertical axes show the expression values of each marker gene, while the different colors and shapes represent their corresponding cluster.

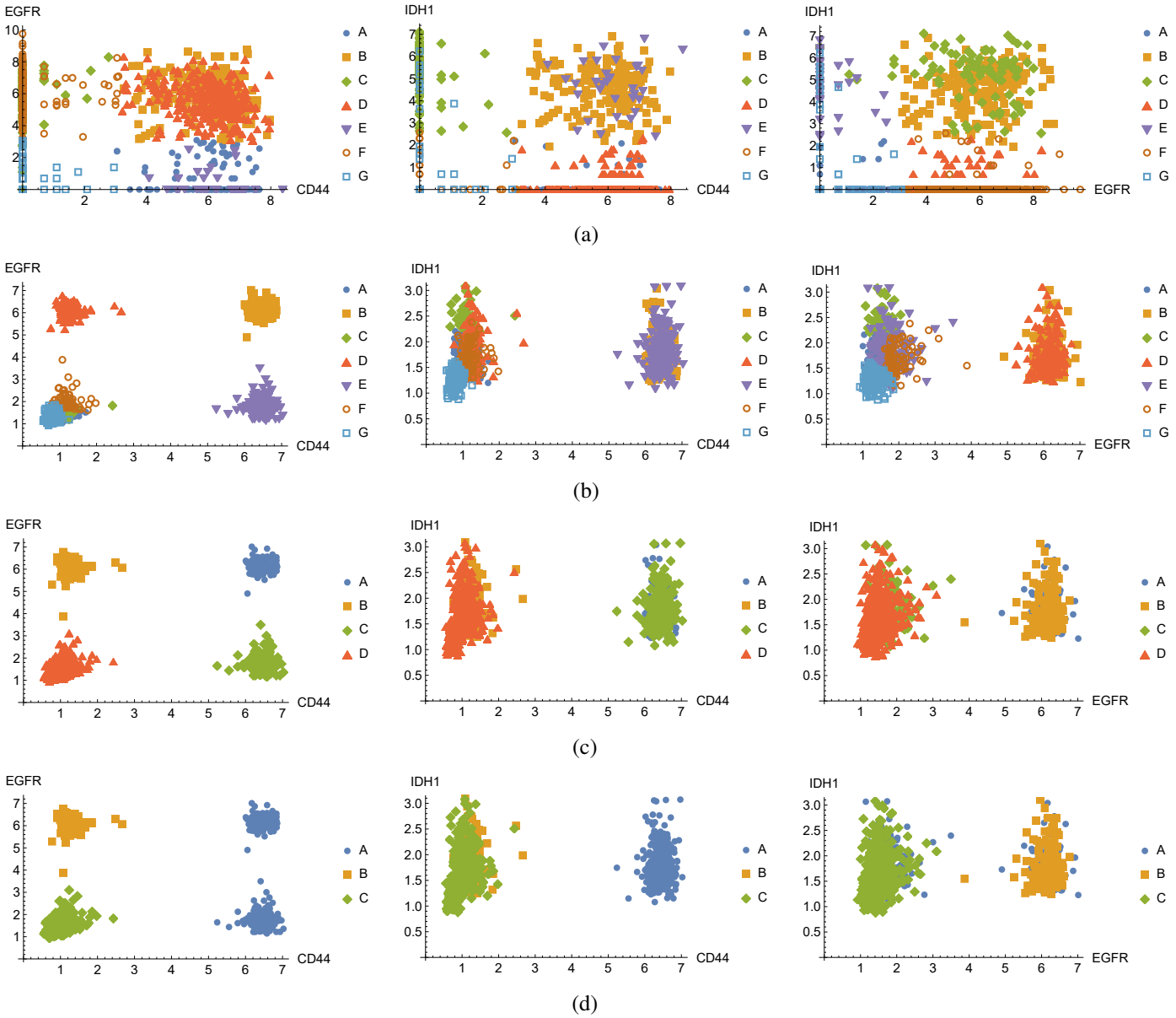


Figure S24: Scatter plots for combinations of marker gene expressions, excluding NEFL due to predominantly zero values. (a) Experimental data with seven NbC clusters; (b) to (d) Simulated data after parameter estimation using NbC clusters. (b) Clustering with k-means; (c) Clustering with NbC; (d) Clustering with Gaussian Mixture. Each scatter plot represents the relationship between the expressions of different marker genes in each clustering method. The horizontal and the vertical axes show the expression values of each marker gene, while the different colors and shapes represent their corresponding cluster.

