International Journal of
Molecular Sciences

MDPI

Article

Unveiling the Dynamics behind Glioblastoma Multiforme Single-Cell Data Heterogeneity

Marcos Guilherme Vieira Junior [1], Adriano Maurício de Almeida Côrtes [2,3], Flávia Raquel Gonçalves Carneiro [4,5,6], Nicolas Carels [7,*] and Fabrício Alves Barbosa da Silva [8,*]

1 Graduate Program in Computational and Systems Biology, Oswaldo Cruz Institute (IOC), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro 21040-900, Brazil
2 Department of Applied Mathematics, Institute of Mathematics, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro 21941-909, Brazil; adricortes@cos.ufrj.br
3 Systems Engineering and Computer Science Program, Coordination of Postgraduate Programs in Engineering (COPPE), Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro 21941-972, Brazil
4 Center of Technological Development in Health (CDTS), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro 21040-361, Brazil; flavia.carneiro@fiocruz.br
5 Laboratório Interdisciplinar de Pesquisas Médicas, Oswaldo Cruz Institute (IOC), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro 21040-900, Brazil
6 Program of Immunology and Tumor Biology, Brazilian National Cancer Institute (INCA), Rio de Janeiro 20231-050, Brazil
7 Laboratory of Biological System Modeling, Center of Technological Development in Health (CDTS), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro 21040-361, Brazil
8 Scientific Computing Program, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro 21041-222, Brazil
* Correspondence: nicolas.carels@fiocruz.br (N.C.); fabricio.silva@fiocruz.br (F.A.B.d.S.); Tel.: +55-21-3882-9234 (N.C.); +55-21-3836-1114 (F.A.B.d.S.)

Abstract: Glioblastoma Multiforme is a brain tumor distinguished by its aggressiveness. We suggested that this aggressiveness leads single-cell RNA-sequence data (scRNA-seq) to span a representative portion of the cancer attractors domain. This conjecture allowed us to interpret the scRNA-seq heterogeneity as reflecting a representative trajectory within the attractor's domain. We considered factors such as genomic instability to characterize the cancer dynamics through stochastic fixed points. The fixed points were derived from centroids obtained through various clustering methods to verify our method sensitivity. This methodological foundation is based upon sample and time average equivalence, assigning an interpretative value to the data cluster centroids and supporting parameters estimation. We used stochastic simulations to reproduce the dynamics, and our results showed an alignment between experimental and simulated dataset centroids. We also computed the Waddington landscape, which provided a visual framework for validating the centroids and standard deviations as characterizations of cancer attractors. Additionally, we examined the stability and transitions between attractors and revealed a potential interplay between subtypes. These transitions might be related to cancer recurrence and progression, connecting the molecular mechanisms of cancer heterogeneity with statistical properties of gene expression dynamics. Our work advances the modeling of gene expression dynamics and paves the way for personalized therapeutic interventions.

Keywords: Glioblastoma Multiforme; epigenetic landscape; parameter sets estimation; single-cell RNA sequencing; heterogeneity; cancer attractors; gene regulatory network dynamics

Citation: Vieira Junior, M.G.; Côrtes, A.M.d.A.; Carneiro, F.R.G.; Carels, N.; Silva, F.A.B.d. Unveiling the Dynamics behind Glioblastoma Multiforme Single-Cell Data Heterogeneity. *Int. J. Mol. Sci.* **2024**, *25*, 4894. https://doi.org/10.3390/ijms25094894

Academic Editor: Raffaella Lazzarini

Received: 8 March 2024
Revised: 2 April 2024
Accepted: 3 April 2024
Published: 30 April 2024

1. Introduction

Notwithstanding the significant advancements in understanding and therapeutics, cancer continues to be a predominant global cause of mortality [1]. Glioblastoma multiforme (GBM; Appendix B) stands as the most common and aggressive brain tumor, characterized by an average survival time of 15 months and a roughly 10% probability of achieving a 5-year overall survival [2,3]. Single-cell RNA sequencing (scRNA-seq) has spotlighted the

pronounced heterogeneity inherent in GBM [4,5]. Intriguingly, this level of complexity in scRNA-seq data is not exclusive to GBM. Many cancers manifest similar intricate patterns in their sequencing data [6,7], underpinning a broader challenge in oncology research. With such heterogeneity possibly driving the aggressiveness of these malignancies [8,9], pressing questions emerge regarding the underlying dynamics. Foremost among these are the interpretations of observed data distributions, their ensuing consequences, and the patterns encapsulated within the data.

Extensive research has delved into the complexities of carcinogenesis, advancing discourses on driver and passenger mutations and the profound influences of epigenetics [10–12]. In the intricate landscape of gene regulatory networks (GRN) dynamics, pivotal studies have elucidated the alignment between cell types or subtypes and stable states, often termed 'attractors' [13,14]. Concurrently, certain oscillatory cellular processes—integral to diverse functions such as circadian rhythms [15], cell cycle progression [16], and NF-κB signaling in response to inflammation [17]—are closely tied to limit cycles in GRNs. Yet, the foundational principles guiding transitions between these states and the mechanisms by which a system traverses within an attractor's domain—a notion sometimes framed as a 'cancer attractor' [18]—remain areas of active research.

The intricacies of tumor evolutionary trajectories further underscore a pressing need for understanding. As cancer progresses or counteracts therapeutic measures, the dynamic shifts in tumor subclonal architectures come to the forefront [19]. Traditional linear evolution models may inadequately capture the complexities of tumor evolution. A dominant clone proliferates in certain scenarios, producing a predominantly homogeneous tumor composition. Conversely, other situations present coexisting subclonal populations, suggesting a more branched evolutionary pattern than a purely linear trajectory [20]. In light of these challenges, the increasing availability of omics data presents an opportunity for deeper investigations. For instance, studies have identified subtypes of GBM, a fundamental leap forward in understanding the disease and a critical step for choosing the applying treatment [2,21–23]. Yet, as the horizon of our knowledge expands, questions about the evolutionary pathways of these subtypes and the dynamic interplay underpinning their classifications persist, beckoning deeper investigations [24–26].

This evolving scenario necessitates a shift in perspective. Instead of remaining anchored in reductionist perspectives, there is a pressing call to embrace a more systemic approach. This perspective provides a comprehensive view of cell type and functionality dynamics and is reinforced by studies such as [27]. This systemic thinking is present in contemporary paradigms that depict cancer as a nuanced series of events leading to a 'state disease' [28], rather than being merely the fallout of isolated mutations. In this context, we define the *state* as the biochemical milieu of a cell, signifying the evolutionary trajectory of states within a complex system. Contrasting starkly with earlier models that portrayed cancer progression as linear, this perspective revels in understanding the multifaceted dynamics of the disease within a broader, multidimensional context [13]. A vivid analogy for this conceptual shift can be traced back to C. Waddington's 1957 metaphor, where cellular differentiation is visualized as a sphere traversing an (epi)genetic landscape (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) of peaks and valleys [29], with the zeniths representing the undifferentiated phenotype.

In the Waddington (epi)genetic landscape, a differentiated state is achieved through developmental paths influenced by both intrinsic cellular factors—as the cell's historical events, such as lineage, gene expression patterns, and epigenetic modifications [30]—and by extrinsic factors such as tissue-level perturbations or environmental influences. According to this perspective, cancer might 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, the attractors are undesirably reached due to the inherent stochastic fluctuations of biological systems at the biomolecular scale [18].

Using the Waddington (epi)genetic landscape as a conceptual framework allows us to understand the extent of heterogeneity within GBM more clearly. As Waddington envisioned, cellular differentiation is deeply intertwined with the nuances of cancer progression. If we extend this framework, it is clear that the various paths and trajectories through which a cell might journey—and eventually culminate in a malignant phenotype—are likely shaped by a combination of genetic, epigenetic, and environmental forces. This understanding reinforces the need to decipher the intricate details behind single-cell data, especially since heterogeneity is both an outcome and an influencer of tumor dynamics. As we delve deeper, the landscape metaphor becomes more than just a conceptual tool; it provides a practical framework that guides our investigation of GBM. By examining how and why certain trajectories or states become prominent in GBM, we lay the groundwork for exploring the underlying factors and mechanisms that drive cellular heterogeneity and what this might signify for our broader understanding of cancer evolution.

From this exploration, it becomes evident that cellular heterogeneity in single-cell data raises crucial questions. These questions concern the interplay between epigenetic regulation, genomic stability, selective pressures, and the inherent GRN governing cancer cell behavior. A thorough investigation into these aspects offers fresh perspectives into GBM's evolving landscape. One of the primary sources of heterogeneity is genomic instability [20]. This instability, marked by a high mutation rate at the DNA level, results in a cellular milieu teeming with diversity [31]. While genomic instability significantly contributes to heterogeneity, it represents only one facet of the complexity. Genetic alterations and epigenetic mechanisms lead to diverse cellular responses, amplifying the heterogeneity and adding layers to our understanding. This way, the dynamics underlying heterogeneity are molded by additional factors adding complexity.

Selective pressures exert a significant influence on cancer heterogeneity. Rather than being merely passive, these pressures actively shape the outcome of the diversity introduced by genomic instability [32]. They curate the cellular environment, favoring the emergence of stable cellular states and modulating the dynamics within the (epi)genetic landscape. Such stable states, called 'attractors', represent characteristic cellular phenotypes. Trajectories within the phase space tend to gravitate towards these attractors, shaped by the so-called 'basins of attraction'. In light of this, the interplay between the chaos of genomic instability and the order introduced by selective pressures, as revealed through the lens of clonal evolution, offers a detailed understanding [32]. It is not merely about heterogeneity; it underscores a structured heterogeneity, reflecting a complex dynamic infused with order.

This narrative implies a representation that captures a cell's transformative journey toward a malignant state. This work proposes modeling this representation that aligns with the snapshot-like data from scRNA-seq (Figure 1). Stemming from a structured stochastic limit cycle [33] (Figure 1I), the cellular trajectory is subject to perturbations from genetic mutations, epigenetic regulations, and selective pressures. We highlight that some dimensions, potentially associated with marker genes, may already exhibit a fixed-point dynamic even in this initial state. As these forces act, the initial trajectory undergoes alterations, manifested as increased stochastic noise that broadens the oscillatory boundary, or 'stochastic tube' (Figure 1II). This amplification in noise effectively populates the state space, thereby increasing the relevance of mean values as accurate representations of the cellular dynamics. Finally, the cell adopts more erratic behavior, eventually taking on characteristics resembling a random wandering around a fixed point (Figure 1III). This transformation might occur for multiple genes, culminating in a tumor's genetic and phenotypic variability, termed intratumor heterogeneity. Importantly, this evolving subclonal architecture is dynamic, undergoing continual shifts throughout disease progression, and thus, presenting challenges for both diagnostics and therapeutic strategies.

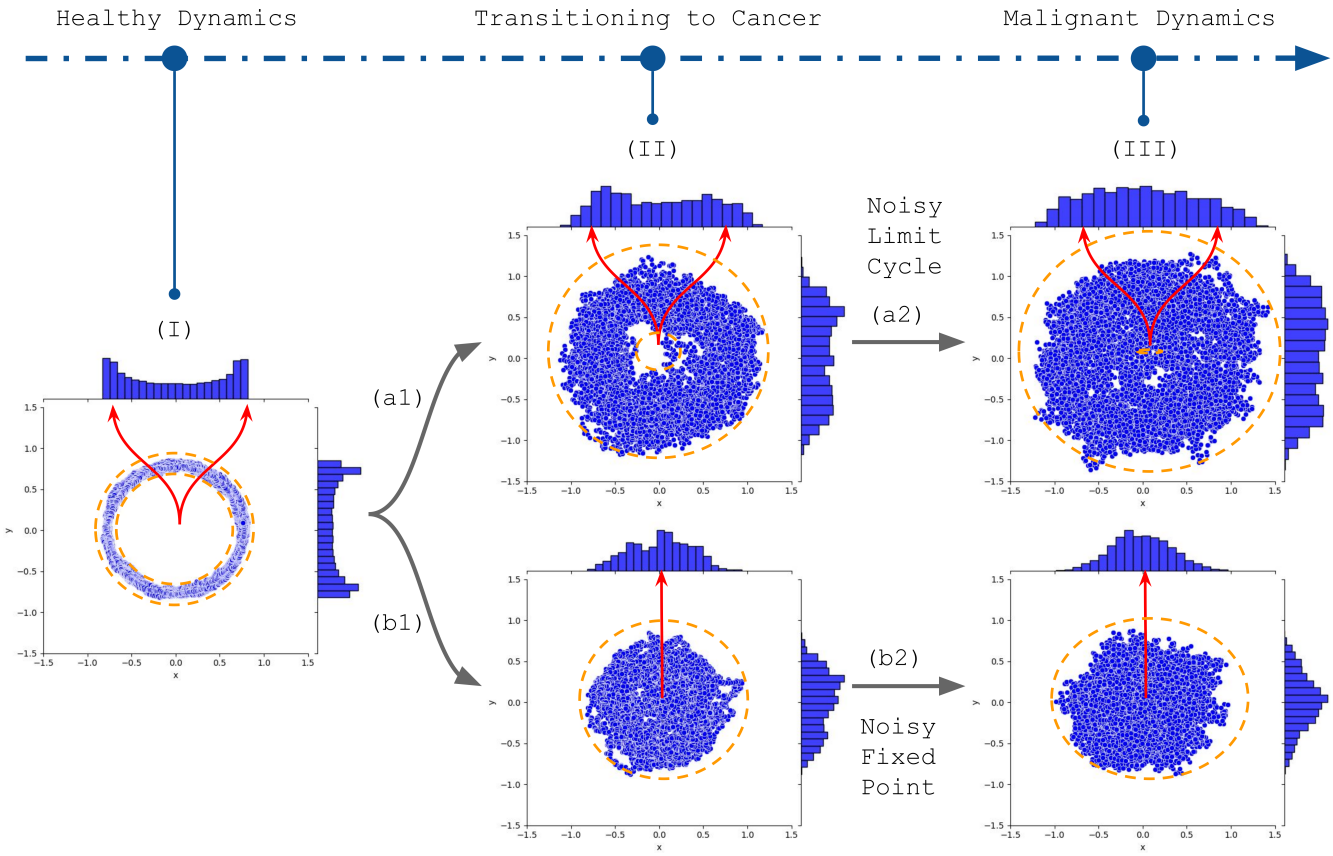


Figure 1. This image depicts our hypothesis about a cellular transition from a 'healthy' limit cycle to a malignant state, a transformation that would enlighten the dynamics underlying the single-cell RNA sequencing data heterogeneity. The 'initial state' (**I**) represents a stable cellular trajectory. However, it is important to note certain dimensions—potentially related to marker genes— might already display a fixed-point dynamic. The outcomes of genetic, epigenetic, and microenvironmental alterations are depicted through directional arrows (a1 to a2, and b1 to b2). The upward trajectory (a1 to a2) underscores an expanded oscillatory boundary called a stochastic tube. Notably, the initial limit cycle distribution (**I**) is characterized by two noncentral peaks, which witness a reduced prominence due to augmented stochastic fluctuations (**II**), resulting in an irregular distribution resembling oscillations around a stable point (**III**). In contrast, the descending trajectory (b1 to b2) emphasizes an amplified fluctuation envelope coupled with parameter shifts associated with a Hopf bifurcation, consolidating the dynamics of a malignancy state around a fixed point. The fixed point histogram displays a peak aligned with the fixed points, highlighting the relevance of features such as clustering centroids in capturing the nuanced hallmarks of malignant transformation.

