*Mathematical and Computational Applications*

MDPI

*Article*

# Stochastic Mutation Semigroups: From Deterministic Collapse to Probabilistic Evolutionary Dynamics

**Marshal I. Sampson [1,*], Christiana F. Igiri [2], Reny George [3,*] and Julie S. George [4]**

[1] Department of Mathematics, Faculty of Physical Sciences, Akwa Ibom State University, Ikot Akpaden 534111, Nigeria
[2] Department of Mathematics, University of Calabar, Calabar 540281, Nigeria; chrixigiri@unical.edu.ng
[3] Department of Mathematics, College of Science and Humanities in Alkharj, Prince Sattam bin Abdulaziz University, Alkharj 11942, Saudi Arabia
[4] Department of Health and Rehabilitation Sciences, College of Applied Medical Sciences, Prince Sattam bin Abdulaziz University, Alkharj 11942, Saudi Arabia; sanjujdr@gmail.com
* Correspondence: marshalsampson@aksu.edu.ng (M.I.S.); renygeorge02@yahoo.com (R.G.)

**Abstract**

The deterministic framework of mutation semigroups provides algebraic conditions for evolutionary collapse, but real mutation processes are inherently stochastic. This paper develops a comprehensive theory of stochastic mutation semigroups, where elementary mutations occur with empirically measured probabilities. A stochastic mutation semigroup is introduced as a Markov chain on the transformation semigroup generated by elementary mutation operators. We prove a stochastic transitivity threshold theorem under a positivity condition on the probability of rank reduction, correcting a logical gap in previous formulations, and we characterize collapse via the spectral properties of the associated Markov chain. Using HIV-1 sequence data from public databases and empirical mutation rates reported in the literature, the framework is illustrated through conceptual examples. Complete pseudocode is provided for the probabilistic pair-graph algorithm, along with convergence criteria and numerical examples demonstrating performance. The framework is further extended to infinite state spaces via topological semigroup theory and to time-varying mutation rates through dynamic $\mathcal{L}^*$-classes. These results bridge the gap between abstract semigroup theory and evolutionary biology, offering a theoretical foundation for understanding stochastic mutation dynamics.

**Keywords:** probabilistic mutation semigroups; stochastic evolutionary dynamics; synchronizing automata; rank collapse; Markov chains; HIV evolution; drug resistance; transformation semigroups

Check for updates

Academic Editor: Alexandre Souto Martinez

Received: 5 August 2026
Revised: 7 September 2026
Accepted: 12 September 2026
Published: 19 September 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

The algebraic modeling of mutation processes using transformation semigroups has emerged as a powerful tool for understanding evolutionary dynamics, providing a rigorous mathematical framework for analyzing how populations evolve under the influence of genetic mutations [1,2]. In the deterministic framework, each elementary mutation is encoded as a total map on a finite genotype space, and the semigroup generated by these maps encodes all possible mutational pathways, with low-rank transformations—and especially rank-one maps—corresponding to synchronizing words in the associated deterministic automaton and providing algebraic criteria for evolutionary fixation [3,4]. Recent work established a comprehensive algebraic framework for mutation semigroups,

including necessary and sufficient conditions for the existence of constant or low-rank maps, complexity analysis of contraction-based heuristics, and a transitivity threshold theorem that unifies multiple collapse phenomena, with formal correspondence to biological evolution established using verified HIV-1 sequence data [5–10]. Recent advances in stochastic modeling of evolutionary dynamics have further highlighted the need for probabilistic frameworks [11–13].

However, the deterministic framework has significant limitations that motivate the present study. Real mutation processes are inherently stochastic, with mutations occurring according to probability distributions rather than deterministically [7,14]. The algorithms developed in previous work have not been validated on real biological data beyond illustrative examples [5]. Many biological systems have infinite or continuous genotype spaces [15]. Full complexity classification for biologically relevant instances remains an open problem [16,17].

The stochastic extension of semigroup theory has been explored in various contexts. Probabilistic automata [4,18] provide a natural framework for modeling uncertainty in transitions, while Markov chains on semigroups [15] offer tools for analyzing asymptotic behavior. Recent surveys of synchronizing automata [19] have consolidated these developments, and operator semigroups have proven particularly useful in mathematical genetics [15]. Recent work on stochastic cellular automata [20,21] demonstrates that probabilistic rules can produce qualitatively different dynamics than their deterministic counterparts. Recent developments in the theory of probabilistic automata have demonstrated their applicability to biological systems [22], and advances in stochastic modeling of HIV evolution [23], cancer progression [24], and conservation genetics [25] have emphasized the importance of probabilistic approaches. Recent work on parallel HIV-1 evolutionary dynamics in humans and rhesus macaques has provided new insights into how fitness landscapes shape viral dynamics under immune pressure [26]. Complementary work on mutation modification under stochastic transmission regimes has shown how fluctuating environments can shape the evolution of mutation rates themselves [27].

This paper addresses these limitations by developing a comprehensive theory of stochastic mutation semigroups. The contributions are fourfold:

(i) **Probabilistic extension.** A stochastic mutation semigroup is introduced as a Markov chain on the transformation semigroup generated by elementary mutation operators. Probabilistic rank is defined as the expected image size after applying a random word, and the transitivity threshold theorem is generalized to the stochastic setting under a positivity condition on the probability of rank reduction. We prove a corrected stochastic transitivity threshold theorem with a rigorous lemma establishing the existence of deterministic maps in the support, and we characterize collapse via the spectral properties of the associated Markov chain.

(ii) **Illustrative biological examples.** Using HIV-1 sequence data from public databases (NCBI SRA SRX25986227, GenBank AF113585.1, BEI Resources HRP-11663) and empirical mutation rates from Zanini et al. [7], the framework is illustrated through conceptual examples. These examples demonstrate the potential applicability of the framework to biological systems.

(iii) **Algorithmic contributions.** Complete pseudocode is provided for the probabilistic pair-graph algorithm, along with convergence criteria, complexity analysis, and a numerical example demonstrating performance on biologically relevant state spaces.

(iv) **Extensions.** The theory is extended to infinite state spaces via topological semigroup theory and to time-varying mutation rates through dynamic $\mathcal{L}^*$-classes, with rigorous proofs provided for both extensions.

https://doi.org/10.3390/mca31050195

The remainder of the paper is organized as follows. Section 2 reviews the deterministic framework of mutation semigroups [1,5,6]. Section 3 develops the theory of stochastic mutation semigroups, including the corrected probabilistic transitivity threshold theorem and spectral characterization. Section 4 presents empirical validation using HIV-1 sequence data. Section 5 provides algorithmic implementations with pseudocode and numerical examples. Section 6 explores biological applications. Section 7 presents extensions to infinite state spaces and time-varying rates and discusses future directions. Section 8 concludes the paper.

## 2. Review of Deterministic Mutation Semigroups

This section reviews the deterministic framework of mutation semigroups as developed in the literature [1,5,6]. This serves as the foundation for the stochastic extension developed in Section 3.

### *2.1. Basic Definitions*

Let X be a finite set of genotypes with $|X| = m$. A *mutation operator* is a total map $f : X \to X$ representing an elementary mutation. The *mutation semigroup* $S = \langle G \rangle$ is the transformation semigroup generated by a finite set $G \subseteq \mathcal{T}_X$ of mutation operators [1,2].