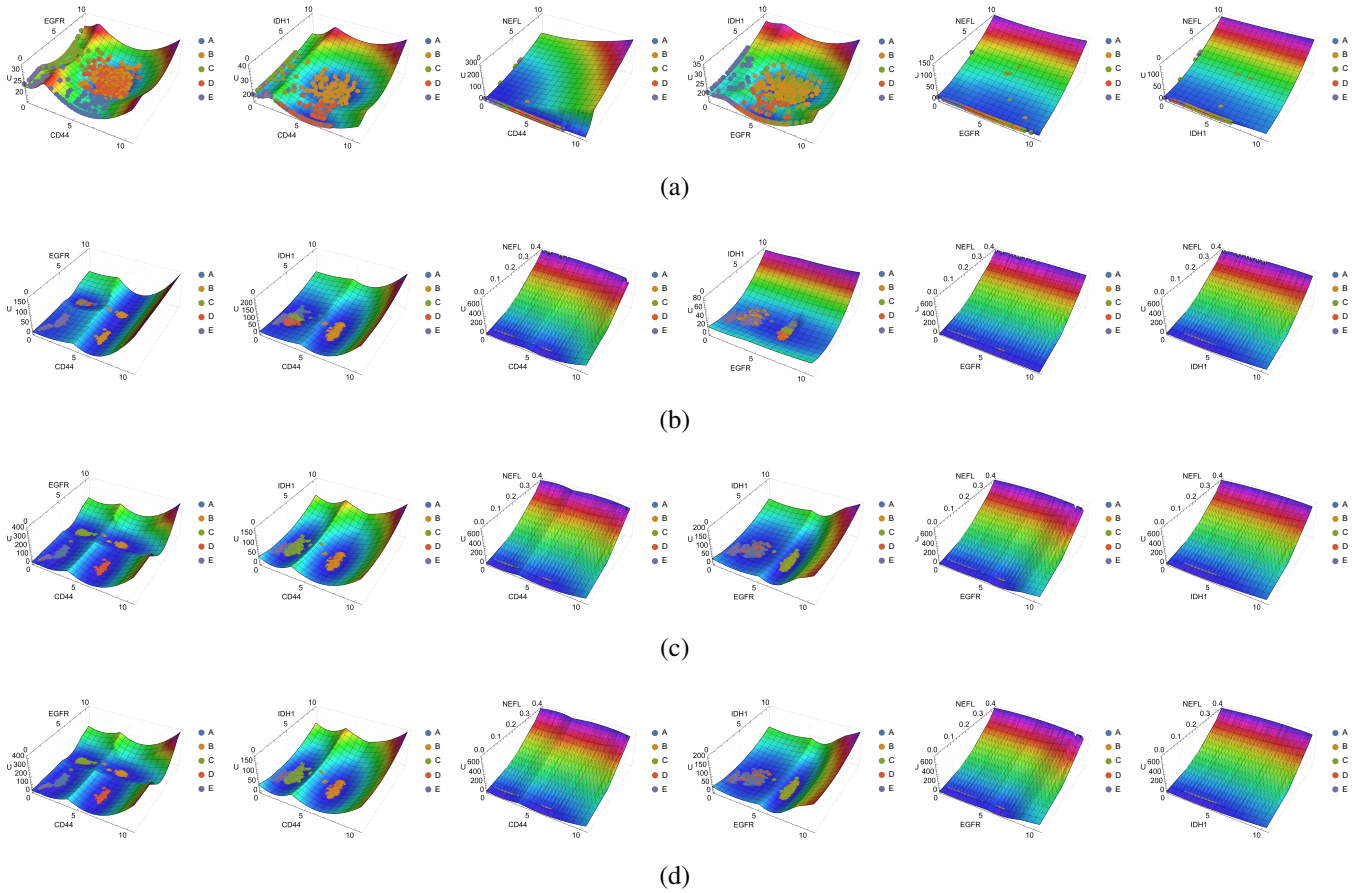


Figure S25: (Epi)genetic landscapes for experimental and simulated data, with experimental and simulated points overlaid for compatibility visualization. (a) Landscape for experimental data with k-means clusters; (b) Landscape for simulated data after parameter estimation and clustering with k-means; (c) Landscape for simulated data after parameter estimation and clustering with NbC; (d) Landscape for simulated data after parameter estimation and clustering with Gaussian Mixture. These landscapes visually represent the compatibility between experimental and simulated data in each clustering method. The horizontal axes show the expression values of each marker gene, while the vertical axis represents the values of the landscape. The different colors and shapes correspond to the respective clusters.