Models aiming to represent a cell's transformative journey, especially through the lens of the (epi)genetic landscape, have emerged in the scientific community [14,34–40]. These models provide insights into cellular states' temporal flow and stability, enhancing our understanding of cancer's evolutionary pathways. However, they face challenges owing to their reliance on time series data for parameter estimation—a requirement often elusive given the intricacies of capturing real-time biological processes. Recognizing these limitations, techniques have emerged that infer temporality (the ordered succession of events) through pseudotime [41]. Though pseudotime provides a promising avenue to infer transitions between attractors using scRNA-seq data, it is not without its challenges. Events such as mutations, deregulation of the cell cycle [42], and cellular heterogeneity can potentially obscure the interpretation of pseudotime trajectories, further complicating the characterization of cancer dynamics. Yet, such complexities underscore the relevance of a stochastic modeling approach.

Temporal data acquisition presents inherent challenges, giving rise to numerous theoretical hypotheses. Among these, a central theme is a system's statistical behavior, tracking its evolution in phase space over time. A distinguishing feature of some of these perspectives is the convergence of time averages to ensemble averages [43], which is particularly important for snapshot data such as scRNA-seq. This approach offers an avenue to bypass the intricacies of temporal sampling. Expanding on this, some theories emphasize examining specific components within a system, such as basins of attraction [44,45]. Translating this to the GBM context, the distinct aggressiveness levels of each subtype can be interpreted as different moments in a broader cellular narrative. We propose that the ensemble averages derived from GBM subtypes might be akin to time averages of a representative trajectory (Figure 1). This correlation might be especially evident in specific gene space dimensions, likely marker genes. Combined with the fixed points modeling, such intertwining of theoretical insights with observed biological phenomena has inspired our hypothesis, driving us toward a comprehensive modeling perspective.

Given these challenges and insights, our hypothesis is grounded in the idea of a system that, over an extended timeframe, will navigate through all accessible states within the bounds of a cancer attractor with a consistent likelihood. While this assumption might simplify certain complexities, it establishes a robust foundation to probe the stochastic essence of single-cell data. Such a model gathers the concepts of an equilibrium state with inherent variability, offers a streamlined approach to estimate parameters, and interprets the variability presented in data. However, should there be complications in estimating these parameters, we may need to reassess the foundation of our hypothesis. Validity would largely depend on comparing simulated outputs with experimental findings, a step that could significantly enhance our hypothesis's reliability.

In validating our hypothesis, we employ experimental GBM scRNA-seq data to generate a computational model, incorporating principles from (epi)genetic landscape modeling, Langevin's dynamics [46], and Hill Functions [47] with modifications for the GRN dynamics. Our methodology centers on dynamic modeling, integrating raw data with insights derived from the underlying biological processes and mechanisms. In the proposed context, the cancer attractor concept suggests a propensity for specific cancer subtypes, recognizable as distinctive regions in phase space. Additionally, while exploring the gene expression space, some areas are expected to be inaccessible due to biological constraints. Our assumptions include metric transitivity, which means that two points in phase space can be connected by a shortest path in the gene expression space. This phenomenon aligns with the idea that intrinsic cellular noise enhances phase space exploration in cancer cells, diminishing the barriers between different basins of attraction [18]. Building upon these insights into the gene expression space and its constraints, we propose a stochastic model in silico, aiming to quantify the (epi)genetic landscapes derived from scRNA-seq data of four GBM subtypes: Classical, Mesenchymal, Proneural, and Neural. This model also contemplates interactions inherent to a GBM-specific GRN.

Our modeling efforts promise more than theoretical research. For instance, they hint at tangible avenues for interpreting the intricacies of genomic instability related to cancer heterogeneity. As various mechanisms that contribute to genomic instability imprint distinct molecular signatures [20,31], by exploring the statistical behaviors explained in our approach, we could potentially correlate mechanisms underlying these signatures, offering a chance to identify novel therapeutic targets [48]. In other words, we aimed to associate alterations in statistical properties observed in scRNA-seq data with molecular-level events that modify the system's exploration of possible states. Distinguishing these unique molecular imprints would allow us to forge stronger connections between the 'geometry of heterogeneity' seen in single-cell data and distinct instability mechanisms, offering a more enriched understanding of tumor biology. Consequently, this statistical property could be viewed as a consequence of the progression of the malignant state.

In light of these insights and their challenges, this report seeks to integrate the theoretical foundations of the Waddington (epi)genetic landscape with the wealth of data

emerging from single-cell technologies. By leveraging a stochastic dynamics model, we aimed to unravel the intricate mechanisms that sculpt the heterogeneity inherent to the GBM landscape. Our methodology provides a data-driven quantification of the (epi)genetic landscape specific to GBM and its respective subtypes. Additionally, we probed the statistical dynamics of our in silico model, establishing a framework for subsequent inquiries and potential practical applications. Furthermore, it contributes to developing studies on a biological system's possible long-term behavior and stability.

2. Results

The results section is organized as follows: (i) We present the gene regulatory network, which defines the vertices (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 dynamics inside the basins of attraction and the hypothesis of equivalence between the sample and time average. The results may also serve as a preliminary study for analyzing experimental data as they become 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.

2.1. Glioblastoma GRN

Our initial step was constructing a GRN, with the methodology outlined in Figure S3 (Supplementary Materials). This network, visualized in Figure 2, captures the intricate interplay among key genes and markers related to GBM subtypes. Primarily based on an initial list of GBM markers, the MetaCore platform autonomously expanded the network, ensuring an objective, bias-free augmentation. The resulting structure comprises 40 vertices and 242 edges: 187 activations, 11 self-activations, 41 inhibitions, and 3 self-inhibitions. Table S1 details the vertices, including those initially selected and those added by MetaCore.

2.2. Clustering Methods and Parameters Estimation

A central question in the scRNA-seq analysis is how to interpret gene expression variations across individual cells. Initial data analysis revealed a group of genes with apparently multimodal distributions (see Supplementary Materials, Figure S1). Such patterns, inherent in complex systems such as GBM, possibly hint at various cellular states, emphasizing the tumor's heterogeneous nature. This observation led us to investigate the extent to which this pattern represented the presence of multiple clusters (see Supplementary Materials, Figure S2). Focusing on the four pivotal marker genes delineating GBM subtypes (as described in Section 4.4), our analytical approach employed dimensionality reduction using t-SNE in Wolfram Mathematica. With a perplexity of 60, this method facilitated an optimal visualization of potential clusters, as depicted in Figure S4 (Supplementary Materials).

We utilized two different clustering methods (k-means and NbC) to perform the clustering analysis. We configured the built-in Mathematica functions for both clustering methods with the 'PerformanceGoal' set to quality, the 'CriterionFunction' set to standard deviation, and the 'DistanceFunction' set to Euclidean distance. By comparing the results of these two clustering methods, we aimed to understand the underlying data structure and identify the optimal number of clusters for the given gene expression data. All clusters' statistics are available in the Table S2.

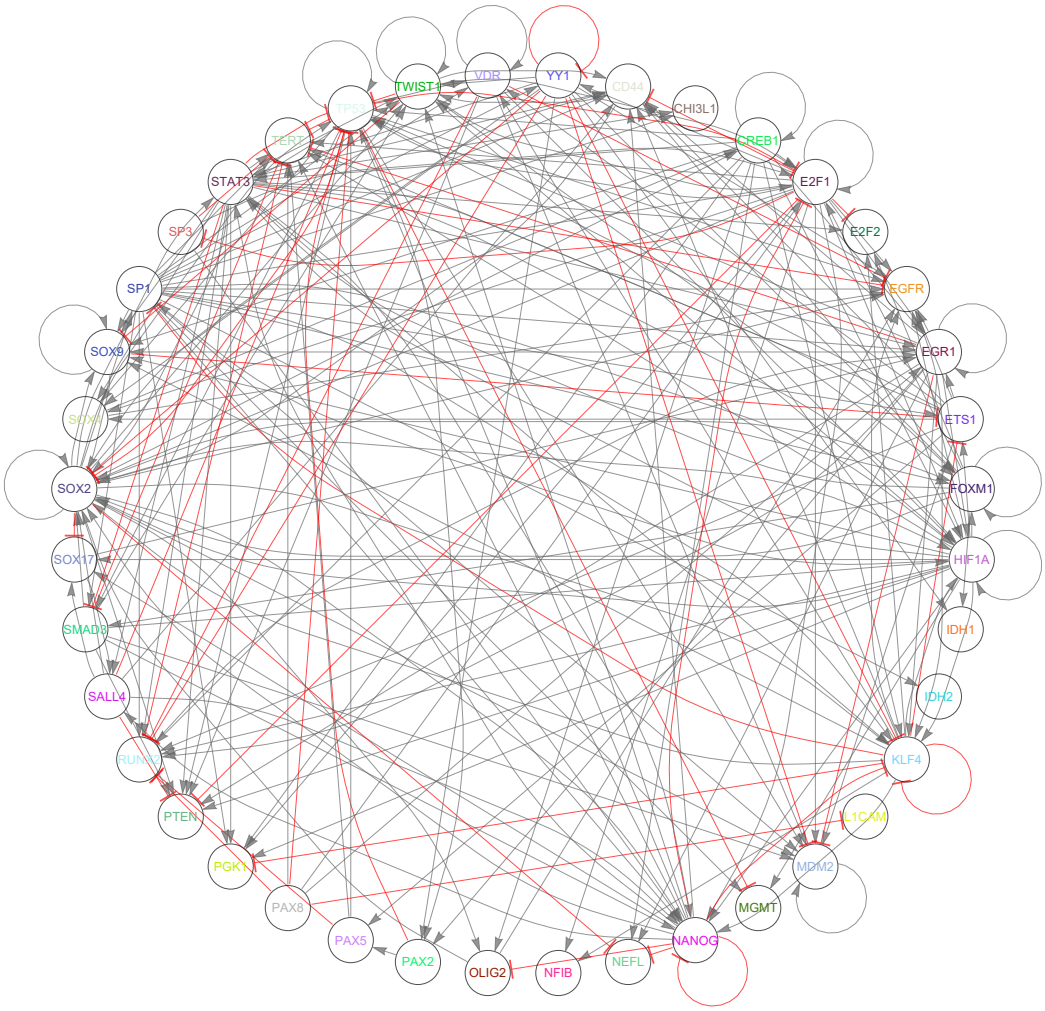


Figure 2. Gene regulatory network for single-cell RNA sequencing of the tumor core of four patients with GBM. Grey lines with flat arrows represent activations, and red lines with arrowheads represent inhibitions. It contains 40 vertices and 242 edges, with 187 activations, 11 self-activations, 41 inhibitions, and 3 self-inhibitions.

Parameter estimation was conducted using the centroid coordinates of all genes within each cluster in an oversized fit, considering $k = 1$. All auxiliary parameter values are available in the Table S3. An interval ranging from 0.01 to 10 was used to adjust 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 (differences in the interval resulting in approximately 1% changes in the indicator values were disregarded). The first estimate was performed using Equation (18), obtaining 784 sets of parameters. The distribution of values for each parameter and both clustering methods can be seen in Figure S5 (Supplementary Materials). 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. In contrast, others may require considerable compensation, without which it could lead to undesirable changes. The second parameter estimation using Equation (19) was performed on top of the first to adjust the values for each interaction. This step's distribution can be seen in Figure S6 (Supplementary Materials). 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 S7 and S8 (Supplementary Materials) 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. In contrast, genes such as *CD44* and *EGFR* exhibited high residuals and variations due to their distant centroid values. While these residuals may indicate a 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. We present all residual values and statistics in the Supplementary Tables S4 and S5.

The final step of the parameter estimation process aimed to assess the compatibility between the cluster centroids derived from experimental data and those obtained from parameter adjustments. To accomplish this, we selected the parameters with the smallest T and R values (as displayed in Table 1). Figure S9a,b (Supplementary Materials) 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 matrix 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 S10 (Supplementary Materials) 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. All parameter values are available in the Supplementary Table S6.

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 superscript 0 and 1 in R_∞ and R_1 inform if it is for the first or second estimation. The right side parameters correspond to the optimization considering stochastic dynamics, with G/C as the mean gene number per cluster when considering the optimization of the parameters.

n	$h(x)$	f_a	f_b	R_∞^0	R_∞^1	R_1^0	R_1^1	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

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, 784 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 S9c,d (Supplementary Materials) 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. After varying the regulation weights in each stochastic simulation, we got attractors A and C with an average of about 15 genes within two sigmas in relation to the observed data, attractor B with an average greater than 20, attractor D with an average close to 20, and attractor E with 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 lead 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.

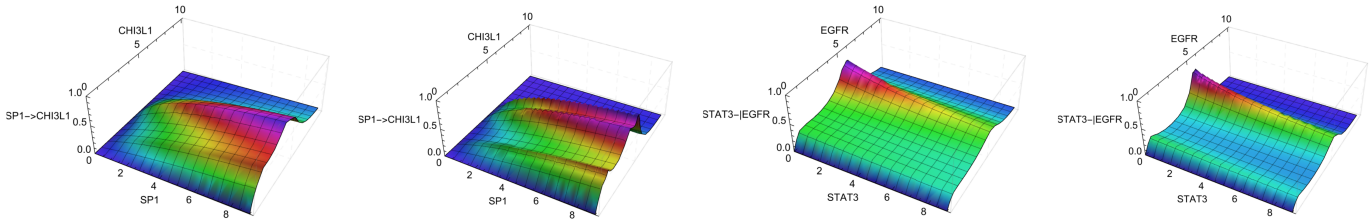
2.3. Regulation Functions

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 Equation (15). For example, Figure S11 (Supplementary Materials) shows the activation and inhibition regulation functions for some 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. Figure S11a,b correspond to the regulation functions of the first clustering choice, while Figure S11c,d to the second choice. Figure 3 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 in the transverse direction 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 (Figure S11—Supplementary Materials).