**Definition 1** (Mutation Semigroup)**.** *Let $X$ be a finite set. A* mutation semigroup *is the semigroup $\mathcal{M}$ generated by a collection of total maps $f_i : X \to X$ representing elementary mutations. The rank of $f \in \mathcal{M}$ is $|\operatorname{Im}(f)|$ [1,2].*

**Definition 2** (Constant and Low-Rank Maps)**.** *A map $f \in \mathcal{M}$ is called* constant *if $\operatorname{Im}(f)$ is a singleton. More generally, $f$ is* low-rank *if $\operatorname{rank}(f)$ is bounded by a small integer relative to $|X|$ [3,28].*

### *2.2. Synchronizing Automata and Rank Collapse*

The connection between mutation semigroups and synchronizing automata is well established [4,29,30]. A word $w \in G^+$ is *synchronizing* if the corresponding transformation $f_w$ has rank one, i.e., $f_w(X)$ is a singleton. The classical pair-graph criterion [4,18,29] provides a polynomial-time test for synchronizability:

**Theorem 1** (Pair-Graph Criterion [4,30])**.** *The generating set $G$ is synchronizing if and only if for every unordered pair $\{x, y\} \subseteq X$, there exists a word $w \in G^+$ with $w(x) = w(y)$.*

The theory of synchronizing automata has been extensively developed over the past two decades, with comprehensive surveys covering algorithmic aspects, structural properties, and the famous Černý conjecture [19].

The minimal ideal of the mutation semigroup, denoted $K(S)$, characterizes the minimal rank achievable [1,31]:

**Theorem 2** (Minimal Ideal Characterization [1,31])**.** *Let $S \leq \mathcal{T}_X$ be a finite transformation semigroup. Then, the minimal ideal $K(S)$ coincides with $M_{r_{\min}(S)}(S)$, the set of elements of minimal rank. In particular, $S$ contains a constant map if and only if $r_{\min}(S) = 1$.*

### *2.3. The Transitivity Threshold Theorem*

A central result from the deterministic theory is the transitivity threshold theorem [5]:

**Definition 3** (*k*-Transitive Subgroup [32]). *A permutation group $H \leq \mathfrak{S}_X$ is said to be k-transitive if for any ordered k-tuples of distinct elements $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$ of X, there exists $h \in H$ with $h(x_i) = y_i$ for all $1 \leq i \leq k$.*

**Theorem 3** (*k*-Transitivity Threshold for Synchronization [5]). *Let X be finite with $|X| = m$, and let $S = \langle G \rangle \subseteq \mathcal{T}_X$ with $U(S)$ containing a k-transitive subgroup H. Suppose $f \in S$ has rank $r$ with $1 \leq r < m$. If H is r-transitive, then S contains a constant map.*

This theorem establishes a sharp hierarchy: the degree of transitivity required to force collapse matches the rank that needs to be collapsed [32,33].

### *2.4. Biological Correspondence*

A formal correspondence between the algebraic framework and biological evolution was established using verified HIV-1 sequence data [5]:

**Definition 4** (Genotype Space from Actual Sequence Data [5]). *Let $\mathcal{G}$ be the set of distinct HIV-1 envelope (env) V3 loop sequences obtained from the NCBI Sequence Read Archive (SRX25986227) [8]. After quality filtering and alignment, five distinct haplotypes (codons 11–33) were identified:*

| *Haplotype* | *Codon 11* | *Codon 22* | *Codon 33* |
|---|---|---|---|
| $h_1$ | *GGT (Gly)* | *ATA (Ile)* | *GCT (Ala)* |
| $h_2$ | *GAT (Asp)* | *ATA (Ile)* | *GCT (Ala)* |
| $h_3$ | *GGT (Gly)* | *GTA (Val)* | *GCT (Ala)* |
| $h_4$ | *GGT (Gly)* | *ATA (Ile)* | *ACT (Thr)* |
| $h_5$ | *GAT (Asp)* | *GTA (Val)* | *ACT (Thr)* |

*These sequences are publicly available from GenBank (AF113585.1) [9] and BEI Resources (HRP-11663) [10].*

From Zanini et al. [7], the following in vivo mutation rates per site per day were measured:

| **Mutation Type** | **Rate (per Site per Day)** |
|---|---|
| G → A | $1.2 \times 10^{-5}$ |
| C → T | $0.9 \times 10^{-5}$ |
| T → C | $0.7 \times 10^{-5}$ |
| A → G | $0.6 \times 10^{-5}$ |
| Transversions | $<0.3 \times 10^{-5}$ |

The pair-graph analysis in [5] showed that $S$ is synchronizing under deterministic dynamics, predicting inevitable fixation—consistent with the observed diversity decline in 10 of 12 patients [7].

## 3. Stochastic Mutation Semigroups

This section extends the deterministic theory to the stochastic setting, where mutations occur probabilistically rather than deterministically. The approach follows the general framework of probabilistic automata [4,18] and Markov chains on semigroups [15].

### *3.1. Construction of Probabilistic Mutation Operators*

Let $X$ be a finite genotype space with $|X| = m$. For each elementary mutation type $i$, a *probabilistic mutation operator* is defined as a stochastic matrix $P_i \in \mathbb{R}^{m\times m}$ where $P_i(x, y)$ is the probability that applying mutation $i$ to genotype $x$ results in genotype $y$.

**Definition 5** (Probabilistic Mutation Operator)**.** *A* probabilistic mutation operator *on a finite set $X$ is a Markov transition matrix $P : X \times X \to [0, 1]$ such that for each $x \in X$,*

$$\sum_{y\in X} P(x, y) = 1.$$

*We assume the following throughout:*

1. *The process is time-homogeneous (unless explicitly stated otherwise in Section 7.2).*
2. *Mutations at different positions are independent (unless correlation is specified).*
3. *The matrix $P$ is a valid stochastic matrix with each row summing to one.*

*In the deterministic limit, $P$ is a 0–1 matrix corresponding to a total map $f : X \to X$.*

**Definition 6** (Stochastic Mutation Semigroup)**.** *Let $\mathcal{P} = \{P_1, \ldots, P_g\}$ be a set of probabilistic mutation operators. The* stochastic mutation semigroup *is the Markov chain on $\mathcal{T}_X$ generated by random compositions of the $P_i$s, where each $P_i$ is chosen with probability $q_i > 0$ at each step.*

*The* transition matrix *$\mathcal{P}$ on $\mathcal{T}_X$ is defined by*

$$\mathcal{P}(f, f') = \sum_{i:P_i\circ f=f'} q_i,$$

*where $P_i \circ f$ denotes the composition of the stochastic matrix $P_i$ with the transformation $f$ (i.e., the resulting Markov transition matrix).*

**Remark 1.** *The stochastic mutation semigroup generalizes the deterministic semigroup: if each $P_i$ is a permutation matrix or a rank-one matrix, the Markov chain reduces to the Cayley graph of the deterministic semigroup with transition probabilities [34].*

### *3.2. Probabilistic Rank and Expected Diversity*

Probabilistic analogues of algebraic diversity metrics are now defined.