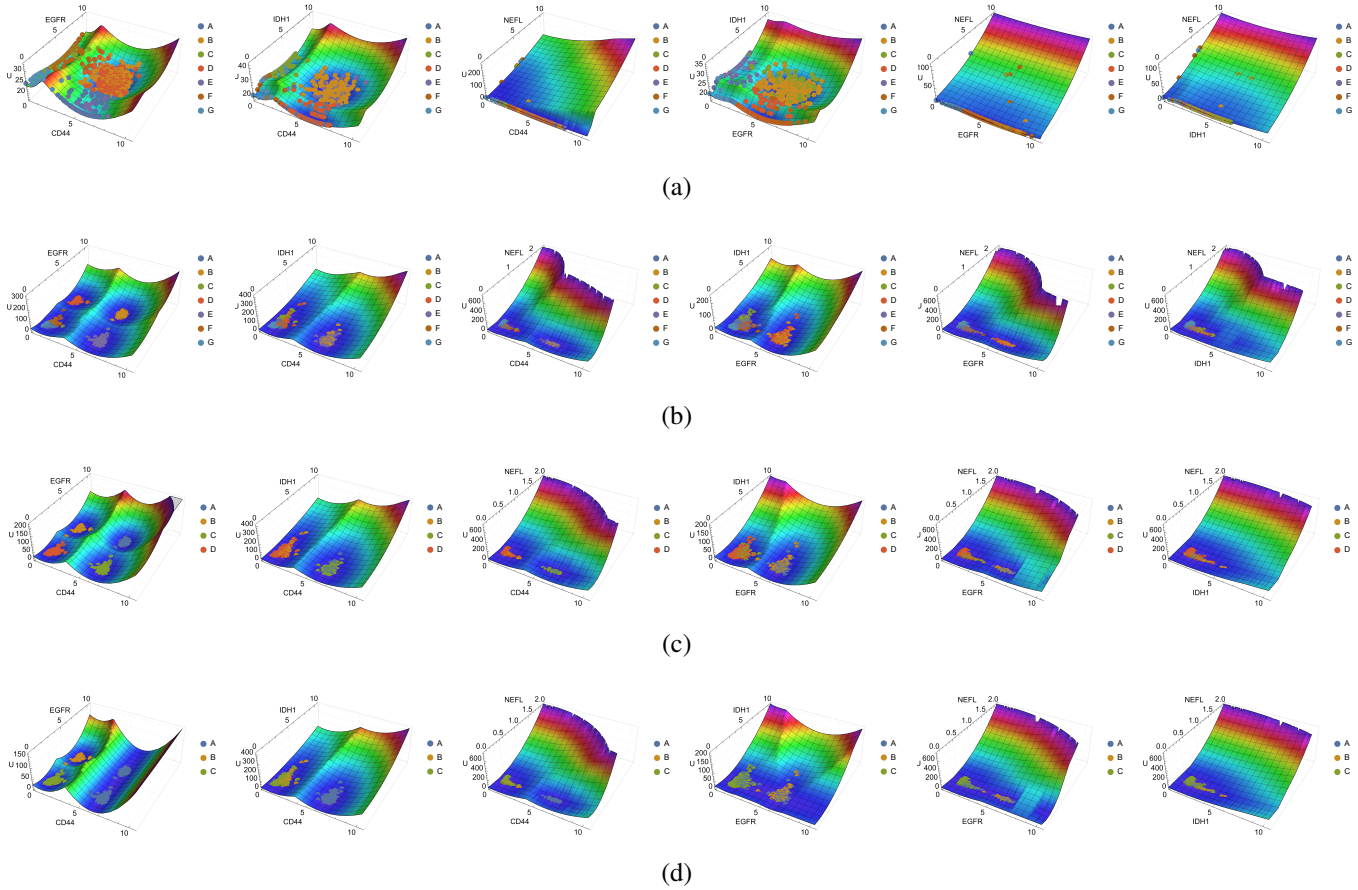


Figure S26: (Epi)genetic landscapes for experimental and simulated data, with experimental and simulated points overlaid for compatibility visualization. (a) Landscape for experimental data with NbC clusters; (b) Landscape for simulated data after parameter estimation and clustering with k-means; (c) Landscape for simulated data after parameter estimation and clustering with NbC; (d) Landscape for simulated data after parameter estimation and clustering with Gaussian Mixture. These landscapes visually represent the compatibility between experimental and simulated data in each clustering method. The horizontal axes show the expression values of each marker gene, while the vertical axis represents the values of the landscape. The different colors and shapes correspond to the respective clusters.

Figure S27: Trajectory plots for each basin, showcasing the system's dynamics. Each panel represents a different gene, with five trajectories, one for each cluster. The horizontal axis represents time, and the vertical axis shows the expression level of the gene. The trajectories illustrate the time evolution from initial conditions as the centroid of each cluster. These visualizations provide insights into the internal dynamics of each basin and help identify potential transitions between clusters.