Considering constant target gene conditions, the regulation functions recover one-dimensional curves, similar to standard Hill functions with $n = 1$. At high target levels, these functions approach zero, mirroring trends in our experimental data. This pattern mimics potential biological mechanisms not explicitly detailed in the network, such as missing environmental signals or missing interactions. Notably, lower inactivation values mean greater inhibitions; a smaller V^b value indicates increased inhibition and decreased basal activation. Furthermore, the modified regulation function does not exhibit peaks at zero target concentrations, a deliberate change to prevent unwanted activation peaks

that might affect observed null data values. Beyond these specific cases, the system's progression will be guided by the interplay of network interactions, parameter values, and noise.



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

Figure 3. 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. The colors represent the vertical axis values, highlighting the representation of the 3D dimension. (**a**,**b**) Activation values using the k-mean and NbC clusters, respectively, ($f_a = 0.1$); (**c**,**d**) Inhibitory interactions using the k-mean and NbC clusters, respectively, ($f_b = 1.3$ and $f_b = 1.1$).

2.4. Comparing Experimental and Simulated Data

2.4.1. Noise and Distributions Compatibility

Data transformation often results in a dependence of the standard deviation on the mean [49]. Part of this dependency can be reduced depending on the normalization process [50]. However, to achieve a better fit between the simulation outcomes and the sequencing data, the function for the multiplicative noise was defined as Equation (26), with $p = 0.23$ and $c_1 = 8$, which were found empirically to fit the data best. Figure S12 (Supplementary Materials) 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 S13 (Supplementary Materials) compares experimental and simulated data. Figure S13a shows the simulation outcome for time 50 (500 steps), Figure S13b shows time 25 (250 steps), Figure S13c shows time 5 (50 steps), and Figure S13d shows 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. Gene expression values for each of the three times are available in the Supplementary Tables S7 and S8.

2.4.2. Clusters Compatibility

To evaluate the congruence between patterns in simulated and experimental data, we employed k-means [51] and Neighborhood Contraction (NbC) [52] on the experimental data and extended the analysis to include Gaussian Mixture [53] for the simulated datasets. The clustering outcomes are depicted in Figure S14 (Supplementary Materials), which represents data simulated from parameters estimated after the k-means and NbC clustering of experimental samples. Clusters were obtained in two reduced dimensions instead of four original marker gene dimensions, facilitating a more tractable visual assessment. Preliminary analyses suggest NbC's superior performance in the reduced dimensional space for both simulation cases.

To advance our comparison, we assessed centroids of simulated data juxtaposed with data distributions. Figures S15 and S16 (Supplementary Materials) illustrate these for k-means and NbC methods. The centroids, regardless of method, particularly between NbC and Gaussian Mixture, displayed notable similarities. These centroids from the simulated

data appeared to represent the centers of the experimental data's multimodal distributions more precisely than direct clustering. Such findings could underscore biologically pertinent insights if these distributions reflect core biological activities.

To further elucidate our findings, we analyzed the proportion of data points in each cluster using pie charts from both experimental and simulated data. Figures S17 and S18 (Supplementary Materials) illustrate these distributions. Clusters A and B demonstrate consistent proportions across most scenarios, except when using the Gaussian Mixture on NbC simulated data. Other clusters displayed method-specific variations, possibly influenced by metastable clusters. Such clusters can challenge clustering techniques, leading to misclassification or indistinct cluster boundaries. Since clustering methods may address metastable clusters differently, discrepancies arise in both qualitative and quantitative outcomes. Consequently, the chosen clustering method is critical in determining clustering results.

We also assessed the distribution within each cluster. Figures S19 and S20 (Supplementary Materials) indicated that standard deviations in every simulated data cluster were consistently lower than in experimental data, suggesting narrow local stability influenced by estimated parameters, and potential noise-induced jumps between attractors. Through boxplots and histograms, we contrasted gene expression distributions within clusters. As evidenced by Figures S21 and S22 (Supplementary Materials), the most congruent gene expression distributions were observed in clusters B, C, and E for post-k-means, and A, B, and G for post-NbC. Figures S23 and S24 (Supplementary Materials) present the specific distributions, excluding *NEFL* due to its near-zero distribution. These findings emphasize our earlier observations about centroid alignment and standard deviation differences. Parameter estimates using NbC produced divergent simulated clustering outcomes, indicating result instability.

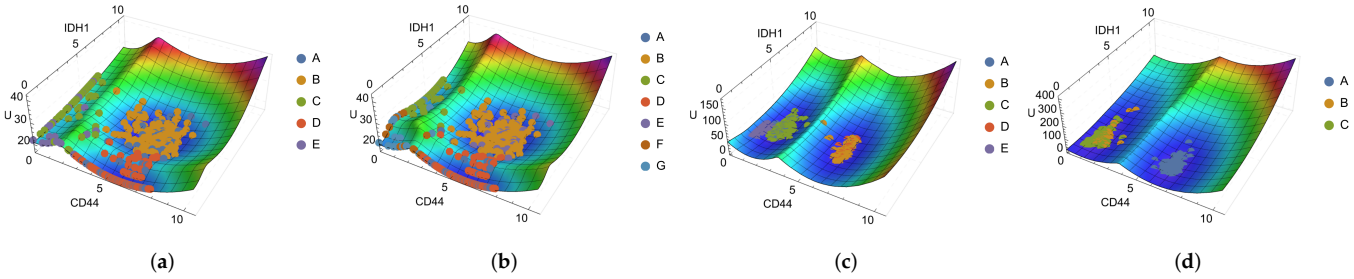
In analyzing scatter plots of the combined marker genes, excluding the *NEFL* gene, transitional points between clusters were evident. As depicted in Figures S25 and S26 (Supplementary Materials), these points might sustain the narrow stability but potentially challenge the sample-time average equivalence if they arise frequently. As these points denote distinct cells at the same time step, further exploration of single trajectories over time is essential. Another observation is that, despite the differences in distribution dispersion, a qualitative analysis (low and high) of the clusters would have compatible results. This suggests that the clusters may still be comparable when focusing on their overall trends rather than the specific dispersion of data points.

2.4.3. 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 S27 and S28 (Supplementary Materials) present the landscapes for experimental and simulated data clustering methods. 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 was significantly smaller than nonnull 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 4a,b), *CD44* × *NEFL*, and *EGFR* × *IDH1*, with the presence of new basins. When comparing these experimental landscapes to the simulated ones (Figure 4c,d), as expected from our previous discussion of Section 2.4.2, 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.



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

Figure 4. (Epi)genetic landscapes for experimental (**a**,**b**) and simulated (**c**,**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.

2.5. Dynamics Inside Basins for the 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 statistical dynamics inside the basins of attraction. First, let us 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 clustering using a Gaussian mixture after the parameters' estimation of the k-means clustered data. Figure S29 (Supplementary Materials) 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 5). Under the sample and time average equivalence hypothesis, such transitions may not be frequent, occurring on a longer timescale 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 statistical properties.

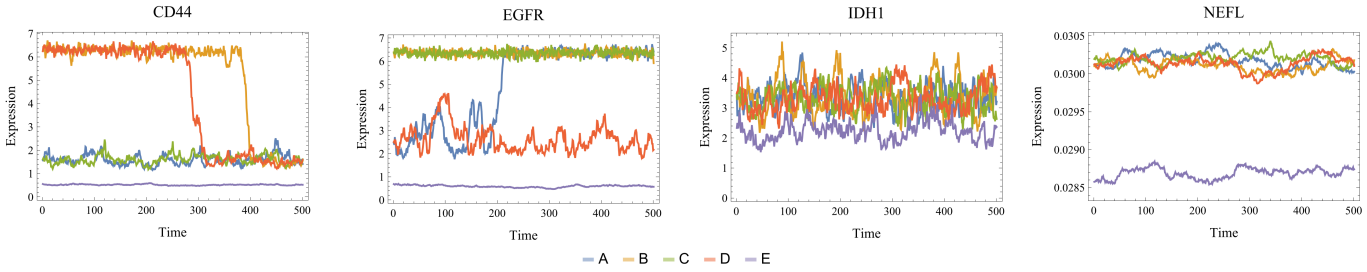


Figure 5. Trajectory plots for each basin (A, B, C, D, and E), showcasing the dynamics of the four marker genes (CD44, EGFR, IDH1, and NEFL). The trajectories illustrate the time evolution from initial conditions chosen as the centroid of the clusters found by the Gaussian mixture algorithm and using the parameters estimated from the k-means centroids. The horizontal axis represents time steps, and the vertical axis shows the expression level of each gene.

To investigate the transitions between components further, we focused on two main features: identifying and quantifying 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 challenges associated with complex dynamic 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 such as distances or standard deviations. Ultimately, 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 S30a,b (Supplementary Materials) show the quantification of the transitions between the basins. Absent basins indicate that there were no transitions to or from them. Figure S30a 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 S29. Figure S30b quantifies the probability of a jump from one cluster to another. The frequencies were normalized by the total outgoing events. For instance, considering the system at cluster A, the figure indicates a 0.85 probability of jumping to cluster C and 0.15 to cluster D. We chose this normalization to allow for the prediction of which cluster the system is more likely to transition to given the present state. An alternative would be pairwise in and out normalization, e.g., A goes more frequently to C than C comes to A. However, given the system is in a specific state, it would not inform the expected transition.

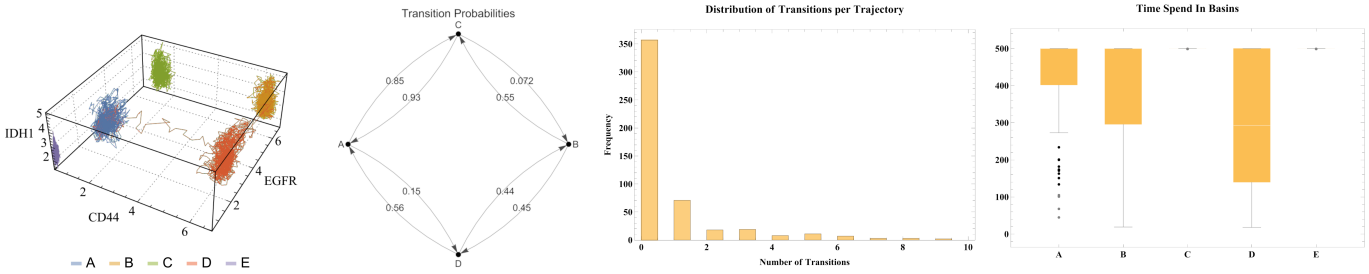
Studying the time spent by a trajectory within each cluster can help us understand the system's stability and the relative significance of each identified state. Long residence times within a cluster suggest a stable state, while shorter times could indicate a transient or metastable state. Figure S31a (Supplementary Materials) 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 S31b 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 hypothesis of low-frequency transitions. However, quantifying a timescale within each basin and basin transitions is still required to verify the hypothesis and further understand the system's behavior.

Before assessing the timescale, let us consider Figures S32–S35 (Supplementary Materials), 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 6 summarizes the results, with (a) one of the 3D trajectories, (b) the transition matrix, (c) 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.

Studying autocorrelations within the trajectories is essential to better understanding the system's dynamics and gaining insights into the timescales. Autocorrelation analysis can reveal the system's temporal structure. By examining these structures, we can verify

timescales that characterize each basin's internal dynamics or are 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 statistical behavior.



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

Figure 6. (**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. Vertices 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 shows 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 represents the correspondent basins.

Figure S36 (Supplementary Materials) 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 timescales, with consistency for both simulations. Besides the visual inspection, we defined another way to quantify this due to the data complexity. We computed the timescale as the minimum time step to reach an autocorrelation value below e^{-1} for each basin, variable, and simulation.

Figures S37–S39 (Supplementary Materials) aim to understand the distribution of these timescales within each internal dynamic and search for possible transition behavior. Figure S37 shows that even with timescales varying between genes, they tend to present similar values for each gene. However, compared to other clusters' timescale patterns, some discrepant cases are observed for *CD44* and *EGFR* genes and for cluster E. We can see in Figure S36a that these cases presented transitions, leading to increased timescales. Figure 7 displays an example of the increased timescale for the *CD44* by comparing the autocorrelation with and without the presence of a transition. These figures highlight the different characteristic timescales within the basins and inter-basins.

Furthermore, Figures S38 and S39 emphasize the values in each basin. Figure S38 shows the significant variations of timescales for the genes within each basin. Still, given the previous discussion, we note that some genes present very narrow timescale 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 S39 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

timescales are interconnected, and the system may somehow compensate for them. The underlying biological processes within the system may lead to interconnectedness that allows for compensation across different timescales. 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.

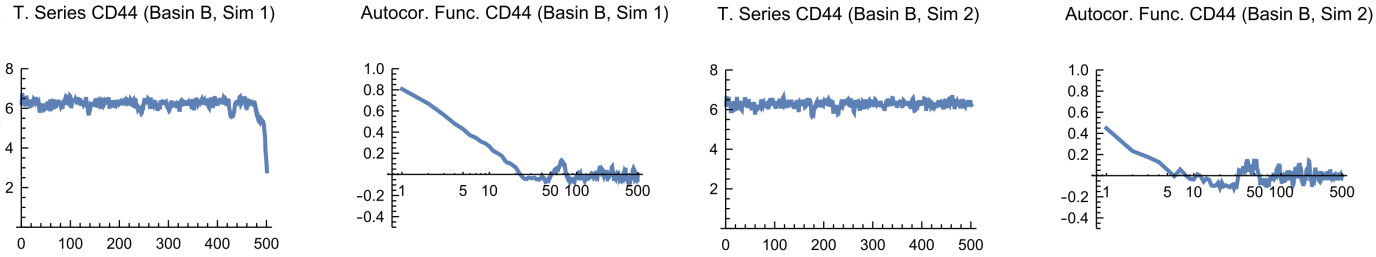


Figure 7. Autocorrelation analysis of time series data for CD44, basin B, and two simulations. Each pair of plots within includes a time series plot (**left**), with the horizontal axis representing time and the vertical axis representing expression values, and an autocorrelation plot (**right**) with the horizontal axis representing time lags and the vertical axis representing autocorrelation values.