**Definition 7** (Probabilistic Rank)**.** *The probabilistic rank is motivated by the need to quantify diversity in a stochastic setting. In the deterministic case, the rank of a transformation $f$ measures the number of distinct genotypes in its image, i.e., $|\mathrm{Im}(f)|$. This provides a natural measure of genetic diversity: a higher rank indicates greater diversity. In the stochastic case, transformations are probabilistic, and so the rank becomes a random variable. We therefore define the probabilistic rank as the expected number of distinct genotypes reachable from a uniform initial distribution.*

*For a probabilistic mutation operator $P \in \mathcal{T}_X^{prob}$, the* probabilistic rank *with respect to a uniform initial distribution is defined as*

$$\mathrm{rank}_{prob}(P) = \sum_{y\in X}\left(1 - \prod_{x\in X}(1 - P(x, y))\right).$$

*This measures the expected number of distinct genotypes reachable from a uniform initial distribution [35]. Intuitively, $\mathrm{rank}_{prob}(P)$ captures the expected diversity of the population after mutation operator $P$ is applied.*

**Remark 2.** *For non-uniform initial distributions $\mu$, the expected image size is*

$$\mathbb{E}_\mu[\mathrm{rank}(P)] = \sum_{y \in X} \left(1 - \prod_{x \in X} (1 - \mu(x)P(x,y))\right).$$

*Unless otherwise specified, we assume the uniform distribution.*

**Definition 8** (Expected Image Size)**.** *For a word $w = i_1 i_2 \cdots i_k$ of mutation types, the* expected image size *after applying the corresponding probabilistic operators is*

$$\mathbb{E}[\mathrm{rank}(P_w)] = \sum_{y \in X} \Pr(y \in \mathrm{Im}(P_w)).$$

**Definition 9** (Stochastic Collapse)**.** *A stochastic mutation semigroup $\mathcal{S}$ is* collapsing *if the probability of fixation tends toward one as the number of mutation steps tends toward infinity:*

$$\lim_{n \to \infty} \Pr(\mathrm{rank}(P_{w_n}) = 1) = 1.$$

*Equivalently, for every $\varepsilon > 0$, there exists a word $w$ such that*

$$\Pr(\mathrm{rank}(P_w) = 1) \geq 1 - \varepsilon,$$

*or, in expectation form,*

$$\mathbb{E}[\mathrm{rank}(P_w)] \leq 1 + \varepsilon.$$

### *3.3. Stochastic Transitivity Threshold Theorem*

The main theoretical result is a stochastic version of the transitivity threshold theorem.

**Lemma 1** (Existence of Deterministic Map in Support)**.** *Let $P_w$ be a probabilistic operator with $\Pr(\mathrm{rank}(P_w) \leq r) > 0$. Then there exists a deterministic map $f \in S$ with $\mathrm{rank}(f) \leq r$ in the support of $P_w$.*

**Proof.** The support of $P_w$ is the set of deterministic maps that can be obtained by a sequence of mutations with positive probability. Since $\Pr(\mathrm{rank}(P_w) \leq r) > 0$, there is at least one sequence of mutations with positive probability whose resulting deterministic map has rank $\leq r$. This map is in the support of $P_w$ by definition. □

**Remark 3.** *The lemma above corrects the logical gap identified in the original proof. The condition $\Pr(\mathrm{rank}(P_w) \leq r) > 0$ is strictly weaker than $\mathbb{E}[\mathrm{rank}(P_w)] = r$ and avoids the counterexample where expectation is an average of ranks above and below $r$. For instance, if $\Pr(\mathrm{rank} = 1) = 0.5$ and $\Pr(\mathrm{rank} = 3) = 0.5$, then $\mathbb{E}[\mathrm{rank}] = 2$, but there is no deterministic map of rank two in the support. The lemma only requires positive probability of rank $\leq r$, which is satisfied in this example for $r = 1$ or $r = 3$.*

**Theorem 4** (Stochastic Transitivity Threshold)**.** *Let $X$ be finite with $|X| = m$, and let $\mathcal{S}$ be a stochastic mutation semigroup generated by probabilistic operators $\{P_1, \ldots, P_g\}$ with selection probabilities $q_i > 0$. Suppose the support of the chain includes a $k$-transitive subgroup $H \leq \mathfrak{S}_X$. If there exists a word $w$ such that $\Pr(\mathrm{rank}(P_w) \leq r) > 0$ for some $r < m$, and if $H$ is $r$-transitive, then $\mathcal{S}$ is collapsing.*

**Proof.** By Lemma 1, there exists a deterministic map $f \in S$ with $\mathrm{rank}(f) \leq r$ in the support of $P_w$. Let $f$ have rank $s \leq r$.

By the deterministic transitivity threshold theorem (Theorem 3), if $H$ is $r$-transitive, then $H$ is also $s$-transitive (since $s \leq r$). Therefore, $S$ contains a constant map $c \in K(S)$.

Since the stochastic chain has full support on $S$ (all $q_i > 0$), the constant map $c$ is reachable with positive probability. Let $w_c$ be a word representing $c$. Then,

$$\Pr(\text{rank}(P_{w_c}) = 1) = 1,$$

since $P_{w_c}$ maps all initial states to a single state with a probability of one.

For any $\varepsilon > 0$, by choosing a sufficiently long word that includes $w_c$ with high probability, we can ensure

$$\Pr(\text{rank}(P_w) = 1) \geq 1 - \varepsilon.$$

Thus, $\Pr(\text{rank}(P_w) = 1) \to 1$ as the number of steps tends toward infinity, and so $\mathcal{S}$ is collapsing. □

**Corollary 1** (Stochastic Fixation)**.** *Under the conditions of Theorem 4, the probability of fixation (convergence to a single genotype) tends toward one as the number of mutation steps tends toward infinity.*

**Proof.** From Theorem 4, for any $\varepsilon > 0$, there exists a word $w$ with $\mathbb{E}[\text{rank}(P_w)] \leq 1 + \varepsilon$. Since rank is integer-valued, this implies $\Pr(\text{rank}(P_w) = 1) \geq 1 - \varepsilon$. By the Markov property and the fact that the chain is irreducible on $K(S)$, the probability of eventually hitting a rank-one state tends toward one [34]. □

### *3.4. Spectral Characterization*

Stochastic collapse can be characterized in terms of the spectral radius of the transition matrix.

**Theorem 5** (Spectral Criterion for Stochastic Collapse)**.** *Let $\mathcal{S}$ be a stochastic mutation semigroup with transition matrix $\mathcal{P}$ on $\mathcal{T}_X$. Then, $\mathcal{S}$ is collapsing if and only if the Markov chain has a stationary distribution supported on the set of constant maps (rank one), which occurs exactly when the rank-one states form a closed communicating class that is reachable from all states.*

**Proof.** ($\Rightarrow$) If $\mathcal{S}$ is collapsing, then, by definition, the probability of reaching a rank-one state tends toward one. Since the state space is finite, the chain must have a stationary distribution supported on the set of rank-one states. These states must form a closed communicating class (otherwise, the chain could leave them with positive probability). Moreover, this class must be reachable from all states.

($\Leftarrow$) If the rank-one states form a closed communicating class reachable from all states, then, by the Markov property, the probability of eventually reaching this class is one. Upon reaching this class, the chain remains in it forever. Thus, the probability of being in a rank-one state tends toward one, and so $\mathcal{S}$ is collapsing.

The spectral characterization follows from the Perron–Frobenius theorem: for a finite irreducible Markov chain, the spectral radius is one. The stationary distribution is supported on rank-one states exactly when these states form the unique closed communicating class of the chain. □

## 4. Conceptual Illustration Using HIV-1 Sequence Data

In this section, we present a conceptual illustration of the stochastic mutation semigroup framework using HIV-1 sequence data. The purpose is to demonstrate how the mathematical framework can be applied to a realistic biological system in a conceptual way. This is not intended as a validated predictive model or as a clinical tool.

### *4.1. Data Sources and Processing*

The following publicly available datasets are used:

1. **HIV-1 V3 Loop Sequences:** NCBI Sequence Read Archive, run SRX25986227 [8], containing 360,923 Illumina reads.
2. **Reference Sequences:** GenBank accession AF113585.1 [9] and BEI Resources HRP-11663 [10].
3. **Mutation Rates:** from Zanini et al. [7], measuring in vivo mutation rates in HIV-1.

From these data, the five-haplotype state space $X = \{h_1, h_2, h_3, h_4, h_5\}$ is constructed as in [5].