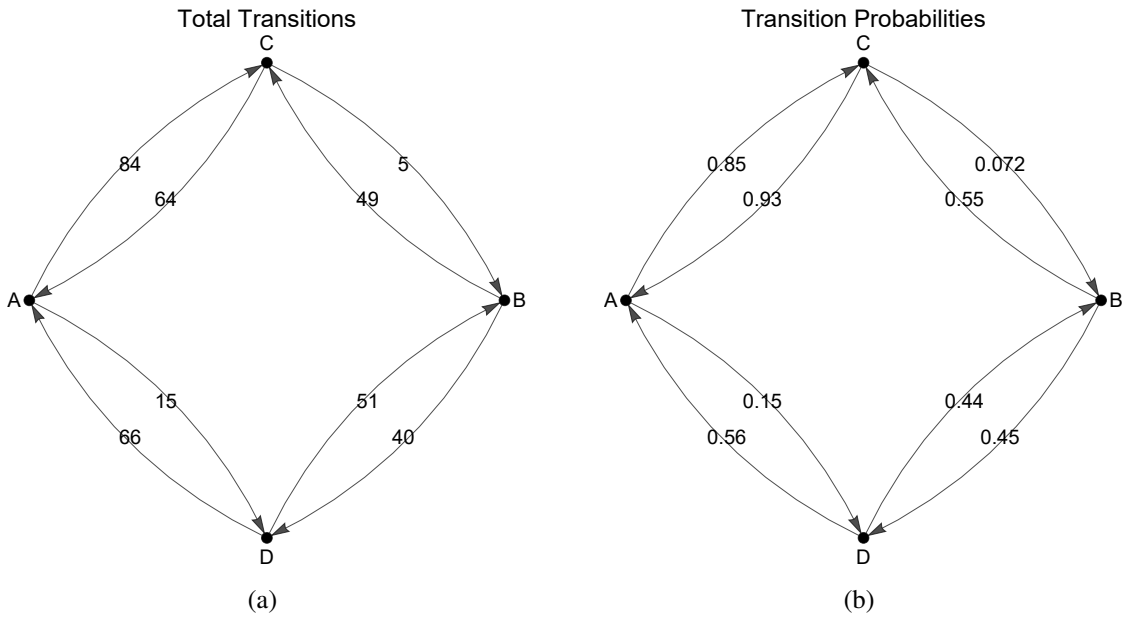
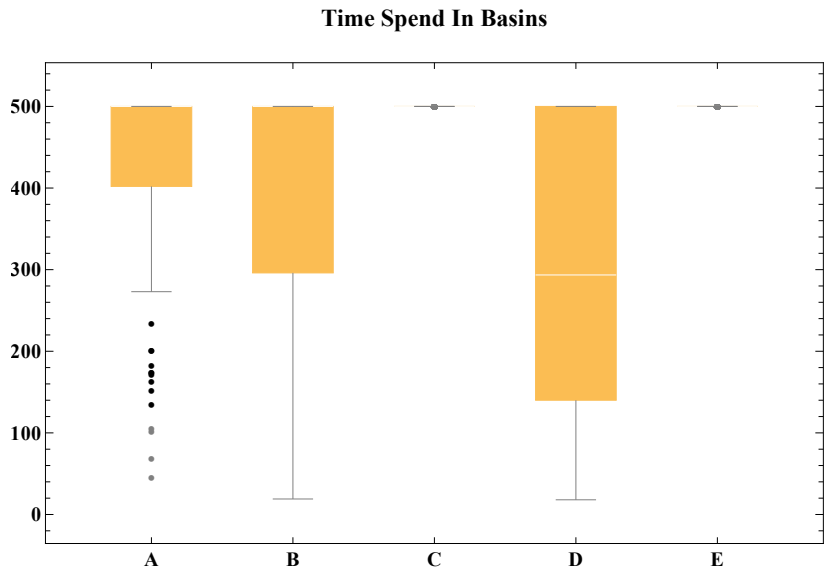
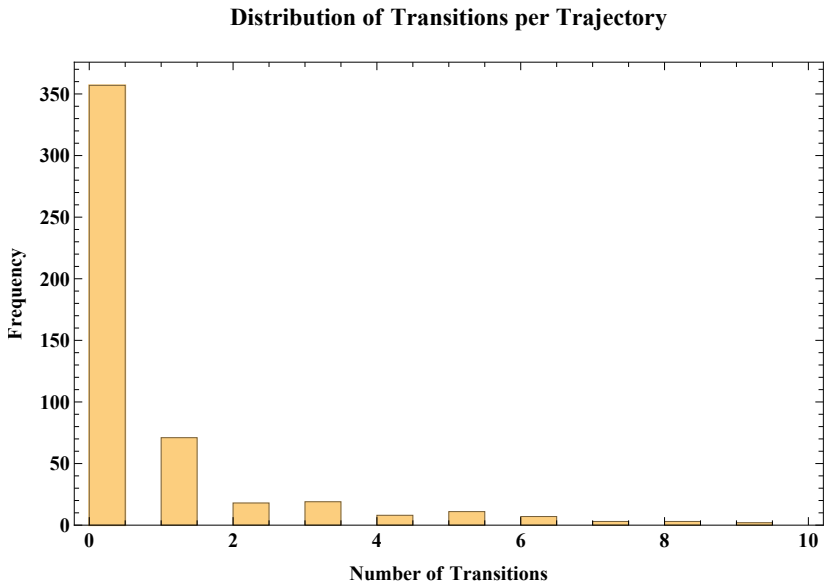


Figure S28: Transition graphs illustrating the connections between different basins. (a) Graph representation of transitions between basins, with nodes representing basins and edges representing observed transitions. (b) The same transition graph as in (a), but with edge weights representing the probabilities of transitions between basins. These visualizations provide insights into the potential transition pathways and their relative likelihood within the system.



(a)



(b)

Figure S29: Analysis of time spent in basins before a transition and the frequency of transitions per trajectory. (a) Box plots of the distribution of time spent in each basin across all trajectories before they present a transition, providing insights into the relative stability of different basins. The vertical axis represents the time spent in the basin, while the horizontal axis the correspondent basins. (b) Histogram showing the frequency of transitions between basins in each trajectory, highlighting that most trajectories do not present any transition, and those that do tend to have a small number of transitions. The vertical axis shows the frequency of each number of transitions per trajectory, while the horizontal axis the number of transitions per trajectory. These plots help to assess the inter-basin dynamics.