After conducting all these analyses, we can finally assess the compatibility between time and sample averages. Figure S40 (Supplementary Materials) demonstrates this compatibility for nearly all clusters and genes. The left panels of Figure S40 represent the average of 100 samples at the final time. In contrast, 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 8). We computed these averages considering their departure states, and they may have jumped to another basin during the process.

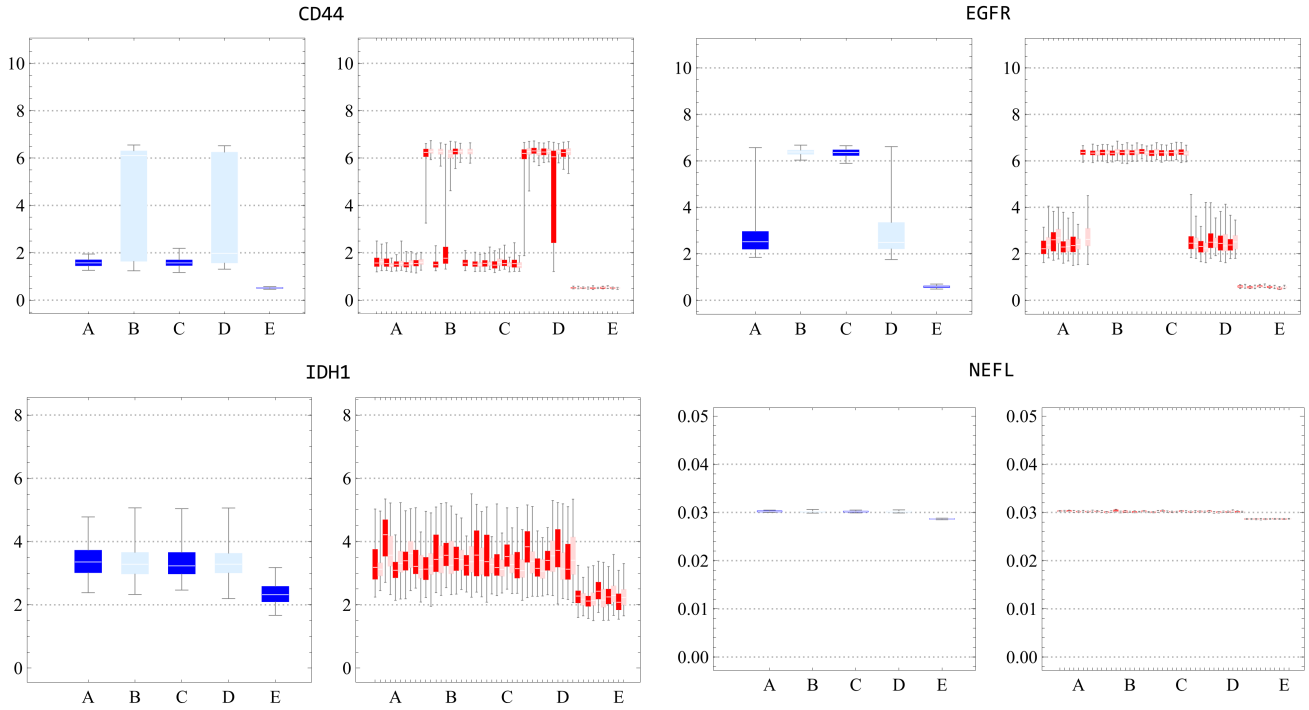


Figure 8. Boxplots for comparison of sample and time averages for all clusters and the four marker genes. The horizontal axis represents the cluster labels, and the vertical axis shows the expression values. The figure displays two sets of plots: the left set shows the sample average of 100 samples at the final time interval, while the right set represents the time average considering 10 trajectories from time 30 to 50 (steps 300 to 500).

In conclusion, these results support our initial assumption of the statistical dynamics behavior inside cancer attractor basins, even if only for simulations. The correspondence with the data suggests that it might also be true for biological systems. 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.

3. Discussion

In an effort to examine scRNA-seq datasets through stochastic modeling, we encountered challenges ranging from a lack of time series data to complex network construction and a vast number of parameters. Here, we discuss our results and interpretation in the context of the present and possible future contributions.

3.1. Dynamics of GBM: Aggressiveness and Heterogeneity

In this report, we introduced an approach to investigate the dynamics underlying tumor heterogeneity as revealed by scRNA-seq data of GBM cancer cells. We strategically focused on GBM due to its pronounced aggressiveness, which allowed us to make specific assumptions about inherent dynamics. Recognizing GBM's aggressive nature, we assumed our sample contained a considerable portion of the cancer attractor domain. We proposed that factors such as genomic instability, epigenetic changes, and selective pressures could alter inherent trajectories, yielding diverse dynamic behaviors. For instance, stochastic fixed points might become increasingly noisy, or a stochastic limit cycle could either resemble a stochastic fixed point or bifurcate towards it (Figure 1).

To our knowledge, this report represents the first application of an integrated framework combining clustering techniques to identify stochastic fixed points and model the cancer attractors dynamics of GBM subtypes using scRNA-seq data. This methodology provided conceptual support for the empirical application of the cancer attractor model. In essence, while our dataset provides a snapshot, this framework provides an underlying theory that shapes this snapshot. This scenario provided a context so we could consider the cluster centroid (sample mean) equivalent to the time average of a representative trajectory within the cancer attractor domain—each cluster symbolizing a distinct cancer subtype phenotype.

We also delineated the clusters and identified the centroids by examining a reduced dimension of specific marker genes, as highlighted in studies such as Sáez et al. [54]. We proposed that these reduced dimensions are particularly amenable to the manifestation of stochastic fixed point dynamics, owing to the inherent characteristics that render them suitable for subtype classification. Our analyses indicated that expression values (both experimental and simulated) are indeed constrained within subregions of these marker gene dimensions, which proved instrumental in studying the dynamics of subtypes. These results reinforced the importance of these subtypes biomarkers. Despite the inability to capture the full breadth of experimental data distributions, our model demonstrates that scRNA-seq data can, to a significant degree, be characterized by our theoretical construct, a premise we plan to expand upon in subsequent sections. We regard this as a progressive step in improving the validity of our hypothesis and demonstrating the viability of our modeling approach to elucidate the dynamics underpinning the experimental data.

3.2. GBM GRN Dynamics: Refining the Model and Centroid-Based Parameter Tuning

Our modeling approach aimed to recreate observed data with stochastic simulations to test how our hypothesis could account for the data patterns. We initiated our model by establishing a GRN and setting up its dynamics. For our GRN, we focused on marker genes and employed the MetaCore platform to establish their interconnections. Although this GRN is a simplified model of the complete biological system, we hypothesized that the dynamics of these marker genes would sufficiently encapsulate the subtypes' dynamics. We proposed modifications to the Hill functions to better reflect these dynamics. Integrat-

ing average values from gene expression clusters into the regulatory model was a key enhancement. This strategy helped to enclose a range of conditions not explicitly included in the model, such as epigenetic differences in patient samples, genetic mutations, missing GRN interactions, or influences from the tumor environment. Our method was designed to construct a dynamics model that closely aligns with experimental data while attenuating the network structure's sensitivity. By adapting the Hill functions, we overcame inherent limitations and successfully captured the centroids observed among different patients. This method proved to be an efficient and effective way to simplify the network's complexity by incorporating these variations through the data.

Building upon the improved Hill functions, linear programming was employed to find the parameters that interpolate the dynamics passing through the centroids in biomarker dimensions. As shown in Figure S1 (Supplementary Materials), patients' data follow similar distribution patterns. Based on this observation, we proposed that a more accurate characterization of the disease landscape could be achieved by considering all patient data to improve the sampling around each cancer attractor. Additionally, tuning the model parameters to fit the centroids of all clusters simultaneously was instrumental in preventing the model from being overfitted to any single cluster. It also allowed us to explore whether a single set of parameters could effectively capture intersecting dynamics between all patients and subtypes. In other words, if different subtypes could coexist within the same biological landscape. Indeed, our findings demonstrated that we successfully identified a set of parameters that not only captured the centroids of the data but also delineated a possible dynamic of the disease subtypes. This aligns with findings by [4], which demonstrate the plasticity of GBM cells and their potential for a single cell to give rise to multiple subtypes. These insights suggest a disease landscape where such cellular states are not only present but also dynamic, with the potential for transitions between subtypes that may contribute to the progression of the disease under certain conditions. Additionally, the selected parameter values leading to these conditions may point to specific, potentially latent configurational states of epigenetic regulation that permit such transitions. In this sense, identifying a singular parameter set operating as an effective representation of biological phenomena signifies underlying uniformity within the biological dynamics. This attribute could exist beyond inter-patient differences. Lastly, using a simplified GRN in our hypothesis does not limit these findings. More complex networks would likely present even more parameter possibilities, of which our model represents a mere subset.

Building upon these deterministic parameters, stochastic features were introduced to match the centroids and distributions observed in the experimental dataset. The challenge lies in the precise calibration of this noise. We sought a noise that neither dominates the dynamics nor is negligible. By achieving this, we assert that the noise model, albeit simplistic, serves as an effective representation. Recognizing the complex factors contributing to this noise, ranging from intrinsic cellular mechanisms to varied patient-specific influences, an effective noise encapsulation arises as a pragmatic choice. Furthermore, the study's aim was not to trace individual cellular trajectories but to discern broader statistical patterns that emerge from the collective trajectories. Since we are trying to observe the 'subtypes envelope', a prior noise was used to get the first idea of the results before its improvement. This iterative process allows for a progressive refinement of the noise model in future investigations.

3.3. Parameters Variability: Heterogeneity and Genomic Instability

As discussed, we have expanded upon traditional cell-type clustering by incorporating dynamic profiling, highlighting the transient nature of cellular states. Our results accentuated the importance of biomarkers, which proved instrumental in delineating stable states and enabled a more subtle understanding of subtype dynamics. However, while our model successfully replicated the centroids of these classifications, it did not achieve the same accuracy for the dispersions. This section focuses on the challenges faced due to the variability of parameters and the interpretation of the underlying causes, including

heterogeneity and genomic instability. Furthermore, their impact on modeling heterogeneity in cancer, including genomic instability. These features are intimately associated with cancer and are directly reflected in scRNA-seq data. They can be incorporated into the gene expression variability, both inter- and intra-clusters.

These mechanisms introduce variability that makes dynamical modeling difficult, messing with the parameters and hyperparameters, making its unique determination difficult. As pointed out in [55,56], multiple parameter combinations could yield results closely mirroring those observed experimentally. We adopted statistical criteria to analyze and select the most meaningful parameters to address this multiplicity. Additionally, by changing the hyperparameters, we constructed an extensive parameter analysis that gave us insights into the distribution of these parameters. In this way, the heterogeneity both complicates and helps parameter estimation. It complicates in the sense that the complete variability probably does not reflect a single parameter set but helps in a way that might end up defining the region of the gene expression space that could contain the centroid of the cancer attractor.

By using different hyperparameters to find the best regulation parameters, we obtained parameter distributions for each gene. These distributions may be more than due to methodological change; they could be about intrinsic biological characteristics associated with the genes and their GRN topology. Genes showing wide variability in parameter values across different hyperparameter settings might indicate that the behavior of these genes is highly sensitive to changes in their regulatory environment. Multiple sets of parameters that lead to a driven force near zero reflect the nature of heterogeneity, showing that different cells would display different but near expression values but still could lie around the same attractor. Genes with parameter values that remain relatively consistent across different hyperparameters might be considered more robust or stable in their behavior. This robustness might be due to GRN's built-in compensatory mechanisms. In the genomic context, it suggests that these gene regulatory networks have evolved to maintain their function despite external perturbations. When such a gene does mutate in cancer, the mutation might have profound effects, given that the gene's behavior is typically so consistent. In these cases, deviations from this narrow window can destabilize the system, potentially leading the cell into aberrant behaviors. This fact might be intimately related to genomic instability.

Additionally, such genes and the GRN regions they comprise might be less resilient to perturbations and more susceptible to disruptions. This highlights where certain genes can act as points of vulnerability within the network, predisposing the system to disequilibrium and chaotic behaviors when altered. This could be an important feature of this modeling approach, revealing potential genes and regulatory mechanisms to study, new diagnostic pathways, and targeted therapies. However, this fact still needs further verification, increasing sampling size and variability, using a validated network, and comparing the results with biological experiments.

Our parameter modeling offers a robust enhancement to biological interpretation, addressing the limitations of prior studies that often rely on arbitrary parameter values [35,36,57]. Furthermore, we demonstrated that residuals, traditionally utilized as indicators for estimation quality, hold potential as network and model goodness measures. This proposition needs comprehensive exploration in subsequent research efforts. Notably, the presence of nonzero residuals was anticipated, asserting that an exact equilibrium characterization would necessitate the incorporation of multiple complex elements, rendering a precise depiction implausible.

3.4. Landscape and Dynamics inside Basins

Central to our investigation is the complexity of modeling dispersions, the varied number of basins, and the emergence of distinct phenotypes. An in-depth comparison of experimental and simulated data enabled us to estimate our model's performance and predictive capability. The observed deviation from the expected dispersion of experimental

data could underscore the need to explore multiple parameter sets following the stochastic fixed-point dynamics. Moreover, technical noise in experimental data—unrelated to core biological processes—might have masked genuine dispersion that our simulated landscape struggled to reproduce. However, an important feature is that the experimental and simulated landscapes identified multiple attractors, reinforcing the hypothesis of multiple stable states. Another point is that the GRN and parameter values might change due to mutations and epigenetic modifications. In that case, it would still support the investigation of short-term dynamics characterization and introduce a possible avenue for investigating tumor progression through longitudinal analysis.

In the final stage of the investigation, we chose suitable parameters. We investigated the in silico dynamics inside each basin, aiming to gain new insights by induction from simulated data. This analysis involved 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. By investigating transitions between attractors, we aimed to understand their stability and the interplay between sample averages and time averages, especially since frequent transitions could challenge the equivalence of these two metrics. To ascertain this, we assessed the frequency of transitions and the timescales within each attractor to better understand the system's dynamics. It also helped to verify the extent to which the stochastic fixed point could achieve sufficient stability to accommodate the wide experimental data distributions.

In conclusion, we stress the value of studying these aspects in more detail by identifying the factors influencing the system's properties. If cancer aggressiveness causes a shift leading to a statistical dynamics regime as hypothesized, investigating the detailed causality behind this could provide critical insights into cancer progression and potential interventions. To our knowledge, the in silico verification of stochastic fixed point dynamics and transitions between GBM attractors using this integrated approach has not yet been reported.