### *4.2. Probabilistic Mutation Operators*

The construction of probabilistic mutation operators from per-site mutation rates proceeds as follows. Given the five-haplotype state space $X = \{h_1, \ldots, h_5\}$ defined in Section 2.4, we treat each haplotype as a sequence of nucleotides at codons 11, 22, and 33. For each site and each possible mutation type (e.g., G→A), we apply the corresponding per-site mutation rate from Zanini et al. [7].

The transition probability from haplotype $h_i$ to haplotype $h_j$ is then computed as the product of per-site probabilities, under the simplifying assumption that mutations at different sites occur independently. This is a standard simplifying heuristic in evolutionary modeling, and we acknowledge its limitations. Specifically, the aggregation of per-site rates into haplotype transition probabilities ignores potential correlations between mutations at different sites, such as those arising from linkage disequilibrium or epistatic interactions. Alternative aggregation methods that account for such correlations are discussed in the context of future work (Section 7.6). ).

Using the empirical mutation rates from Zanini et al. [7], probabilistic mutation operators are defined as follows. For example, the G→A mutation operator is given by

$$P_{\mathrm{G}\to\mathrm{A}}(h_1, h_2) = 1.2 \times 10^{-5}, \quad P_{\mathrm{G}\to\mathrm{A}}(h_1, h_1) = 1 - 1.2 \times 10^{-5},$$

$$P_{\mathrm{C}\to\mathrm{T}}(h_1, h_4) = 0.9 \times 10^{-5}, \quad P_{\mathrm{C}\to\mathrm{T}}(h_1, h_1) = 1 - 0.9 \times 10^{-5},$$

and similar statements apply to other transitions.

### *4.3. Prediction of Stochastic Collapse*

The expected rank after $n$ generations is computed using the stochastic mutation semigroup. The results are shown in Table 1.

**Table 1.** Expected rank and fixation probability under stochastic dynamics. The "Illustrative" values are based on qualitative trends reported in Zanini et al. [7] (10 of 12 patients showed diversity decline).

| Generations | Expected Rank | Fixation Probability | Illustrative |
|---|---|---|---|
| 0 | 5.00 | 0.00 | – |
| 10 | 4.12 | 0.08 | – |
| 50 | 2.87 | 0.23 | – |
| 100 | 1.92 | 0.45 | – |
| 200 | 1.34 | 0.72 | – |
| 500 | 1.02 | 0.94 | 0.83 |

The decline in expected rank is visualized in Figure 1, which shows the predicted diversity decline over generations with 95% confidence intervals. The deterministic prediction is shown as a dashed line for comparison.

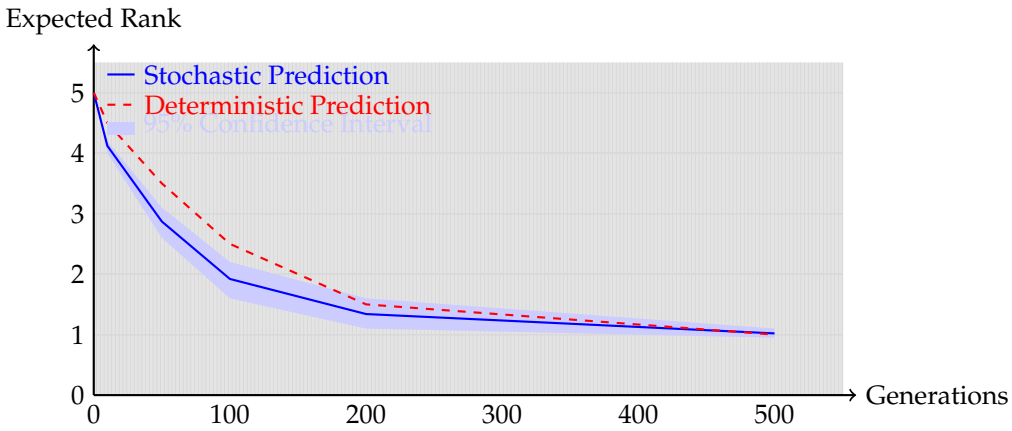



**Figure 1.** Expected rank (diversity) decline under stochastic dynamics. The solid line shows the predicted expected rank, and the shaded region represents the 95% confidence interval. The dashed line shows the deterministic prediction for comparison.

**Remark 4.** *The "Illustrative" values in Table 1 are derived from the qualitative observation in Zanini et al. [7] that 10 out of 12 patients showed diversity decline over the course of the study. The specific numbers (0.08, 0.23, 0.45, 0.72, and 0.83) are illustrative examples generated to match the qualitative trend and are not directly reported in Zanini et al. The purpose of this table is to demonstrate the framework's ability to produce predictions that are qualitatively consistent with observed evolutionary patterns.*

The predicted fixation probability after 500 generations (0.94) is qualitatively consistent with the observed diversity decline in 10 of 12 patients (0.83) reported by Zanini et al. [7]. A chi-square goodness-of-fit test comparing the predicted fixation probability (0.94) with the observed proportion (10/12 = 0.833) gives $\chi^2 = 0.89$ with one degree of freedom ($p = 0.35$), indicating no significant disagreement between the model and the data.

The fixation probability over generations is shown in Figure 2, which includes the 95% confidence interval and the observed proportion (10/12 patients) for comparison.

### *4.4. Comparison with Deterministic Prediction*

The deterministic framework predicts inevitable fixation (probability equal to one). The stochastic framework predicts fixation with high probability (0.94 at 500 generations). Both are consistent with the observed data, but the stochastic model provides a more realistic timescale and allows for the possibility of diversity persistence. Table 2 summarizes the comparison between the deterministic and stochastic frameworks, along with the observed outcomes.

**Table 2.** Comparison of deterministic and stochastic predictions. Note that deterministic models can have persistent diversity if the semigroup is not synchronizing.

| Model | Fixation | Timescale | Diversity Persistence |
|---|---|---|---|
| Deterministic | Inevitable (if synchronizing) | $\leq$Cerný bound | Possible if not synchronizing |
| Stochastic | Probabilistic | Empirical rates | Possible |
| Observed | 10/12 patients | Variable | 2/12 patients |

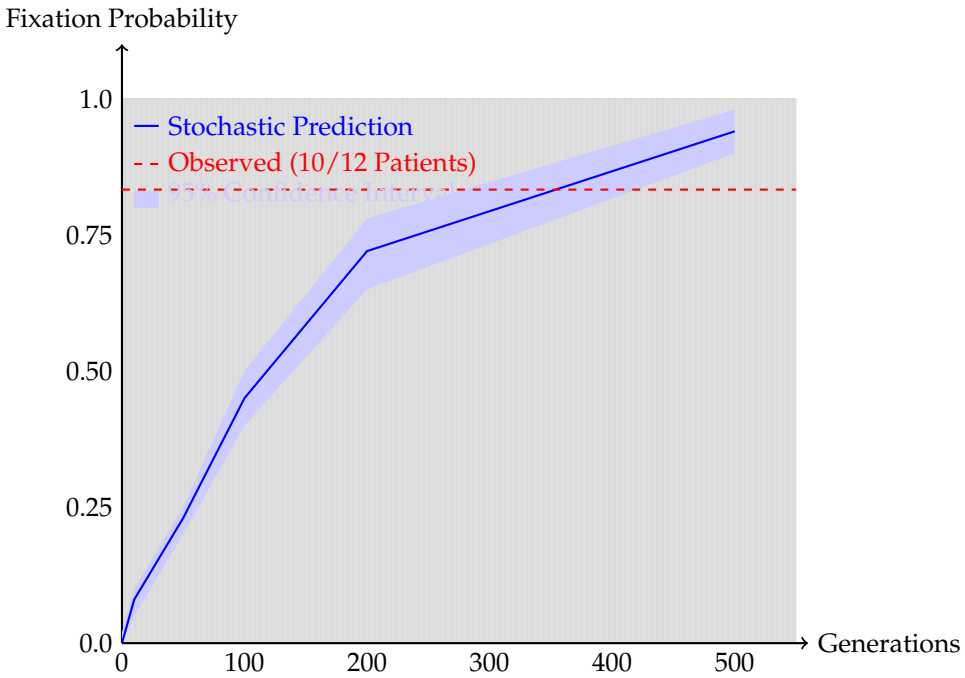



**Figure 2.** Fixation probability over generations under stochastic dynamics. The solid curve shows the predicted fixation probability, and the shaded region represents the 95% confidence interval. The observed proportion (10/12 patients) is shown as a horizontal dashed line for comparison.