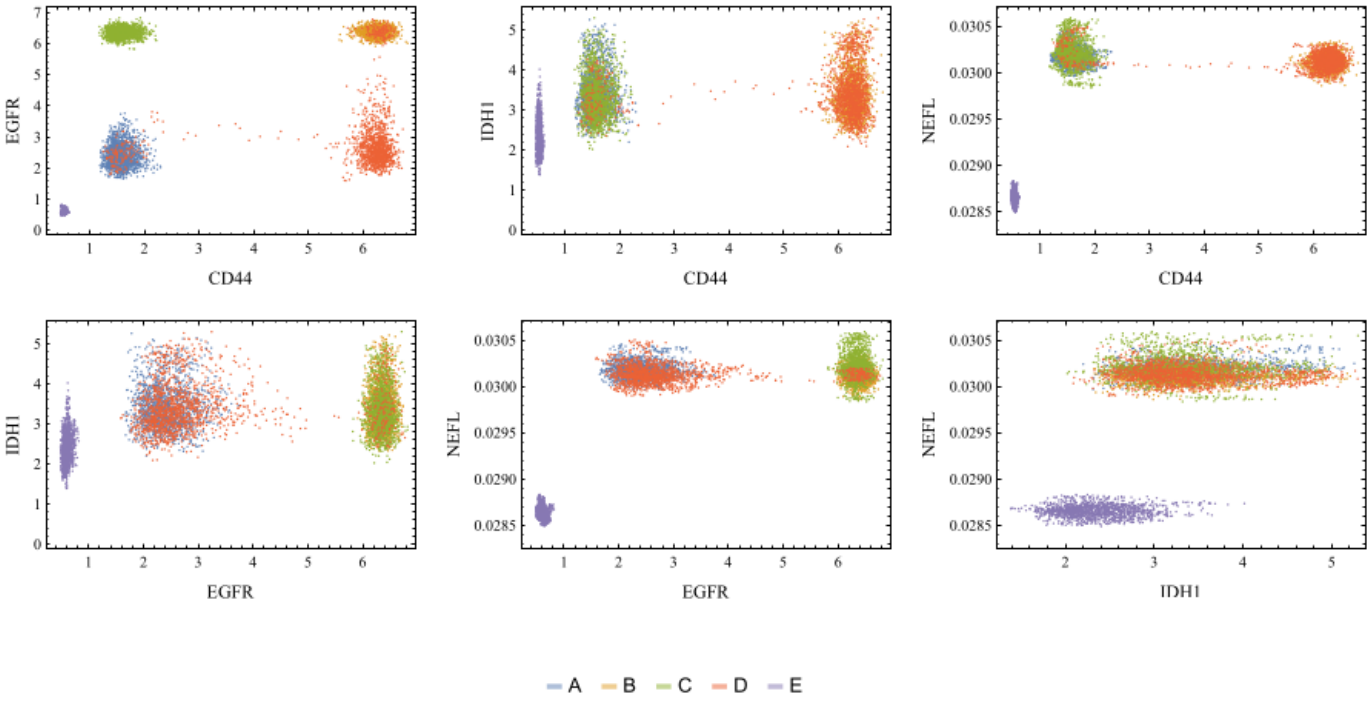


Figure S30: Two-dimensional visualization of the trajectories in the markers combination space. The horizontal and vertical axes display the expression values of each marker gene. Each point represents a single time step of three considered trajectories. Each color/letter indicates its respective basin. This figure helps to illustrate the trajectories' paths and the system's dynamics in a simplified 2D space.

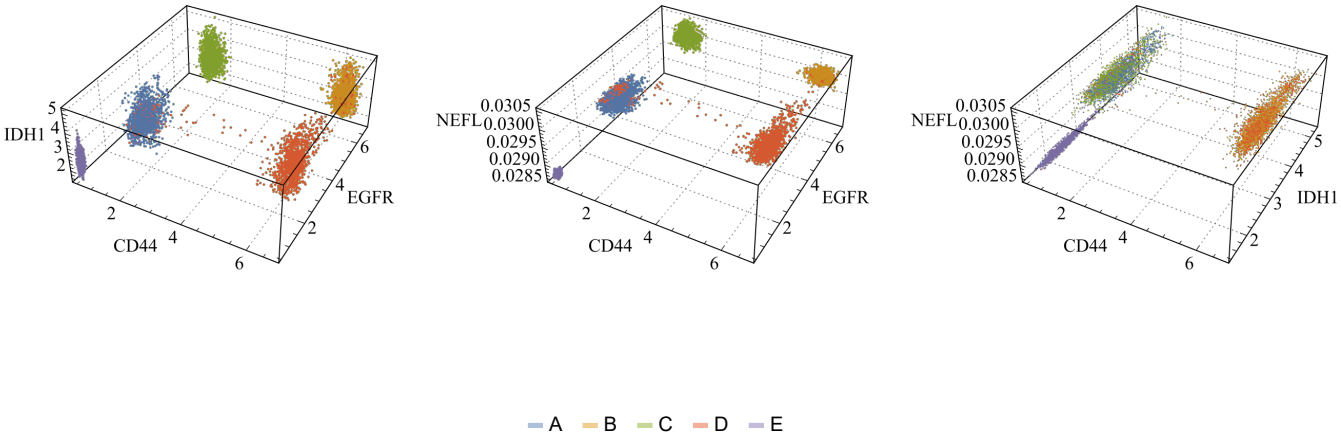


Figure S31: Three-dimensional visualization of the trajectories in the markers combination space. All axes display the expression values of each marker gene. Each point represents a single time step of three considered trajectories. Each color/letter indicates its respective basin. This figure offers a more detailed view of the system's dynamics and trajectories in a 3D space.