4. Materials and Methods

Historically, biology has utilized models to interpret complex biological phenomena. Traditional approaches often employed model organisms or cell lines for in vitro studies. However, with recent advancements in computational methods and mathematics, there has been a notable shift towards abstract mathematical models. These models, acting as approximations, allow researchers to navigate and hypothesize within controlled digital environments, simulating the complexities of biological systems. While in silico experiments may not conclusively validate general biological principles, they offer insights into hypothesis outcomes and afford preliminary validation via induction [58].

Gene regulation is one of the most intricate and fundamental biological processes that benefit from computational modeling. Given its inherent complexity and dynamic nature, mathematical modeling has emerged as an indispensable tool for elucidating its nuances.

4.1. Model Background

The regulation of gene expression is a complex process that involves multiple layers and mechanisms [13]. One possible measurement of gene expression is the number of messenger RNA (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 [59].

There are several possible functions to parameterize the interactions of a nonlinear model, but the common choice is sigmoidal functions. Among them, the Hill function is

the most frequent as it has many experimentally observed required characteristics [47]. An example of a driving force F using Hill functions can be seen in Wang et al. [34], with a more general form described by:

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

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 parameter, 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 Equation (1), the gene activation (a) and inhibition (b) parameters may vary for each interaction or even as a function of time (nonautonomous 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 using deterministic differential equations to study general behavior is adequate, biological systems are inherently stochastic [59]. Thermal fluctuations and varying conditions affect the likelihood of interactions and make these systems probabilistic. Consequently, the number of molecules over time 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 [46]. However, solving the CME can be computationally challenging, especially for large systems. An alternative is to use Langevin dynamics [46], which serves as an approximation of the CME, described by a deterministic term $\mathbf{F}(\mathbf{x})$ and a stochastic term $\boldsymbol{\xi}(t)$. In Langevin dynamics, we can treat $\boldsymbol{\xi}(t)$ as random fluctuations (without memory) due to its much smaller timescale compared to $\mathbf{F}(\mathbf{x})$. The dynamics became:

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

where $\mathbf{x}(t)$ is the expression level as a function of time (implicit dependence) relative to random variables of $\mathbf{X}$, $\mathbf{F}(\mathbf{x})$ is the deterministic term representing regulation due to network interactions, and $\boldsymbol{\xi}(t)$ is the stochastic term with average $\langle \boldsymbol{\xi}(t) \rangle = 0$ and amplitude given by its autocorrelation function $\langle \xi_i(t) \xi_j(t') \rangle = 2D_{ij}\delta_{ij}\delta(t - t')$ [37], 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 (Equation (A2)). This equation provides a continuous approximation to the CME [59], 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). In this context, we consider a system described by [14]:

$$d\mathbf{X} = \boldsymbol{\mu}(\mathbf{X}, t)\, dt + \boldsymbol{\sigma}(\mathbf{X}, t)\, dW\,, \tag{3}$$

where $\boldsymbol{\mu}(\mathbf{X}, t)$ is the drift, $\boldsymbol{\sigma}(\mathbf{X}, t)$ is the noise parameter, and dW is the Wiener standard process. We assumed $\boldsymbol{\mu}(\mathbf{X}, t) = \mathbf{F}(\mathbf{X})$ by considering the drift due to the driven force. The

noise $\sigma(\mathbf{X}, t)$ 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 $\sigma(\mathbf{X}, t) = \mathbf{g}(\mathbf{X}, t)$ so that the fluctuations may be described by different timescales and constrained by the defined regulation function. The complete definition of the system is given by the parameters of the deterministic and stochastic components, with the specifics of the multiplicative noise and the regulation function being detailed later in this report.

4.2. scRNA-Seq Data

While theoretical models provide a conceptual procedure to understand gene regulation, capturing accurate data remains paramount. 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.

4.2.1. GBM and Single-Cell Data

Transitioning from the broader picture of scRNA-seq to its specialized utility, GBM stands out as a compelling case study. GBM, renowned for its profound cellular heterogeneity, exemplifies the challenges researchers grapple with when studying complex disease landscapes [5]. Considering the nuances of cellular evolution in tumor environments, a deeper dive into the roles of selective pressures and genomic instability in shaping GBM's intricate heterogeneity became imperative. Yet, it is precisely this complexity that makes GBM a fertile ground for scRNA-seq explorations. In this context, single-cell datasets serve as "temporal snapshots", chronicling the multifaceted expression patterns of GBM's cellular ensemble at distinct timeline intervals. Although these snapshots might appear isolated, a deeper dive reveals they often resonate with the broader dynamism governing cellular behavior. Figure 1 encapsulates this idea of the richness of information each "snapshot" brings to the table. It suggests that while each scRNA-seq dataset offers a temporally distinct perspective, collectively, they can traverse the entire phase space, capturing the essence of GBM's intricate dynamics over time. Such insights emphasize scRNA-seq's transformative potential in unveiling tumor heterogeneity dynamics.

4.2.2. GBM Dataset

We utilized the dataset curated and analyzed by Darmanis et al. [60], which encompasses single-cell resolution RNA sequencing outputs from patients diagnosed with diverse GBM subtypes. The study scrutinized tumor heterogeneity, contrasting the tumor core with its periphery. This dataset aggregates samples from four patients, all diagnosed with primary GBM and characterized by a negative *IDH1* signature (indicating an absence of mutations in the *IDH* gene). Following stringent quality control measures, the dataset retained information from 3589 cells, including various cell types from the central nervous system, such as vascular, immune, neuronal, and glial cells.

The analytical framework employed by Darmanis et al. [60] identified cellular clusters from the dimensionality reduction with tSNE, layered over a dissimilarity matrix. Subsequent clustering via the k-means algorithm refined cellular groupings. A meticulous gene expression audit identified the signature genes of each cluster, the results of which were cross-referenced against healthy tissue data to chart cellular identities. Residual clusters were cataloged as neoplastic, predominantly localized to the tumor core and marked by heightened expression of genes such as *EGFR* and *SOX9*. Further validation against independent datasets from healthy brain tissue and GBM bulk RNA-Seq reinforced the study's findings. An intriguing observation was the conspicuous absence of astrocytes within the tumor core. Furthermore, a consistent expression profile for tumor cells in the peripheral zones was documented across all patient samples [60].

4.3. GRN Construction and Implementation

Our research transitions into its computational modeling phase. The central goal was to create a representative model of the GRNs to understand the dynamics of GBM's subtypes and their inherent heterogeneity. In the subsequent section, we detail the methodology that forms the foundation of this computational framework.

Biological Criteria and Methodological Approach

The challenge we embraced was formulating a GRN that captures the essential dynamics behind GBM subtypes and heterogeneity. The GRN needs to be composed of regulatory interactions (edges) between the genes (vertices) of the system, which are supposed to be representative of the case under analysis (GBM). To begin this endeavor, the first step involved setting up clear biological criteria to guide the curation of these interactions. Our primary focus revolved around molecular mechanisms that directly or indirectly affect the number of mRNA molecules. Our approach embarked on an initial survey of direct and indirect interactions. *Direct interactions*, highlighted at the transcriptional level, are characterized by the binding of transcription factors (TF) to their target gene promoters. 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. Conversely, *indirect interactions* encompass mechanisms that modulate the TF's ability to bind to a gene promoter, such as the ubiquitination of a TF culminating in 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. As we delve into the complexities of these interactions, it becomes evident that a structured approach is needed to assemble the GRN.

To construct this survey of interactions, we began by pinpointing genes and markers pivotal for GBM subtypes [2]. The subtype classification was anchored in the schema presented by Verhaak et al. [21]. We then employed the MetaCore [61] (accessed on 16 April 2022) platform to search for interactions among these genes and markers. Specifically, we used the MetaCore transcription regulation network construction algorithm to build the network, identifying new vertices that bridge the initial gene list (provided in Supplementary Table S9). However, the output generated consisted of several disjoint subnetworks, with the initial genes scattered among them. We addressed this issue using the initial and added genes as a new input list. We repeated this until we obtained a connected network, ensuring that all the genes of interest were included in a single structure (Figure S3 - Supplementary Materials). With the GRN in place, the next step was to cross-reference it with the scRNA-seq expression data.

To achieve this alignment, an algorithm [62] in R [63] (version 4.1.2) was developed to match the scRNA-seq expression data with the network generated by MetaCore. The data preprocessing was performed using the Seurat package (version 4.1.1) [64], from which a *sctransform* normalization was applied to reduce technical bias and recover biologically significant distributions [50,65]. Cell cycle effects were not removed, as such information may lose its accuracy in tumor cells [42]. 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 IDs. With this adaptive network established, the subsequent step involved refining our working model to more accurately represent the cell types and their spatial localization within the GBM landscape.

Building upon our initial network, we adopted the cell type classifications proposed in Darmanis et al. [60]. These authors identified various cell types present in GBM samples, such as astrocytes, oligodendrocytes, and neurons, among others, through clustering and other methods. Centering our study on neoplastic cells nested within the tumor core, we delved into the dispersion of expression value for every patient consolidated in our dataset

(Figure S1, Supplementary Materials). 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 features observed for neoplastic cells present in the periphery of the tumor. The filtration process culminated in an aggregate of 1027 cells, offering a robust foundation for downstream analyses.

To enable the automatic integration of the network with mathematical and computational models, we developed an algorithm available in the provided code repository [66], simplifying the process. The algorithm takes the network as input in a tabular format, representing the connectivity list between each vertex. 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:

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

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

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

and $\mathbf{V}^b$ 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 $\text{rowsum}(\cdot)$ returns the vector with the row-wise sums of the matrix.

To observe the effects of perturbations, we represented the original adjacency matrices positions corresponding to activation (a), self-activation (sa), inhibition (b), and self-inhibition (sb) parameters as arbitrary values using symbolic computation. This method allowed further replacement of these symbols with numerical values. For example, $a = sa = b = sb = 1$ during the parameter's estimation, and a broader parameter space exploration for studying the basins' stability. We could achieve the same by using multiplicative factors for each matrix element when $i = j$ or $i \neq j$.

4.4. Data Analysis: Dynamics Underlying Heterogeneity

Upon constructing our GRN, we have delineated the requisite genes and established the targeted scenario suitable for an in-depth investigation into GBM dynamics. The diverse clusters within our dataset underscore the heterogeneity of tumor evolution. This heterogeneity, especially when viewed through single-cell data, hints at the complex evolutionary trajectories of gene expressions.

At the heart of this exploration is the concept of the cancer attractor [18]. This paradigm posits that, despite genetic differences and the evident heterogeneity, cancer cells often gravitate towards a common state. It is this theoretical backbone that justifies our application of dynamic systems theory. Cellular states, acting as nonlinear evolving trajectories in a multidimensional space [13], are not merely random paths but rather can be viewed as potential courses directed towards these cancer attractors, converging at *basins of attraction.*

Given the challenge of real-time tracking, our focus shifts to snapshot-like data, which reflects cellular states at distinct temporal intervals. Such snapshots can be construed as

potential distributions surrounding these cancer attractors for each GBM subtype. This interpretation allows studying system dynamics indirectly by observing the data variability [55]. Representations akin to Figure 1 enhance our understanding of the underlying dynamics and equip us with refined methodologies for data interpretation and parameter estimation. This integrated approach allows for a more nuanced perspective on the observed heterogeneity, paving the way for an optimized analytical framework.

In light of the cancer attractor concept, which underpins the dynamics evident in our Figure 1 hypothesis, the premise that ensemble and time averages converge becomes central. For example, in the case of aggressive malignancies, such as GBM, mutation accumulation and disease progression might lead to a swifter exploration of phase space. It is imperative to note that not all cells are destined to traverse every available state over extended time intervals. Our hypothesis applies to the system's intrinsic components ('basins of attraction'), as delineated in Palmer [44] and expanded upon in later studies, such as [45,67]. Given GBM's inherent aggressiveness, it is posited that the heterogeneity discerned from our spatial samples provides a panoramic view of this attractor's topography. The cellular subsets within the spatial samples are believed to span a substantial segment of the attractor.

To initiate, we discern clusters within the data (refer to Section 4.2), considering the centroids of these clusters as approximations of equilibrium states. With this presumption, the driving forces (i.e., expression level change rates) near cluster centroids are assumed to be negligible. Consequently, the system can be characterized by:

$$\mathbf{F} = \frac{d\mathbf{X}}{dt} \approx \mathbf{0}\,, \tag{7}$$

which enables the system's stability characterization through parameter estimation (see Section 4.6.2).

Our characterization of clusters relies on dimensions associated with four GBM subtype markers: *EGFR* (classic), *IDH1* (proneural), *NEFL* (neural), and *CD44* (mesenchymal) [2]. Unlike genes linked with the cell cycle or prone to high variability, markers must exhibit constrained expression ranges, ensuring clustering and landscape visualization reliability. While many clusters likely represent distinct cell types and subtypes, others may be spurious due to artifacts, noise, or external factors. Some clusters could also reflect metastable states (short-lived configurations within the phase space). This necessitates careful analysis and validation when interpreting clustering outcomes within the GRN framework.

To operationalize the aforementioned considerations, we employed Mathematica (Version 13.1) [68] to analyze the gene expression data from Darmanis et al. [60]. We applied a dimensionality reduction (t-SNE) to identify the clusters, followed by two clustering methods, k-means and Neighborhood Contraction (NbC). K-means is a popular and computationally efficient algorithm partitioning data into spherical groups [51]. NbC is a density-based method that identifies clusters of varying shapes and densities without a prior cluster number definition [52].

4.5. Mapping the Landscape: Subtypes Distributions

Given GBM's complex landscape, it is crucial to define the framework for our study. This landscape can be visualized as a hierarchical structure, where each layer or 'envelope' captures different cellular dynamics. At a high level, the landscape encapsulates cell-type attractors. Delving deeper, it reveals the intricacies of dynamics within cell types, highlighting subtypes. An even more granular approach would reflect processes within these subtypes, such as metabolism.

Our analysis targets the intermediary level—that of dynamics among subtypes, situated within the broader GBM 'basin of attraction'. This choice enables us to provide detailed insights while maintaining a manageable computational scope. Through our modeling, distinct basins of attraction have been identified, emphasizing GBM's inherent

heterogeneity. These basins reflect genomic instability, epigenetic regulations, and selective pressures driving the observed diversity.

We applied the central limit theorem and the law of large numbers, considering the centroids of the clusters as means of Gaussian distributions and representing gene expression levels associated with potential GBM subtypes. Additionally, we computed and stored each cluster's standard deviation based on the respective averages, alongside the proportions of cells they contain.

Therefore, we modeled gene expression as a Gaussian distribution given by:

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

where $\bar{X}$ represents the sample mean and σ represents the (unbiased) sample standard deviation.