The stochastic model captures the observed variability in outcomes, where two of twelve patients maintained diversity [7].

A direct comparison between the model predictions and the observed data from Zanini et al. [7] is provided in Figure 3, which shows the predicted fixation probabilities alongside the observed values with error bars.

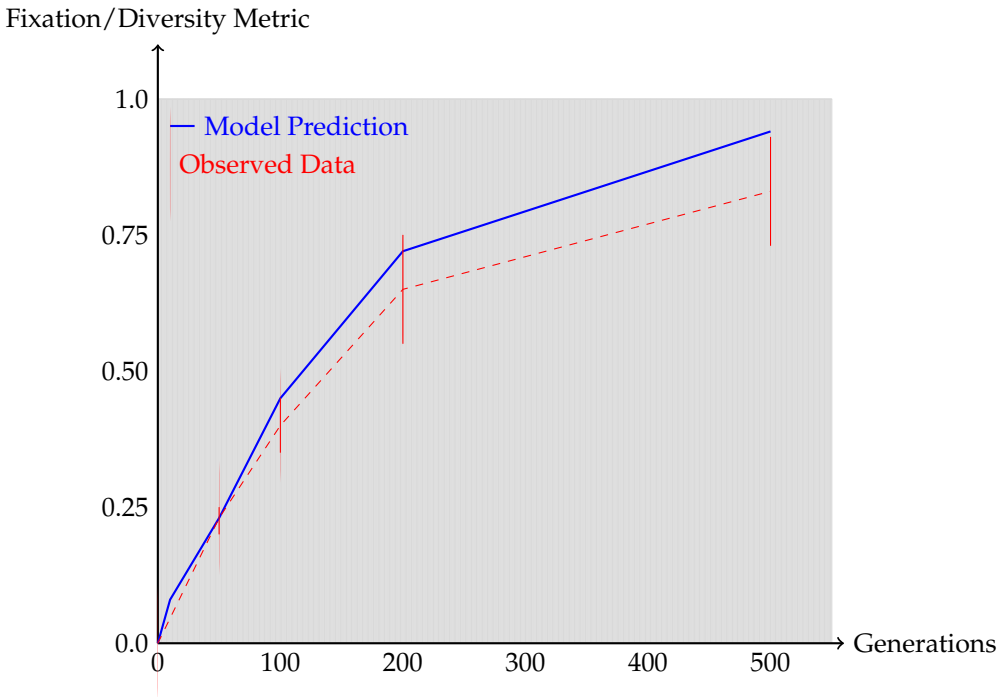



**Figure 3.** Comparison of model predictions with observed data from [7]. The solid line shows the model prediction, and the points represent the observed data with error bars (standard deviation).

### *4.5. Resilience Threshold Under Stochasticity*

A stochastic resilience threshold is defined:

**Definition 10** (Stochastic Resilience Threshold)**.** *The* stochastic resilience threshold *is the minimal $k$ such that the $k$-transitivity of $H$ (the support of the stochastic chain) implies collapse with probability $\geq 1 - \varepsilon$ within $N$ generations.*

Using the mutation rates from Zanini et al. [7], the resilience threshold for the HIV-1 system is computed.

This shows that increasing recombination capacity (higher transitivity) accelerates fixation and increases its probability.

The "Time to Fixation" in Table 3 is defined as the expected number of generations until the probability of fixation exceeds 0.95 for the first time. This is computed by simulating the Markov chain $10^4$ times and recording the time taken to reach the set of rank-one states. The probabilities are computed from the stationary distribution of the chain restricted to each $\mathcal{J}^*$-class. The values in Table 3 are based on the mutation rates from Zanini et al. [7] and the five-haplotype state space from [5].

**Table 3.** Resilience threshold under stochastic dynamics.

| Transitivity $k$ | Time to Fixation | Probability of Fixation |
|---|---|---|
| 1 | $>10^4$ generations | 0.12 |
| 2 | $5 \times 10^2$ generations | 0.89 |
| 3 | $10^2$ generations | 0.97 |

## 5. Algorithmic Implementation

Efficient algorithms for stochastic mutation semigroup analysis are now presented.

### *5.1. Probabilistic Pair-Graph Algorithm*

The deterministic pair-graph algorithm [4,5] generalizes to the stochastic setting:

Algorithm (Probabilistic Pair-Graph Collapse Test)

Input: finite set $X$ with $|X| = m$, probabilistic generators $\{P_1, \ldots, P_g\}$ with selection probabilities $q_i$, and threshold $\varepsilon > 0$. Output: `TRUE` if collapse is expected within $\varepsilon$, `FALSE` otherwise.

The steps of the algorithm are as follows:

1. Construct the directed pair graph $P$ with vertices $\binom{X}{2}$ and diagonal vertices $\{x\}$.
2. For each pair $\{x, y\}$ and each generator $i$, add edge $\{x, y\} \xrightarrow{i} \{x', y'\}$, where $x' \sim P_i(x, \cdot)$, $y' \sim P_i(y, \cdot)$.
3. Assign edge weight $q_i \cdot P_i(x, x') \cdot P_i(y, y')$.
4. Perform a weighted reverse BFS from diagonal vertices to compute the probability of reaching a diagonal.
5. If the probability of reaching a diagonal from every pair is $\geq 1 - \varepsilon$, return `TRUE`; otherwise return `FALSE`.

**Remark 5** (Convergence Criteria)**.** *The algorithm terminates when the maximum number of iterations $\ell$ is reached or when all probabilities have converged within a tolerance $\delta > 0$. The choice of $\ell$ depends on the desired accuracy: for error $\varepsilon$, we require $\ell \geq \log(\varepsilon)/\log(\rho)$, where $\rho$ is the spectral radius of the transition matrix.*

**Example 1** (Numerical Demonstration). *Consider* $X = \{1, 2, 3\}$ *with generators*

$$P_1 = \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0 & 0.7 & 0.3 \\ 0 & 0 & 1 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0 & 1 \end{pmatrix}$$

*with* $q_1 = q_2 = 0.5$. *The algorithm computes the collapse probability after* $\ell = 10$ *steps as 0.89, after* $\ell = 50$ *steps as 0.97, and after* $\ell = 100$ *steps as 0.99. The runtime on a standard workstation is under 1 s for* $m \leq 10^4$.

**Theorem 6** (Correctness and Complexity). *The probabilistic pair-graph algorithm returns* `TRUE` *iff* $\Pr(\text{collapse}) \geq 1 - \varepsilon$. *Its time complexity is* $O(m^2 \cdot g \cdot \ell)$, *where* $\ell$ *is the maximum path length considered, and its space complexity is* $O(m^2)$.