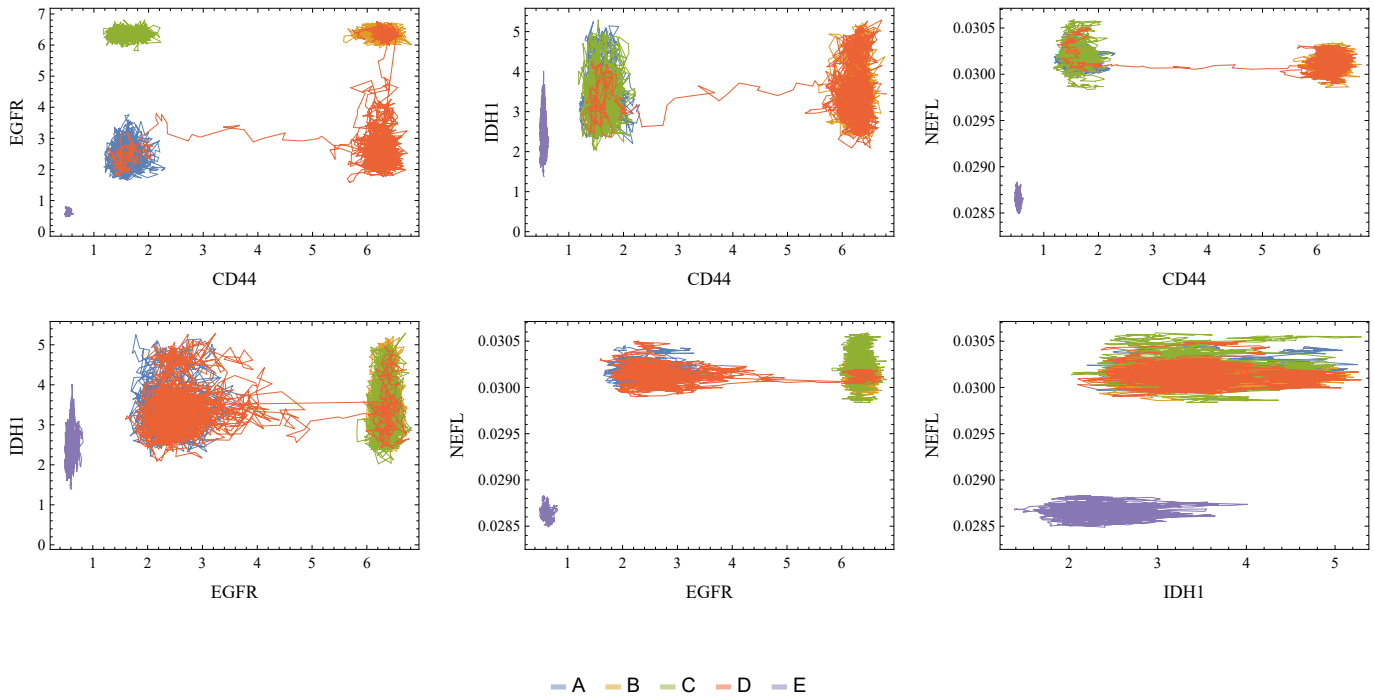


Figure S32: Two-dimensional visualization of full trajectories in the markers combination space. The horizontal and vertical axes display the expression values of each marker gene. Each line represents an entire time of the three considered trajectories. Each color/letter indicates its respective basin. This figure provides an overview of the paths and dynamics of the system's trajectories in a 2D space.

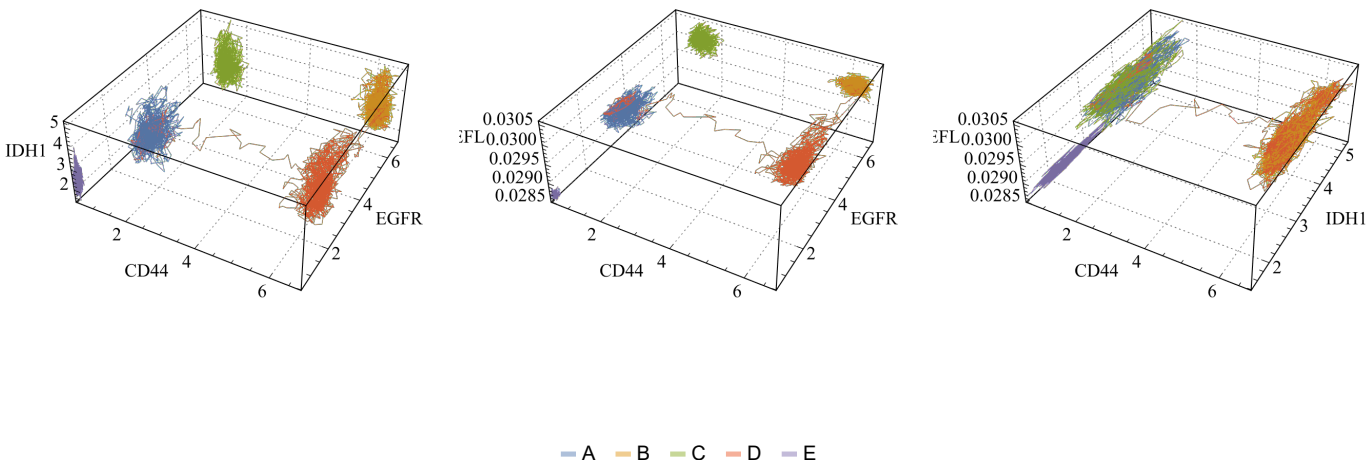
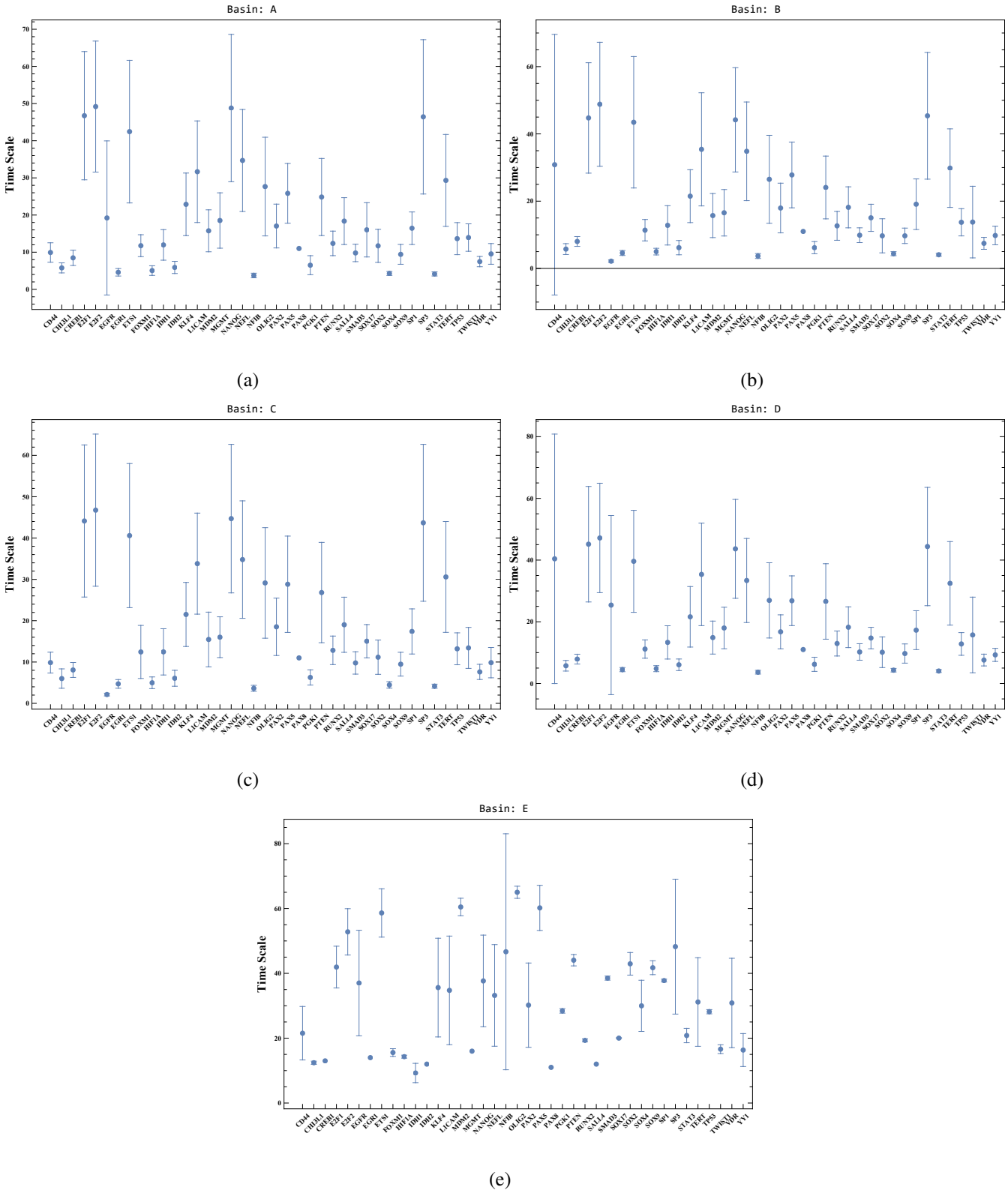