Gene interactions exhibit intricate dynamics, with alterations in one gene influencing others. Nevertheless, in steady-state systems, these interactions are reflected in the statistical distribution, enabling independent consideration of each gene (mean-field approximation). By focusing on the moments that depict the resultant distributions from the system's collective interactions, we can manage a more tractable analysis, capturing the essential behavior of the stochastic dynamics.

The mean-field approximation helps to compute the Waddington (epi)genetic landscape, enabling the study of state transitions and the probability of observing specific gene expression distributions. This approximation allows each attractor α probability to be described by [37]:

$$\mathcal{P}_\alpha(X_1, \ldots, X_N) = \prod_{i=1}^{N} P_{\alpha,i}(X_i), \tag{9}$$

where i is the index of genes, N is the number of genes, and $P_{\alpha,i}$ is defined as (8), each one with sample mean $\bar{X}_i$ and sample standard deviation σ_i.

The probability of a cellular state will be given by the steady-state probability $\mathbb{P}_{ss}$ [37]:

$$\mathbb{P}_{ss} = \sum_{\alpha=1}^{N_{clusters}} w_\alpha \mathcal{P}_\alpha, \tag{10}$$

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

The total steady-state probability (Equation (10)) can be estimated for experimental by clustering the scRNA-seq data and for the simulated data by allowing sufficient time for the in silico system to evolve from sampled initial conditions towards steady states. Once we obtain the steady-states, the process is the same as for scRNA-seq data.

Computing steady-state probabilities enables the system's global behavior to be studied using (epi)genetic landscapes. This theory is based on the flow theory of nonequilibrium dynamical systems, as discussed in Wang [69]. Obtaining steady-state probability $\mathbb{P}_{ss}$ through sampling rather than an analytical solution of the Fokker–Planck equations leads to the *populational landscape.* This landscape U is defined as derived from the negative logarithm of the total probability of the system [37]:

$$U = -\ln \mathbb{P}_{ss}, \tag{11}$$

where U is a dimensionless potential that quantifies the probabilities of states and their 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 cases where the system is in equilibrium, with detailed balance preserved, the Boltzmann relation holds. For these cases, the landscape corresponds to the equilibrium probability. However, in cases out of equilibrium, as considered here, rates of transitions

between states do not balance each other. This causes trajectories not to follow the gradient but to be characterized by a probability flux, as discussed in Wang et al. [34].

Despite these complexities, the (epi)genetic landscape can still provide valuable insights into the behavior of systems out of equilibrium. By investigating the nonequilibrium (epi)genetic landscape, we expect to gain insights into the impact of stochastic fluctuations and transient states. This can help guide further investigations and develop more accurate models for the system's behavior.

4.6. In Silico Dynamics and Simulated Landscape

Having delineated the intricate (epi)genetic landscape of GBM data, we transition to constructing an in silico dynamical model. This endeavor seeks to simulate a compatible landscape and evaluate the emergent properties of the in silico system compared to our initial observations. Central to this modeling process is the accurate representation of gene interactions within the GRN. To achieve this, we employ regulation functions, with a particular emphasis on Hill functions.

4.6.1. Specifying Dynamics: Regulation Functions

Hill functions are widely used for modeling GRN interactions, accounting for the influence of activators and inhibitors on regulation strength. Their steepness reflects the gene expression's sensitivity to variations in regulatory molecule concentrations. They are typically characterized by constant coefficients for each interaction, resulting in a uniform regulation strength when applied to all target genes of the same TF. As monotonic functions, they exhibit increasing regulation strength in response to increasing TF concentrations. These features can lead to obstacles in modeling cancer heterogeneity.

Cancer heterogeneity implies cells might present different regulations. In other words, the same TF concentration might lead to different regulation strengths for different cells. For example, one gene may exhibit high expression for a specific number of TF molecules while another presents low expression. Typical features of Hill functions cannot model this dependence on the target gene.

Another important aspect is that they might face challenges modeling complex regulatory mechanisms when the network is incomplete. For example, gene expression levels are typically biologically constrained. This could be due to internal regulatory interaction or even microenvironment responses. In any case, both constraints are typically intricate to implement [70]. One reason is the need to know a priori all interactions that could constrain gene expression levels. Another is that even knowing these constraints, it would still need to implement a very complex model with complicated network dynamics.

To overcome these limitations, we introduced modified Hill functions, which provided a flexible framework for modeling the relationships between genes and regulators, allowing experimental data to constrain the dynamics. We proposed 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 is a modifier function used to obtain shapes with desired properties.

We tested several options for h, including the original model with $h(x) = 0$ to determine the best model. 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 gene average considering each attractor α. The first term of Equations (15) and (16) was chosen to ensure appropriate behavior for values close to zero with the constants in the denominator empirically verified to optimize the results.

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. Lower levels could rapidly increase 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 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 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.

4.6.2. GRN Edges and Parameter Estimation

Within the context of GRN and cancer, it is observed that parameter values do more than modulate gene interactions. When certain values approach near-zero levels, they can alter the GRN topology, removing specific network connections altogether. This reflects the potential for cancer to reconfigure cellular regulatory landscapes and allows the model to indirectly capture GRN topology from scRNA-seq data in the parameter estimation process.

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.

We proposed different ways of addressing the challenges associated with estimating the deterministic parameters in our model. These challenges arise from the 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 crucial, as selecting unique parameters estimated from the available data can be difficult. 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.

Considering these factors, we explored various parameter estimation strategies that balance model complexity, computational efficiency, and biological interpretability. We considered the following three scenarios: (i) with two parameters per Equation (one for activation and one for inhibition); (ii) with one parameter related to each input vertex (divided between activation and inhibition), that is, $2n$ parameters per Equation (including null values); and (iii) with a combination of (i) and (ii), i.e., $n \times n$ parameters (including null values).

Departing from Equation (7), Equation (18) illustrates the first case:

$$k_i X_i = a_i \left(\sum_j V_{ij}^a \right) + b_i \left(\sum_j V_{ij}^b \right), \tag{18}$$

where a possible biological interpretation is an activation and inhibition intensity proportional to the target gene, for example, due to epigenetic regulations. Equation (19) brings the second case:

$$k_i X_i = \left(\sum_j a_j V_{ij}^a \right) + \left(\sum_j b_j V_{ij}^b \right), \tag{19}$$

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. Equation (20) brings the last case:

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

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}.

To estimate the parameters of cases (i) and (ii), we used the L_1-norm robust regression that can be solved as a linear programming problem [71]. We used the Simplex algorithm in the Mathematica environment [68]. Assuming uniform and constant degradation coefficients for all mRNA molecules, we have for all gene i, $k_i = k$, and Equations (18) and (19) can be rewritten in the form of the following equation:

$$k\mathbf{X} = \mathbf{V}\mathbf{c}\,, \tag{21}$$

where $\mathbf{V} = (\mathbf{V}^a \mid \mathbf{V}^b)$, $\mathbf{c} = (\mathbf{c}^a \mid \mathbf{c}^b)$, for $(\mathbf{c}^a)_i = a_i$ and $(\mathbf{c}^b)_i = b_i$.

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 avoided overfitting individual clusters, potentially hindering other attractors' representation. Math-

ematically, for each centroid vector $\mathbf{X}_\alpha$, we build the matrices $\mathbf{V}_\alpha$ and the vectors $\boldsymbol{\beta}_\alpha = k\mathbf{X}_\alpha$, and stack them as:

$$\mathbf{M} = [\mathbf{V}_1 \mid \cdots \mid \mathbf{V}_{N_{clusters}}]^T, \tag{22}$$

$$\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \mid \cdots \mid \boldsymbol{\beta}_{N_{clusters}}]^T. \tag{23}$$

By doing so, we solve the L_1-norm minimization problem:

$$\min \ ||\mathbf{Mc} - \boldsymbol{\beta}||_1 \,, \tag{24}$$

$$\mathbf{c} \geq \mathbf{0} \,, \tag{25}$$

then we compute and store the maximum residual $R_\infty = ||\mathbf{Mc} - \boldsymbol{\beta}||_\infty$ and the total $R_1 = ||\mathbf{Mc} - \boldsymbol{\beta}||_1$ for each fit as a regression quality measurement.

We obtain the solutions for cases (i) and (ii) for combinations of n and S, the regulation functions coefficients. 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 proposed regulation functions for each parameter combination. After estimating the parameters, the absolute residuals list, R_∞, and R_1 were saved. We performed another estimation for each set of parameters obtained using Equation (19). This additional estimation allowed us to combine the newly estimated parameters with the initial set, as in Equation (20).

4.6.3. Noise Characterization and Stochastic Simulation

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 [56]. We employed a multiplicative noise function $g(X, t)$, which accounts for possible mean and standard deviation dependence after a logarithm transformation [49,50], as shown in Equation (26):

$$g(x; c_0, c_1, p) = x\, c_0 \binom{x + c_1 - 1}{x - 1} p^x (1 - p)^{c_1} \,. \tag{26}$$

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 it belongs to or maintain trajectories around its average value. We performed stochastic simulations to investigate the system dynamics with a time interval of 50 a.u. (time steps of $\Delta t = 0.1$) 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 (27). In essence, given:

$$z = \frac{|\bar{x} - \mu|}{\sigma} \,, \tag{27}$$

where μ represents the cluster centroid of each gene, $\bar{x}$ denotes the time average within an interval from time 30 to 50 (allowing the system to fluctuate around the centroid), and σ is 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. The initial condition was sampled from a uniform distribution for computing the landscapes. This sampling method was employed on the notion that the trajectories of the system, given sufficient time, would populate regions of the phase space in a manner consistent with their statistical significance.

4.6.4. Clusters Comparison and Simulated Landscape

At the culmination of our previous section, we underscored our strategy of analyzing the in silico data akin to the experimental scRNA-seq data. This involved utilizing cluster centroids as a close approximation of fixed points to delineate the attractors and their respective basins. Such an approach serves as an analytical framework and paves the way for the critical validation step: a juxtaposition of in silico and experimental data distributions. To what extent could we reproduce data centroids and heterogeneity considering stochastic fixed point dynamics? Given the necessity of this validation, an essential preliminary step was undertaken—endeavoring to align simulated clusters with experimental ones.

To facilitate the clustering comparison, we rearranged each simulated data cluster to match its closest experimental data cluster. The rearranging starts with (i) computing the centroids of experimental and simulated clustered data. Both cases contain grouped cells linked to each cluster stored in a list of lists manner, with each sublist representing the cells of an α cluster. This way, the centroids for each case are stored in an α-ordered list. With two lists of centroids, we then (ii) calculated the Euclidean distances between each element using the experimental centroids as a reference. This step ends up creating a distance matrix. (iii) The reordering starts with an algorithm that identifies the simulated centroid corresponding to the smallest distance to one of the experimental ones. (iv) These indices are mapped together and removed from the distance matrix. This step ensures that each element in the simulated set is associated with a unique experimental cluster. (v) This process is repeated until all experimental centroids are mapped to one of the simulated data. (vi) The code returns the reordered simulated clusters in a list of lists format. (vii) For differing numbers of clusters, the ones not matched are appended at the end. This code pseudocode is presented in Algorithm 1. This process aids in identifying similarities and differences between the two sets and helps to understand the underlying structure of the data.

Algorithm 1 Reorder Clusters

Require: *clusters1*, *clusters2*
Ensure: Combination of sorted and remaining clusters
1: *centroids1* ← computeCentroids(*clusters1*)
2: *centroids2* ← computeCentroids(*clusters2*)
3: **for all** *centroid1* in *centroids1* **do**
4: Calculate the distance to each *centroid2* in *centroids2*
5: **end for**
6: Find the smallest distance and corresponding *clusters* indices
7: Create new ordering of *clusters2* based on these indices mapping
8: *sortedClusters* ← Newly ordered *clusters2*
9: *remainingClusters* ← *clusters* in *clusters2* not in *sortedClusters*
10: *combinedClusters* ← Combine *sortedClusters* with *remainingClusters* appended at the end
11: **return** *combinedClusters*

Finally, we use different ways to assess the compatibility between experimental and simulated distributions, including their respective landscapes, computed using Equations (9) to (11).

4.7. Dynamics Inside Basins of Attraction

Within the framework of our model, we sought to understand the depth of GBM's heterogeneity by examining the basins of attraction. Our primary focus was to assess the

model's capability to capture data heterogeneity evident in GBM. We aimed to ascertain the stability and scope of the cancer subtype attractors.

Additionally, we ventured into the statistical dynamics inherent within each attractor. This involved evaluating the consistency between sample and time averages, a vital aspect of the used framework. We employed autocorrelation functions, which facilitated the determination of the system's inherent timescales and the propensity for transitions between attractors.

4.7.1. Gene Expression Dispersion in Cancer

The dispersion observed in cancer data prompts inquiries about the underlying dynamics causing such patterns [72]. In this study, we postulate that stochasticity, chaotic dynamics, or a combination of both might influence the observed dispersion. Regardless of the underlying reason, we assume these dynamics unfold within specific basins of attraction (see Appendix A).

We view gene expression dispersion as a consequence of cancer progression. Genetic mutations, epigenetic shifts, and changing cellular microenvironments amplify the system's diversity, resulting in more intricate and varied expression distributions. Such diversity mirrors the myriad ways cancerous cells can organize and interact, mirroring the increasing complexity and heterogeneity of the tumor environment [73].

Past research has delved into the rising entropy associated with these phenomena [74,75]. This increase in entropy points to varied cellular states and possibly fluctuating energy levels. Given that differences in energy levels between cell types closely intertwine with genetics and epigenetics, our aim is to examine the stability of cancer subtypes through the lens of these varied states' attraction basins. We utilize the Waddington landscape concept to study the system's behavior upon settling into these configurations instead of focusing on the journey leading to them.

Our approach centers on observing the system's properties over time to comprehensively capture its dynamics and fluctuations. Moreover, sampling must span a broad spectrum of potential states. We conceive each basin of attraction as a distinct subsystem, wherein the movement between them is minimal yet possible. Such a perspective aligns with the notion of attractors and the definition of subtypes; frequent transitions between subtypes would undermine their distinct categorization.

Furthermore, rapidly evolving malignancies, such as GBM, are expected to explore the phase space more extensively than healthy cells' constrained dynamics or less aggressive tumors. Our analysis, thus, focuses primarily on these basins of attraction, adding rigidity to our suppositions.