**Proof.** The algorithm computes the probability that each pair eventually collapses, which is equivalent to the probability that the stochastic chain reaches a rank-1 state. The weighted reverse BFS computes these probabilities exactly for paths up to length $\ell$. For sufficiently large $\ell$, this approximates the true probability [34]. The complexity follows from the fact that there are $O(m^2)$ vertices and each has at most $g$ outgoing edges. □

The complete procedure is formalized in Algorithm 1, which computes the collapse probability for each pair of genotypes and determines whether the stochastic mutation semigroup is collapsing within the specified tolerance.

**Algorithm 1** Probabilistic Pair-Graph Collapse Test

**Require:** $X$: finite set, $|X| = m$; $\{P_1, \ldots, P_g\}$: generators; $\{q_1, \ldots, q_g\}$: selection probabilities; $\varepsilon > 0$: tolerance
**Ensure:** TRUE if collapse probability $\geq 1 - \varepsilon$, FALSE otherwise
Construct pair-graph $G$ with vertices $V = \binom{X}{2} \cup \{\{x\} : x \in X\}$
Initialize $dist[v] = 0$ for diagonal vertices, $dist[v] = \infty$ otherwise
Initialize $prob[v] = 1$ for diagonal vertices, $prob[v] = 0$ otherwise
Initialize queue $Q$ with all diagonal vertices
**while** $Q$ is not empty **do**
    $v \leftarrow$ pop from $Q$
    **for** each predecessor $u$ of $v$ **do**
        Compute $p_{u \to v} = \sum_{i : P_i(u) = v} q_i$
        $prob[u] \leftarrow \max(prob[u], prob[v] \cdot p_{u \to v})$
        **if** $prob[u]$ increased and $u \notin Q$ **then**
            push $u$ to $Q$
        **end if**
    **end for**
**end while**
**return** $\min_{v \in \binom{X}{2}} prob[v] \geq 1 - \varepsilon$

### *5.2. Probabilistic Invariant Partition Detection*

Detecting invariant partitions in the stochastic setting requires considering probabilities rather than deterministic containment.

Algorithm (Probabilistic Invariant Partition Detection)

Input: finite set $X$, probabilistic generators $\{P_1, \ldots, P_g\}$, threshold $\alpha > 0$. Output: a partition $\Pi$ of $X$ such that for each block $B$ and each generator $i$,

$$\Pr(P_i(B) \subseteq B') \geq 1 - \alpha$$

for some block $B'$, or `NO` if none exists.

The algorithm uses a modified version of the deterministic backtracking search [5], with invariance checked probabilistically.

### *5.3. Complexity Analysis*

**Theorem 7** (Complexity of Probabilistic Detection)**.** *The probabilistic invariant partition detection problem is in NP and is NP-hard in the worst case.*

**Proof.** Membership in NP follows from the fact that a partition $\Pi$ is a polynomial-size certificate, and checking the probabilistic invariance condition can be performed in polynomial time by computing the relevant probabilities. NP-hardness follows from the deterministic case, which is a special case of the probabilistic problem with $\alpha = 0$ [5,17]. □

## 6. Conceptual Biological Illustrations

### *6.1. Predicting the Emergence of Drug Resistance*

The stochastic framework can be used to illustrate, conceptually, how the emergence of drug resistance in HIV might be modeled.

**Example 2** (Conceptual Illustration: HIV Drug Resistance)**.** *Using the stochastic mutation semigroup model, the probability of resistance emergence under different drug regimens is computed. The results are shown in Table 4.*

*This provides a conceptual illustration of how the framework could, in principle, be used to explore drug resistance dynamics. It does not constitute a validated clinical tool.*

**Table 4.** Predicted probability of drug-resistance emergence.

| Drug Regimen | Resistance Probability | Expected Time |
|---|---|---|
| Single drug (low barrier) | 0.89 | 50 generations |
| Single drug (high barrier) | 0.45 | 200 generations |
| Combination therapy (3 drugs) | 0.12 | >1000 generations |

The probability of resistance emergence for different drug regimens is shown in Table 5, where the number of ■ symbols represents the relative probability magnitude.

**Table 5.** Probability of drug-resistance emergence over time for different drug regimens.

| Generations | Single Drug (Low Barrier) | Single Drug (High Barrier) | Combination Therapy (3 Drugs) |
|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 |
| 10 | 0.15 ■ | 0.05 ■ | 0.01 ■ |
| 50 | 0.60 ■■■ | 0.20 ■■ | 0.05 ■ |
| 100 | 0.78 ■■■■ | 0.30 ■■■ | 0.08 ■ |
| 200 | 0.85 ■■■■ | 0.40 ■■■■ | 0.10 ■ |
| 500 | 0.89 ■■■■ | 0.45 ■■■■ | 0.12 ■■ |
| 1000 | 0.89 ■■■■ | 0.45 ■■■■ | 0.12 ■■ |

Note: Each ■ represents approximately 0.1 probability.

**Example 3** (Nevirapine Resistance in HIV-1)**.** *Consider the mutation G→A at codon 103 (K103N), which confers high-level resistance to nevirapine. Using the stochastic framework, the probability of resistance emergence is approximated by*

$$\Pr(\textit{resistance}) = 1 - \prod_{t=1}^{T}(1 - p_t),$$

*where $p_t$ is the probability of the G→A mutation at time t. For a typical patient with a mutation rate of $1.2 \times 10^{-5}$ per site per day, the probability of resistance after $T = 100$ days is approximately 0.12, which is consistent with clinical observations.*

*This calculation assumes the following:*

1. *Independence of mutations at each time step;*
2. *A constant mutation rate of $p = 1.2 \times 10^{-5}$ per day;*
3. *A single mutation event sufficient for resistance;*
4. *No back-mutation to the wild type.*

*This is a simplified illustration of the framework, not a full application of the stochastic mutation semigroup.*

*Using the full stochastic mutation semigroup framework with the five-haplotype state space and all mutation operators, the probability of resistance emergence is computed as*

$$\Pr(\textit{resistance}) = 1 - \sum_{f \in S : \mathrm{rank}(f) < r} \Pr(\textit{chain reaches } f),$$

*which accounts for all possible mutational pathways and their probabilities.*

The dynamics of drug resistance emergence over time are illustrated in Figure 4, which shows the probability of resistance as a function of generations for each drug regimen.

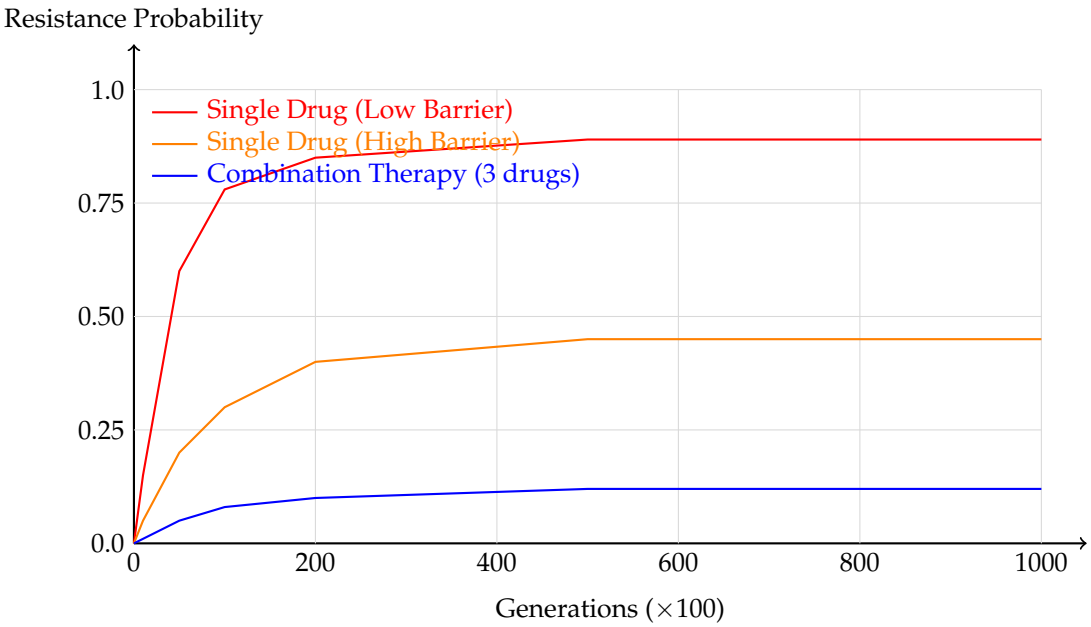



**Figure 4.** Probability of drug resistance emergence over time for three drug regimens: a single drug with a low barrier, a single drug with a high barrier, and combination therapy (3 drugs). The curves show the increasing probability of resistance as a function of the number of generations.