Figure S33: Three-dimensional visualization of full trajectories in the markers combination space. All axes display the expression values of each marker gene. Each line represents an entire time of the three considered trajectories. Each color/letter indicates its respective basin. This figure offers a comprehensive view of the system's dynamics and trajectories in a 3D space.

Figure S34: Autocorrelation analysis of time series data for different genes, basins, and repetitions. (a) to (d) Autocorrelation plots for four representative genes, illustrating the dependence structure of the time series data. Each pair of plots within (a) to (d) includes a time series plot (left) with the horizontal axis representing time and the vertical axis representing expression values and an autocorrelation plot (right) with the horizontal axis representing time lags and the vertical axis representing autocorrelation values. This figure helps to assess the temporal dependence of gene expressions and the potential impact of time series structure on the autocorrelation analysis.

Figure S35: Distribution of time scales for different genes and their basins. The horizontal axis represents the cluster labels, and the vertical axis represents the time scale values, defined as the minimum time lag at which the autocorrelation falls below e^{-1}. The figure presents the distribution of time scales for each gene, with error bars indicating the standard deviation within all repetitions. This visualization helps to understand the characteristic time scales within and across different genes, providing insights into the internal dynamics and possible transition behaviors of the system.



(a) (b) (c) (d) (e)

Figure S36: Time scale distributions for different genes within their basins. The horizontal axis represents the gene labels, and the vertical axis represents the time scale values, defined as the minimum time lag at which the autocorrelation falls below e^{-1}. The figure emphasizes the differences within each basin, showcasing the distinct characteristics of each gene's time-scale distribution across different basins.