4.7.2. Fixed Points: Sample vs. Time Averages

In dynamical systems modeling GBM, attractors represent stable cellular states the system naturally gravitates towards. These cancer attractors, identified through cluster centroids, elucidate the system's tendencies. When analyzing variations around these attractors, the time average and sample average become crucial. Ideally, the time average should align with the sample average from various trajectories for a system closely orbiting an attractor. This alignment sheds light on GBM's inherent dynamics and the stability of its states.

To elucidate this phenomenon mathematically, let us contextualize our variables. The expression level of gene i at a particular time point within the interval T is denoted by $X_i(t)$. The spatial domain or basin, in which these expression levels predominantly lie and which is closely associated with the specific cancer attractor $\mathbf{A}$, is termed $\mathcal{B}(\mathbf{A})$ (refer to Appendix A for a comprehensive definition). Within this schema, we propose a model suggesting that the alignment between the time and sample averages becomes increasingly robust as the temporal subinterval $T_{12} = t_2 - t_1$ enlarges and the sample size, N, increases. This alignment resonates with our understanding of GBM's aggressive progression and manifestation over different timescales. The underlying principle is straightforward: our

mathematical approximation is closer to empirical reality as we gather more data samples and witness increased cellular aggression (indicating shorter times to span the associated basins). This time average expression for gene i can then be expressed as $\langle X_i \rangle_{T_{12}}$, given by:

$$\langle X_i \rangle_{T_{12}} = \frac{1}{T_{12}} \int_{t_1}^{t_2} X_i(t) dt \tag{28}$$

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

$$\langle X_i \rangle_A = \frac{1}{N_{samples}} \sum_{j=1}^{N_{samples}} (X_i)_j, \tag{29}$$

where $N_{samples}$ is the sample size in the basin of attraction $\mathcal{B}(\mathbf{A})$ and $(X_i)_j$ represents the j-th sample of the gene expression level X_i within the basin of attraction $\mathcal{B}(\mathbf{A})$.

Our foundational assumption rests on the idea that:

$$\lim_{T_{12} \to \infty, N \to \infty} \left| \langle X_i \rangle_{T_{12}} - \langle X_i \rangle_A \right| = 0, \tag{30}$$

where T_{12} is the length of the subinterval and the sample size N approaches infinity; the difference between the time and space average approaches zero.

4.7.3. Timescales and the Propensity for Transitions

Finally, timescale separation helps assess the relaxation time of X_i within $\mathcal{B}(\mathbf{A})$ and outside of it. We propose this by examining the decay rates of the autocorrelation function for $X_i(t)$. The autocorrelation function measures the similarity between a variable and its lagged version, with a faster decay implying a shorter relaxation time. This behavior would suggest a propensity for reaching equilibrium. The autocorrelation function of X_i is given by [76]:

$$r_{X_i X_i}(\delta) = \frac{\sum_n^{T_{max}-\delta} (X_{i,n} - \bar{X})(X_{i,n+\delta} - \bar{X})}{\sum_n^{T_{max}} (X_{i,n} - \bar{X})^2} \tag{31}$$

where $r_{X_i X_i}(\delta)$ is the autocorrelation at lag δ, $X_{i,n} \approx X_i(t_n)$ is the value of the time series at time step n ($t_n = n\Delta t$, where Δt is the time step interval), $X_{i,n+\delta} \approx X_i(t_n + \delta)$ is the value of the time series at time step $n + \delta$, $\bar{X}$ is the mean of the time series, and T_{max} is the total number time series points. The autocorrelation ratio normalizes the measure to stay between -1 and 1.

The autocorrelation function provides valuable insights into the system dynamic guided by a driving force. Derivatives greater than zero imply a system driven by network interactions. In this case, the autocorrelation values at different lags capture a pattern and may not oscillate around zero. As the lag increases, the autocorrelation values might decrease, but still with the influence of the driving force.

On the other hand, the driving force's influence becomes constant or negligible for a system in a steady state. Being mainly affected by stochastic fluctuations, the autocorrelation at different lags likely oscillates around zero. It indicates little or no correlation between points and more unpredictable behavior, less dependent on past values. We expect to achieve this behavior fast once centroids (i.e., attractors) are used as starting points. In this case, discrepancies would possibly relate to nonstability or transitions.

To proceed with the autocorrelation analysis, we defined a vector $\mathcal{K}$ with each element representing the autocorrelation at a specific lag δ. Thus, $\mathcal{K}$ can be computed as follows:

$$\mathcal{K} = \left[r_{X_i X_i}(1), r_{X_i X_i}(2), \ldots, r_{X_i X_i}(T_{max} - 1) \right] \tag{32}$$

with each value $r_{X_i X_i}(\delta)$ computed using Equation (31). By computing the autocorrelation values for increasing lags, we measure if time series values that are further apart are linearly related or correlated.

To quantify the relaxation time, we defined the timescale by:

$$\delta^* = \min \delta \mid \mathcal{K} < e^{-1} \tag{33}$$

where δ^* is the lowest lag δ, related to a characteristic timescale, at which the autocorrelation value falls below the threshold e^{-1}.

In conclusion, these methods offer a strategy to validate the aforementioned criteria, delving into the consistent gene expression patterns within the simulated dataset. This examination underscores how such assumptions can significantly enhance our understanding of GBM GRN dynamics derived from scRNA-seq data.

5. Conclusions

Despite advances in the analysis and availability of scRNA-seq data, the dynamics of GRNs are still primarily investigated using arbitrary parameter values. Our work goes beyond descriptive and analytical methods, proposing a conceptual and methodological investigation using scRNA-seq to interpret the heterogeneity and quantify parameter values related to the dynamics in the context of GBM. By investigating statistical properties of GBM aggressiveness, e.g., cluster centroids as fixed points describing the cancer attractor, our method connected biological features (heterogeneity and aggressiveness) and statistical properties of the stochastic dynamics of a GRN model. We proposed that the aggressive nature of GBM influences the dynamics and determines how extensively the cancer attractor is sampled. In this context, the sample average (used to identify the cluster centroid) is equivalent to the time average of a representative sample traversing the cancer attractor for an extended duration.

Our findings demonstrated compatibility between experimental and simulated data centroids. We also investigated the stability and transitions between attractors. Despite the model simplifications, our results offered insights into the dynamics of GBM. The transitions between the basins revealed a possible interplay between subtypes, potentially uncovering factors that drive cancer recurrence and progression. Further exploring the dynamics within the (epi)genetic landscape of GBM subtypes can help understand the path leading to the differentiation of each subtype. This investigation might help uncover dynamical principles underlying cancer development and correlate with molecular mechanisms. We expect that determining unique molecular mechanisms related to the statistical properties of cancer heterogeneity might help develop potential diagnostic tools and personalized therapeutic interventions. Finally, connecting the 'geometry of heterogeneity' with instability mechanisms could offer a different perspective on tumor biology. This approach could be used to assess and monitor the evolution of malignant states, serving as an instrument for diagnosis and treatment.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25094894/s1.

Author Contributions: M.G.V.J. designed the analysis, developed the codes, analyzed/interpreted the data, and wrote the manuscript. A.M.d.A.C. revised the mathematical model and its implementation. N.C. and F.R.G.C. ensured biological accuracy. F.A.B.d.S. provided structural critiques and improvements. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Data Availability Statement: The codes developed for the analysis presented in this report are available in the provided GitHub repository. All data utilized to generate the plots are shared in the Supplementary Materials.

Acknowledgments: We would like to thank Mariana Boroni for reading the manuscript and her suggestions concerning patient-specific analyses in this report.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Basins of Attraction: Deterministic vs. Stochastic Modeling

Gene regulation can be modeled as a dynamical system, represented by a set of equations $\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^n$ characterizing the state vector and $\mathbf{F}(\mathbf{x})$ a vector field defining the dynamics within the gene regulatory network (GRN) model [34]. We consider an attractor, denoted as $\mathbf{A}$, as a set of points in the phase space that the system converges to under a specific set of initial conditions [77–82].

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 [77–82]. In this context, a basin of attraction represents a set of points in the phase space that lead to similar long-term behavior. We considered that each basin of attraction relates to a group of genes associated with specific conditions, such as cancer or cancer subtypes. A more formal definition could be:

$$\mathcal{B}(\mathbf{A}) = \{\mathbf{x}_0 \in \mathbb{R}^n \mid \lim_{t\to\infty} \mathbf{x}(t;\mathbf{x}_0) \in \mathbf{A}\}\,, \tag{A1}$$

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

In the outlined deterministic framework, trajectories converge to attractors given a set of initial conditions. However, biological systems are subject to fluctuations due to inherent stochasticity in gene expression. This randomness leads expression values to be described by probability distributions [73]. The temporal evolution of these probability distributions $p(\mathbf{x}, t)$ is described by the Fokker–Planck Equation [83]:

$$\frac{\partial}{\partial t}p(\mathbf{x},t) = -\sum_{i=1}^{N}\frac{\partial}{\partial x_i}[\mu_i(\mathbf{x},t)p(\mathbf{x},t)] + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial^2}{\partial x_i,\partial x_j}\left[D_{ij}(\mathbf{x},t)p(\mathbf{x},t)\right]\,, \tag{A2}$$

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

$$D_{ij}(\mathbf{x},t) = \frac{1}{2}\sum_{k=1}^{M}\sigma_{ik}(\mathbf{x},t)\sigma_{jk}(\mathbf{x},t)\,. \tag{A3}$$

where $D_{ij}(\mathbf{x}, t)$ is the diffusion tensors' elements.

In summary, the Fokker–Planck Equation (A2) describes the evolution of probability distributions subjected to deterministic and stochastic dynamics. It models gene expression fluctuation's influence on biological systems over time. Studies have shown that in the low noise regime, the solution is a Gaussian and that the first-order moment corresponds to the deterministic model [84]. Therefore, one alternative is to characterize its first and second moments relationships and solve the corresponding system of differential equations.

To analyze the long-term behavior of gene expression levels in GBM GRN dynamics within the context of our stochastic system, we recognized the inherent assumption that individual trajectories converge to the same behavior when averaged over time. Given the possible complex ways that noise could shape the basin of attraction size/geometry, we introduce a custom distance function $d(\mathbf{x}, \mathbf{A})$ [77] to adjust its definition. It allows for being generic and still captures the complexity of the problem. It becomes:

$$\mathcal{B}(\mathbf{A}) = \{\mathbf{x}_0 \in \mathbb{R}^n \mid \lim_{t\to\infty}\min_{\tau\geq 0} d(\mathbf{x}(t+\tau;\mathbf{x}_0),\mathbf{A}) \leq \epsilon\}\,. \tag{A4}$$

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 might depend on the characteristics of the system and the specific dynamics around the attractor. For simplicity, we assumed symmetrical geometries with Euclidean distances and a tolerance level ϵ proportional to the distribution standard deviation. Given our Gaussian distribution assumption, we expect it should deal with in silico model fluctuations.

Analyzing the stability of basins and their inter-transitions, we utilized clustering algorithms and inherent assumptions about system behavior over extended periods. By doing so, we sought to understand the possible relations between GRN dynamics and expression patterns observed in experimental data. Essentially, for such systems, 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 [85]:

$$\langle x_i \rangle_t = \lim_{T \to \infty} \frac{1}{T} \int_0^T x_i(t) dt \, , \tag{A5}$$

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 [85]:

$$\langle x_i \rangle_{ss} = \sum_j x_{i,j} P(x_{i,j}) \quad or \quad \langle x_i \rangle_{ens} = \frac{1}{N_{obs}} \sum_{j=1}^{N_{obs}} x_{i,j} \, , \tag{A6}$$

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

Appendix B. Glioblastoma Multiforme

Among the different types of cancer, Glioblastoma Multiforme is the most common and aggressive type of brain tumor. It has an average survival of around 15 months with only approximately 10% chances of 5-year overall survival and 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 related to patients over 65 years old.

GBM can occur as both a primary tumor and a secondary tumor. Primary tumors account for more than 90% of GBM cases, having a worse prognosis. Secondary tumors account for approximately 5%, being rarer in younger individuals [2]. Both the primary and the secondary tumors have different characteristics. Primary tumors present 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]. GBM can still divide into subtypes according to various criteria and markers, which is challenging. 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*) [2,21–23].

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 presents with both mesenchymal and astrocyte markers; and the Proneural presents with high expression of oligodendrocyte genes [21]. In general, the Classic subtype has high levels of expression for *EGFR*, the Mesenchymal has low levels 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 exhibiting mutations in *TP53*, high expressions of *EGFR*, and deletions in inhibitor cyclin-dependent kinase 2A (*CDKN2A*). Additionally, they found a relationship between the subventricular zone (SVZ) proximity and subtypes' characteristics: the Neural and Proneural types are closer to the SVZ, displaying faster progression, lower survival rates, and 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 studies capable of proposing viable alternatives [3].