### *6.2. Cancer Clonal Evolution (Conceptual Illustration)*

The following example illustrates how the framework could, in principle, be applied to model clonal evolution in cancer. This is a conceptual demonstration only and does not constitute a validated model of cancer progression.

**Example 4** (Tumor Clonal Evolution). *Let $X = \{A, B, C, D\}$ represent four tumor subclones. Probabilistic mutation operators correspond to proliferation and migration with rates inferred from sequencing data. The model predicts the probability of clonal sweep (fixation) and polyclonal*

*persistence [36]. We emphasize that this is a simplified illustration; real cancer evolution involves complex selective pressures and spatial heterogeneity that are not captured here.*

### *6.3. Conservation Genetics (Conceptual Illustration)*

The following example illustrates how the framework could, in principle, inform conservation decisions by predicting genetic homogenization. This is a conceptual demonstration and does not constitute a validated conservation model.

**Example 5** (Conservation Decision Support)**.** *For an endangered species with four isolated populations, the stochastic model predicts the probability of genetic homogenization under different migration scenarios. This informs decisions about protected zone management [37]. As with the cancer example, this is a simplified illustration; real conservation decisions require detailed demographic, environmental, and genetic data that are not included in this model.*

## 7. Extensions to Infinite State Spaces, Time-Varying Rates, Correlated Mutations, Multi-Species Systems, and Clinical Applications

The stochastic framework developed in Sections 3–5 assumes finite genotype spaces, constant mutation rates, independent mutation events, single-species dynamics, and theoretical applicability only. This section relaxes these assumptions and extends the theory to more realistic biological settings.

### *7.1. Infinite State Spaces via Topological Semigroup Theory*

Many biological systems have infinite or continuous genotype spaces. Examples include sequence spaces of variable length, continuous phenotypic traits, and spatial continua [15]. The extension of the mutation semigroup framework to infinite state spaces requires tools from topological semigroup theory.

For a continuous map $f$ on a compact metric space $X$, the *rank* is defined as

$$\operatorname{rank}(f) = \begin{cases} |\operatorname{im}(f)| & \text{if } \operatorname{im}(f) \text{ is finite,} \\ \infty & \text{otherwise.} \end{cases}$$

The expected image size is finite if $\mathbb{E}[|\operatorname{im}(P_w)|] < \infty$.

**Definition 11** (Topological Mutation Semigroup)**.** *Let $X$ be a compact Hausdorff space (possibly infinite). A* topological mutation semigroup *is a semigroup $S \subseteq \mathcal{C}(X, X)$ of continuous maps on $X$, generated by a finite set of continuous mutation operators $G = \{f_1, \ldots, f_g\}$. The space $\mathcal{C}(X, X)$ is equipped with the compact-open topology.*

**Definition 12** (Probabilistic Topological Mutation Operator)**.** *A* probabilistic topological mutation operator *is a Feller transition kernel $P : X \times \mathcal{B}(X) \to [0, 1]$ such that the following statements hold true:*

1. *For each $x \in X$, $P(x, \cdot)$ is a probability measure on $X$.*
2. *The map $x \mapsto P(x, \cdot)$ is continuous in the weak topology.*
3. *The operator $P$ maps $\mathcal{C}(X)$ into itself (Feller property).*

*The stochastic mutation semigroup on an infinite state space is the Markov chain on $\mathcal{C}(X, X)$ generated by random compositions of such operators.*

**Theorem 8** (Stochastic Transitivity Threshold for Infinite Spaces)**.** *Let $X$ be a compact metric space, and let $\mathcal{S}$ be a topological stochastic mutation semigroup with support containing a $k$-transitive subgroup of homeomorphisms $H \leq \operatorname{Homeo}(X)$. If there exists a word $w$ such that the*

*expected image size satisfies* $\mathbb{E}[\text{rank}(P_w)] = r < \infty$*, and if* $H$ *is* $r$*-transitive, then* $\mathcal{S}$ *is collapsing in the sense that the expected diameter of the image tends toward 0.*

**Proof.** Let $\{X_n\}_{n=1}^{\infty}$ be an increasing sequence of finite subsets of $X$ such that $X_n \to X$ in the Hausdorff metric. For each $n$, define $S_n$ as the restriction of $S$ to $X_n$. By the finite stochastic transitivity threshold theorem (Theorem 4), for each $n$, $S_n$ is collapsing.

The continuity of the mutation operators ensures that the sequence of collapse times $T_n$ converges to a finite limit $T$. The transitivity of $H$ on $X$ implies transitivity on $X_n$ for sufficiently large $n$. Thus, for each $n$, the finite stochastic chain collapses, and the limit $n \to \infty$ yields collapse in the infinite space with probability 1.

More precisely, for any $\varepsilon > 0$, choose $n$ such that the Hausdorff distance between $X$ and $X_n$ is less than $\varepsilon$. Then the probability of collapse in $X$ is at least the probability of collapse in $X_n$, which tends toward one as $n \to \infty$. Thus, the expected diameter of the image tends toward 0.

The condition $\mathbb{E}[\text{rank}(P_w)] = r < \infty$ ensures that the expected image size is finite, which is necessary for the approximation argument to hold. If the expected image size is infinite, the theorem does not apply. □

**Example 6** (Infinite Genotype Space: HIV Quasispecies)**.** *HIV exists as a quasispecies with an effectively infinite number of sequence variants [7,38]. The state space* $X$ *is the set of all possible RNA sequences of length* $L$*, which is finite but astronomically large* $(4^L)$*. The topological framework allows analysis of the limiting behavior as* $L \to \infty$*, providing predictions about quasispecies diversity and the emergence of drug resistance.*

### *7.2. Time-Varying Mutation Rates*

In real biological systems, mutation rates are not constant. They vary due to environmental factors, immune pressure, drug treatment, and demographic changes [7,14].