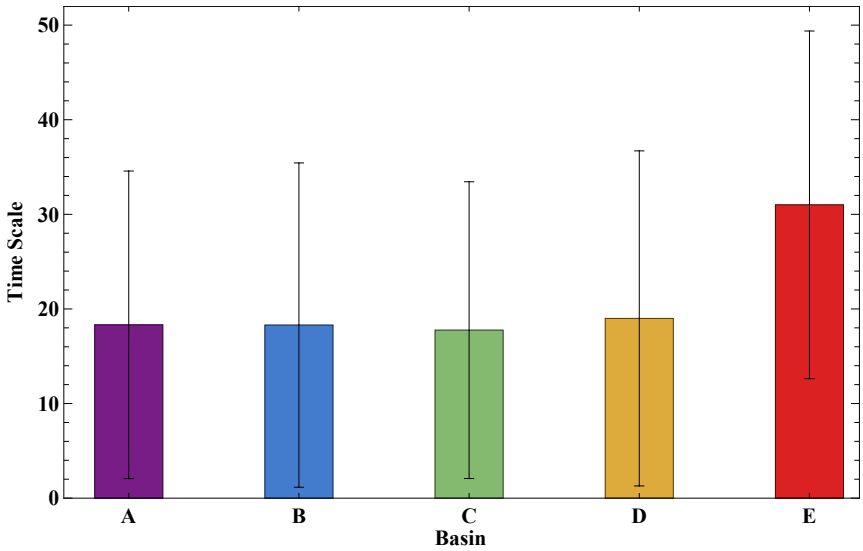


Figure S37: Final average time scales of each basin across all genes and repetitions. The horizontal axis represents the cluster labels, and the vertical axis represents each basin's final average time scale values.

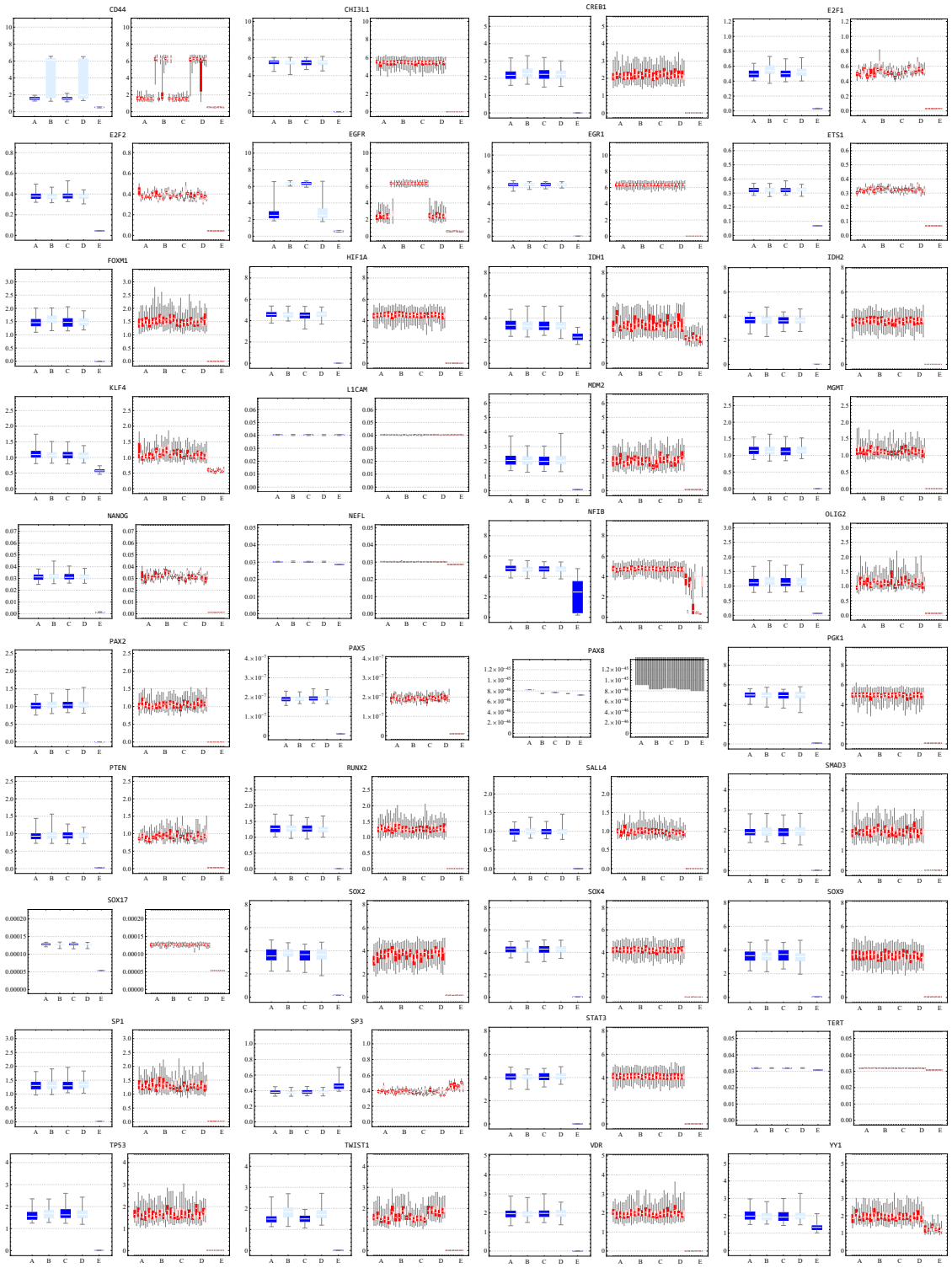


Figure S38: Comparison of time and sample averages for various clusters and genes using boxplots. The horizontal axis represents the cluster labels, and the vertical axis shows the expression values. The figure displays two sets of plots: the left set shows the sample average of 100 samples at the final time interval, while the right set represents the time average considering 10 trajectories from time 30 to 50 (steps 300 to 500). The boxplots illustrate the distribution of expression values within each cluster, with the box representing the interquartile range (IQR), the line inside the box showing the median, and the whiskers extending to the minimum and maximum data points within 1.5 times the IQR. This comparison helps to assess the compatibility between time and sample averages, supporting the internal ergodicity assumption and providing insights into potential discrepancies between different basins and genes.