References

1. WHO. Cancer Overview. 2022. Available online: https://www.who.int/health-topics/cancer#tab=tab_1 (accessed on 3 August 2022).
2. Sasmita, A.O.; Wong, Y.P.; Ling, A.P.K. Biomarkers and therapeutic advances in glioblastoma multiforme. *Asia-Pac. J. Clin. Oncol.* **2017**, *14*, 40–51. [CrossRef]
3. Gallego, O. Nonsurgical Treatment of Recurrent Glioblastoma. *Curr. Oncol.* **2015**, *22*, 273–281. [CrossRef] [PubMed]
4. Neftel, C.; Laffy, J.; Filbin, M.G.; Hara, T.; Shore, M.E.; Rahme, G.J.; Richman, A.R.; Silverbush, D.; Shaw, M.L.; Hebert, C.M.; et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **2019**, *178*, 835–849.e21. [CrossRef] [PubMed]
5. Patel, A.P.; Tirosh, I.; Trombetta, J.J.; Shalek, A.K.; Gillespie, S.M.; Wakimoto, H.; Cahill, D.P.; Nahed, B.V.; Curry, W.T.; Martuza, R.L.; et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **2014**, *344*, 1396–1401. [CrossRef] [PubMed]
6. Tirosh, I.; Izar, B.; Prakadan, S.M.; Wadsworth, M.H.; Treacy, D.; Trombetta, J.J.; Rotem, A.; Rodman, C.; Lian, C.; Murphy, G.; et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **2016**, *352*, 189–196. [CrossRef] [PubMed]
7. Kim, C.; Gao, R.; Sei, E.; Brandt, R.; Hartman, J.; Hatschek, T.; Crosetto, N.; Foukakis, T.; Navin, N.E. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **2018**, *173*, 879–893.e13. [CrossRef] [PubMed]
8. McGranahan, N.; Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **2017**, *168*, 613–628. [CrossRef] [PubMed]
9. Marusyk, A.; Tabassum, D.P.; Altrock, P.M.; Almendro, V.; Michor, F.; Polyak, K. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **2014**, *514*, 54–58. [CrossRef] [PubMed]
10. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.; Kinzler, K.W. Cancer Genome Landscapes. *Science* **2013**, *339*, 1546–1558. [CrossRef]
11. Shen, H.; Laird, P.W. Interplay between the Cancer Genome and Epigenome. *Cell* **2013**, *153*, 38–55. [CrossRef]
12. Esteller, M. Epigenetics in Cancer. *N. Engl. J. Med.* **2008**, *358*, 1148–1159. [CrossRef] [PubMed]
13. Huang, S. Systems biology of stem cells: Three useful perspectives to help overcome the paradigm of linear pathways. *Philos. Trans. R. Soc. B Biol. Sci.* **2011**, *366*, 2247–2259. [CrossRef] [PubMed]
14. Li, Q.; Wennborg, A.; Aurell, E.; Dekel, E.; Zou, J.Z.; Xu, Y.; Huang, S.; Ernberg, I. Dynamics inside the cancer cell attractor reveal cell heterogeneity, limits of stability, and escape. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 2672–2677. [CrossRef] [PubMed]
15. Takahashi, J.S. Transcriptional architecture of the mammalian circadian clock. *Nat. Rev. Genet.* **2016**, *18*, 164–179. [CrossRef] [PubMed]
16. Pomerening, J.R.; Kim, S.Y.; Ferrell, J.E. Systems-Level Dissection of the Cell-Cycle Oscillator: Bypassing Positive Feedback Produces Damped Oscillations. *Cell* **2005**, *122*, 565–578. [CrossRef] [PubMed]
17. Nelson, D.E.; Ihekwaba, A.E.C.; Elliott, M.; Johnson, J.R.; Gibney, C.A.; Foreman, B.E.; Nelson, G.; See, V.; Horton, C.A.; Spiller, D.G.; et al. Oscillations in NF-κB Signaling Control the Dynamics of Gene Expression. *Science* **2004**, *306*, 704–708. [CrossRef] [PubMed]
18. Huang, S.; Ernberg, I.; Kauffman, S. Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* **2009**, *20*, 869–876. [CrossRef] [PubMed]
19. Álvarez-Arenas, A.; Podolski-Renic, A.; Belmonte-Beitia, J.; Pesic, M.; Calvo, G.F. Interplay of Darwinian Selection, Lamarckian Induction and Microvesicle Transfer on Drug Resistance in Cancer. *Sci. Rep.* **2019**, *9*, 9332. [CrossRef] [PubMed]
20. Burrell, R.A.; McGranahan, N.; Bartek, J.; Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **2013**, *501*, 338–345. [CrossRef]

21. Verhaak, R.G.; Hoadley, K.A.; Purdom, E.; Wang, V.; Qi, Y.; Wilkerson, M.D.; Miller, C.R.; Ding, L.; Golub, T.; Mesirov, J.P.; et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **2010**, *17*, 98–110. [CrossRef]
22. Phillips, H.S.; Kharbanda, S.; Chen, R.; Forrest, W.F.; Soriano, R.H.; Wu, T.D.; Misra, A.; Nigro, J.M.; Colman, H.; Soroceanu, L.; et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **2006**, *9*, 157–173. [CrossRef] [PubMed]
23. Jiao, Y.; Killela, P.J.; Reitman, Z.J.; Rasheed, B.A.; Heaphy, C.M.; de Wilde, R.F.; Rodriguez, F.J.; Rosemberg, S.; Oba-Shinjo, S.M.; Marie, S.K.N.; et al. Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas. *Oncotarget* **2012**, *3*, 709–722. [CrossRef] [PubMed]
24. Wang, Q.; Hu, B.; Hu, X.; Kim, H.; Squatrito, M.; Scarpace, L.; de Carvalho, A.C.; Lyu, S.; Li, P.; Li, Y.; et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **2017**, *32*, 42–56.e6. [CrossRef] [PubMed]
25. Sidaway, P. Glioblastoma subtypes revisited. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 587. [CrossRef] [PubMed]
26. Rajapakse, V.N.; Herrada, S.; Lavi, O. Phenotype stability under dynamic brain-tumor environment stimuli maps glioblastoma progression in patients. *Sci. Adv.* **2020**, *6*, aaz4125. [CrossRef] [PubMed]
27. Strauss, B.; Bertolaso, M.; Ernberg, I.; Bissell, M.J. *Rethinking Cancer: A New Paradigm for the Postgenomics Era*; MIT Press: Cambridge, MA, USA, 2021.
28. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **2022**, *12*, 31–46. [CrossRef] [PubMed]
29. Waddington, C. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*; Allen & Unwin: Crows Nest, Australia, 1957.
30. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **2007**, *447*, 425–432. [CrossRef] [PubMed]
31. Guo, S.; Zhu, X.; Huang, Z.; Wei, C.; Yu, J.; Zhang, L.; Feng, J.; Li, M.; Li, Z. Genomic instability drives tumorigenesis and metastasis and its implications for cancer therapy. *Biomed. Pharmacother.* **2023**, *157*, 114036. [CrossRef] [PubMed]
32. Greaves, M.; Maley, C.C. Clonal evolution in cancer. *Nature* **2012**, *481*, 306–313. [CrossRef]
33. Biswas, S.; Rounak, A.; Perlikowski, P.; Gupta, S. Characterising stochastic fixed points and limit cycles for dynamical systems with additive noise. *Commun. Nonlinear Sci. Numer. Simul.* **2021**, *101*, 105870. [CrossRef]
34. Wang, J.; Xu, L.; Wang, E.; Huang, S. The Potential Landscape of Genetic Circuits Imposes the Arrow of Time in Stem Cell Differentiation. *Biophys. J.* **2010**, *99*, 29–39. [CrossRef]
35. Wang, J.; Zhang, K.; Xu, L.; Wang, E. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 8257–8262. [CrossRef]
36. Ferrell, J.E. Bistability, Bifurcations, and Waddington's Epigenetic Landscape. *Curr. Biol.* **2012**, *22*, R458–R466. [CrossRef]
37. Li, C.; Wang, J. Quantifying Cell Fate Decisions for Differentiation and Reprogramming of a Human Stem Cell Network: Landscape and Biological Paths. *PLoS Comput. Biol.* **2013**, *9*, e1003165. [CrossRef]
38. Li, C.; Wang, J. Quantifying the underlying landscape and paths of cancer. *J. R. Soc. Interface* **2014**, *11*, 20140774. [CrossRef]
39. Li, C.; Wang, J. Quantifying the Landscape for Development and Cancer from a Core Cancer Stem Cell Circuit. *Cancer Res.* **2015**, *75*, 2607–2618. [CrossRef]
40. Verd, B.; Crombach, A.; Jaeger, J. Classification of transient behaviours in a time-dependent toggle switch model. *BMC Syst. Biol.* **2014**, *8*, 43. [CrossRef]
41. Saelens, W.; Cannoodt, R.; Todorov, H.; Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **2019**, *37*, 547–554. [CrossRef]
42. Witkiewicz, A.K.; Kumarasamy, V.; Sanidas, I.; Knudsen, E.S. Cancer cell cycle dystopia: Heterogeneity, plasticity, and therapy. *Trends Cancer* **2022**, *8*, 711–725. [CrossRef]
43. Moore, C.C. Ergodic theorem, ergodic theory, and statistical mechanics. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 1907–1911. [CrossRef]
44. Palmer, R. Broken ergodicity. *Adv. Phys.* **1982**, *31*, 669–735. [CrossRef]
45. Mauro, J.C.; Smedskjaer, M.M. Statistical mechanics of glass. *J. Non-Cryst. Solids* **2014**, *396-397*, 41–53. [CrossRef]
46. Gillespie, D.T. The chemical Langevin equation. *J. Chem. Phys.* **2000**, *113*, 297–306. [CrossRef]
47. Santillán, M. On the Use of the Hill Functions in Mathematical Models of Gene Regulatory Networks. *Math. Model. Nat. Phenom.* **2008**, *3*, 85–97. [CrossRef]
48. da Costa, A.A.B.A.; Chowdhury, D.; Shapiro, G.I.; D'Andrea, A.D.; Konstantinopoulos, P.A. Targeting replication stress in cancer therapy. *Nat. Rev. Drug Discov.* **2022**, *22*, 38–58. [CrossRef]
49. Bioconductor.org. RNA-seq Workflow: Gene-Level Exploratory Analysis and Differential Expression, 2023. Available online: https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html (accessed on 19 July 2022).
50. Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, *20*, 296. [CrossRef] [PubMed]
51. Wolfram Research, Inc. KMeans, 2023. Available online: https://reference.wolfram.com/language/ref/method/KMeans.html (accessed on 9 April 2023).

52. Wolfram Research, Inc. Neighborhood Contraction, 2023. Available online: https://reference.wolfram.com/language/ref/method/NeighborhoodContraction.html (accessed on 9 April 2023).
53. Wolfram Research, Inc. Gaussian Mixture, 2023. Available online: https://reference.wolfram.com/language/ref/method/GaussianMixture.html (accessed on 9 April 2023).
54. Sáez, M.; Blassberg, R.; Camacho-Aguilar, E.; Siggia, E.D.; Rand, D.A.; Briscoe, J. Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions. *Cell Syst.* **2022**, *13*, 12–28.e3. [CrossRef]
55. Weinreb, C.; Wolock, S.; Tusi, B.K.; Socolovsky, M.; Klein, A.M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E2467. [CrossRef]
56. Coomer, M.A.; Ham, L.; Stumpf, M.P. Noise distorts the epigenetic landscape and shapes cell-fate decisions. *Cell Syst.* **2022**, *13*, 83–102.e6. [CrossRef]
57. Li, C.; Wang, J. Quantifying Waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. *J. R. Soc. Interface* **2013**, *10*, 20130787. [CrossRef]
58. Voit, E.O. Perspective: Dimensions of the scientific method. *PLoS Comput. Biol.* **2019**, *15*, e1007279. [CrossRef] [PubMed]
59. Meister, A.; Du, C.; Li, Y.H.; Wong, W.H. Modeling stochastic noise in gene regulatory systems. *Quant. Biol.* **2014**, *2*, 1–29. [CrossRef] [PubMed]
60. Darmanis, S.; Sloan, S.A.; Croote, D.; Mignardi, M.; Chernikova, S.; Samghababi, P.; Zhang, Y.; Neff, N.; Kowarsky, M.; Caneda, C.; et al. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **2017**, *21*, 1399–1410. [CrossRef] [PubMed]
61. Clarivate Analytics. MetaCore, 2019. Available online: https://portal.genego.com (accessed on 16 April 2022).
62. Vieira, M. *Gene Expression Network Analysis, 2023*; GitHub Repository. Available online: https://github.com/marcosgvjunior/gene-expression-network-analysis (accessed on 16 April 2022).
63. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: https://www.R-project.org/ (accessed on 16 April 2022).
64. Satija, R.; Farrell, J.A.; Gennert, D.; Schier, A.F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, *33*, 495–502. [CrossRef] [PubMed]
65. Lab, S. Using Sctransform in Seurat, 2022. GitHub Repository. Available online: https://satijalab.org/seurat/articles/sctransform_vignette.html (accessed on 17 July 2022).
66. Vieira, M. *Graph Matrix and Combinatorics, 2023*; GitHub Repository. https://github.com/marcosgvjunior/graph-matrix-and-combinatorics (accessed on 17 July 2022).
67. Mauro, J.C.; Gupta, P.K.; Loucks, R.J. Continuously broken ergodicity. *J. Chem. Phys.* **2007**, *126*, 184511. [CrossRef] [PubMed]
68. Wolfram Research, Inc. *Mathematica*, Version 13.1; Mathematica: Champaign, IL, USA, 2022.
69. Wang, J. Landscape and flux theory of non-equilibrium dynamical systems with application to biology. *Adv. Phys.* **2015**, *64*, 1–137. [CrossRef]
70. Strogatz, S.H. Exploring complex networks. *Nature* **2001**, *410*, 268–276. [CrossRef] [PubMed]
71. Wolfram Research, Inc. Constrained Optimization, 2023. Available online: https://library.wolfram.com/infocenter/Books/8506/ConstrainedOptimization.pdf (accessed on 12 July 2022).
72. Uthamacumaran, A. A review of dynamical systems approaches for the detection of chaotic attractors in cancer networks. *Patterns* **2021**, *2*, 100226. [CrossRef]
73. Chen, C.; Wang, J. A physical mechanism of cancer heterogeneity. *Sci. Rep.* **2016**, *6*, 20679. [CrossRef]
74. Nijman, S.M. Perturbation-Driven Entropy as a Source of Cancer Cell Heterogeneity. *Trends Cancer* **2020**, *6*, 454–461. [CrossRef]
75. Tarabichi, M.; Antoniou, A.; Saiselet, M.; Pita, J.M.; Andry, G.; Dumont, J.E.; Detours, V.; Maenhaut, C. Systems biology of cancer: Entropy, disorder, and selection-driven evolution to independence, invasion and "swarm intelligence". *Cancer Metastasis Rev.* **2013**, *32*, 403–421. [CrossRef] [PubMed]
76. Wolfram Research, Inc. Correlation Function. 2023. Available online: https://reference.wolfram.com/language/ref/CorrelationFunction.html (accessed on 9 April 2023).
77. Perko, L. *Differential Equations and Dynamical Systems*; Springer: New York, NY, USA, 2001.
78. Wiggins, S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*; Texts in Applied Mathematics; Springer: New York, NY, USA, 2003.
79. Strogatz, S. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2018.
80. Guckenheimer, J.; Holmes, P. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*; Springer: New York, NY, USA, 1983.
81. Alligood, K.; Sauer, T.; Yorke, J. *Chaos: An Introduction to Dynamical Systems*; Textbooks in Mathematical Sciences; Springer: New York, NY, USA, 2000.
82. Pitzer, E.; Affenzeller, M.; Beham, A. A closer look down the basins of attraction. In Proceedings of the 2010 UK Workshop on Computational Intelligence (UKCI), Colchester, UK, 8–10 September 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1–6. [CrossRef]
83. Kampen. *Stochastic Processes in Physics and Chemistry*; Elsevier: Amsterdam, The Netherlands; Boston, MA, USA; London, UK, 2007.

84. Bover, D. Moment equation methods for nonlinear stochastic systems. *J. Math. Anal. Appl.* **1978**, *65*, 306–320. [CrossRef]
85. Gardiner, C. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*; Springer complexity; Springer: Berlin, Germany, 2004.