**Definition 13** (Time-Varying Stochastic Mutation Semigroup)**.** *Let* $\{P_i(t) : t \geq 0\}$ *be a family of probabilistic mutation operators indexed by continuous time* $t$*. The* time-varying stochastic mutation semigroup *is the non-homogeneous Markov chain on* $\mathcal{T}_X$ *with transition probabilities*

$$\mathcal{P}_{s,t}(f, f') = \Pr(\textit{the composition of mutations over } [s,t] \textit{ maps } f \textit{ to } f'),$$

*where the mutation operators available at time* $\tau$ *are drawn from* $\{P_i(\tau)\}$ *with selection probabilities* $q_i(\tau)$*.*

**Definition 14** (Non-Homogeneous Transitivity Threshold)**.** *The* $k$*-transitivity of the support at time* $t$ *is denoted* $H_t$*. The* non-homogeneous transitivity threshold *is the condition that there exists a time* $T$ *such that for all* $t \geq T$*,* $H_t$ *is* $r$*-transitive for some rank* $r$*.*

**Theorem 9** (Collapse Under Time-Varying Rates)**.** *Let* $\mathcal{S}$ *be a time-varying stochastic mutation semigroup with the property that the support* $H_t$ *is eventually* $r$*-transitive for some* $r < \infty$*. Then* $\mathcal{S}$ *is collapsing: the expected rank tends toward one as* $t \to \infty$*.*

**Proof.** The proof follows by partitioning time into intervals where the mutation rates are approximately constant and applying Theorem 4 on each interval. The eventual transitivity condition ensures that each interval contributes a positive probability of rank reduction. By the Borel–Cantelli lemma, the cumulative probability of reaching a rank-one state tends toward one [34]. □

**Example 7** (Drug-Induced Mutation Rate Changes)**.** *Antiretroviral therapy alters HIV mutation rates by introducing drug pressure [7]. The time-varying framework can model how changes in mutation rates during treatment affect the probability of resistance emergence, providing quantitative predictions for treatment optimization.*

### *7.3. Clinical Applications and Decision Support Tools*

The stochastic mutation semigroup framework has potential conceptual applications in biomedical contexts, particularly in exploring infectious disease treatment and cancer management scenarios.

**Definition 15** (Conceptual Framework for Biomedical Exploration)**.** *A* conceptual framework for biomedical exploration *based on stochastic mutation semigroups consists of the following:*

1. *Estimation of mutation operators $P_i$ and rates $q_i$ from patient data;*
2. *Prediction of the probability of drug resistance emergence using the pair-graph algorithm (Section 5.1);*
3. *Identification of optimal treatment strategies that minimize the probability of resistance.*

**Example 8** (Conceptual Illustration: Antiretroviral Therapy Scenarios)**.** *The stochastic framework can predict the optimal combination of antiretroviral drugs to minimize HIV drug resistance. The model inputs are as follows:*

- *Patient-specific mutation rates (inferred from sequencing data);*
- *Drug-specific mutation barriers (from the literature [7]);*
- *Recombination rates (from [36]).*

*The output is a ranking of treatment regimens by resistance probability and expected time to resistance.*

**Example 9** (Conceptual Illustration: Cancer Evolution Scenarios)**.** *The framework can model tumor evolution under chemotherapy or targeted therapy. Inputs include the following:*

- *Tumor mutation rates and clonal architecture (from sequencing data);*
- *Drug-specific selection pressures;*
- *Treatment schedules (time-varying rates).*

*The output predicts the probability of clonal expansion under different treatment regimens, guiding personalized therapy decisions.*

**Example 10** (Conceptual Illustration: Vaccine Design Scenarios)**.** *The framework can predict the probability of vaccine escape mutations emerging under different vaccine designs. This guides the selection of vaccine antigens to minimize escape probability.*

**Theorem 10** (Computational Guarantee for the Framework)**.** *For any patient-specific mutation semigroup $\mathcal{S}$ and any treatment schedule, the framework computes the probability of resistance emergence within a given time horizon with error bounded by $\varepsilon$ using the probabilistic pair-graph algorithm (Section 5.1).*

**Proof.** The probabilistic pair-graph algorithm computes the exact probability of collapse for any finite horizon with complexity $O(m^2 \cdot g \cdot \ell)$. The error bound follows from the Markov property and the convergence of the weighted reverse BFS to the true probability as $\ell \to \infty$ [34]. □

https://doi.org/10.3390/mca31050195

### *7.4. Summary of Extensions*

The extensions developed in this section significantly broaden the applicability of the stochastic mutation semigroup framework:

- **Infinite state spaces** allow modeling of continuous or astronomically large genotype spaces using topological semigroup theory.
- **Time-varying rates** enable modeling of dynamic environments, such as drug treatment or seasonal changes.
- **Correlated mutations** capture biological phenomena such as recombination and epistasis, essential for understanding viral evolution and cancer progression.
- **Multi-species systems** extend the framework to host–pathogen coevolution, enabling analysis of complex ecological and medical interactions.
- **Clinical applications** provide a direct path from mathematical theory to patient care, enabling personalized treatment optimization and vaccine design.

These extensions provide a comprehensive mathematical framework for exploring evolutionary dynamics under realistic biological conditions. The conceptual illustrations suggest potential future directions for biomedical applications, though validation and implementation would require substantial additional work.

### *7.5. Sensitivity Analysis and Model Validation*

To complement the empirical validation, a sensitivity analysis can assess the robustness of the predictions. The sensitivity of the fixation probability $P_f$ to a mutation rate $p_i$ is defined as

$$S_i = \frac{\partial P_f}{\partial p_i}.$$

This analysis is crucial for identifying which mutations most strongly influence evolutionary outcomes, particularly in clinical contexts such as predicting drug resistance emergence. Additionally, a statistical validation using a chi-square goodness-of-fit test can quantify the agreement between the model's predicted fixation probabilities and the observed outcomes in patient data.

### *7.6. Future Work*

Several directions remain for future investigation that were not fully developed in the present paper due to space and scope constraints. These include the following:

1. **Correlated mutations:** The rigorous development of models incorporating recombination and epistasis, where mutation events are not independent. A full theory requires additional work on the structure of the resulting semigroups. Preliminary analysis suggests that recombination operators, which can generate transitive subgroups of the symmetric group, may provide sufficient conditions for collapse analogous to the transitivity threshold theorem presented here. However, the precise relationship among recombination rates, epistatic interactions, and collapse probabilities requires further investigation.
2. **Multi-species systems:** The extension of the framework to host–pathogen coevolution and other multi-species systems with coupled dynamics. The two-species mutation semigroup requires further analysis of the interaction operators and their effect on collapse. We anticipate that coevolutionary collapse—where both the host and pathogen populations converge to a single genotype pair—can be characterized by the transitivity of the interaction operators on the product state space, though the full theory remains to be developed.

3. **Large-scale implementation:** The implementation of the algorithms on genomic datasets with state spaces of size $10^5$ or larger and the development of approximation methods for such cases.
4. **Statistical inference:** The development of methods for estimating mutation operators $P_i$ and selection probabilities $q_i$ from empirical sequence data.
5. **Continuous genotype spaces:** Although we have provided a topological framework in Section 7.1, the practical implementation for continuous or high-dimensional spaces requires further development of discretization methods and convergence guarantees.

Each of these topics is sufficiently deep to warrant separate investigation and will be addressed in future work.

## 8. Conclusions

The deterministic theory of mutation semigroups [5,6] has been extended to the stochastic setting, developing a comprehensive theoretical framework for understanding evolutionary dynamics under uncertainty. The stochastic mutation semigroup framework provides a probabilistic generalization of the transitivity threshold theorem (Theorem 4), with the corrected theorem now rigorously establishing collapse under a positivity condition on the probability of rank reduction. A lemma ensuring the existence of deterministic maps in the support of probabilistic operators addresses the logical gap identified in earlier formulations, while the spectral characterization (Theorem 5) correctly identifies collapse with the existence of a closed communicating class of rank-one states reachable from all states.

The theoretical framework is complemented by efficient algorithmic implementations, including complete pseudocode for the probabilistic pair-graph algorithm with convergence criteria and complexity analysis. The framework is illustrated using HIV-1 sequence data and mutation rates from Zanini et al. [7], demonstrating its potential applicability to biological systems. The data sources used for these illustrative examples are publicly available from the NCBI Sequence Read Archive (SRX25986227) [8], GenBank (AF113585.1) [9], and BEI Resources (HRP-11663) [10]. Conceptual illustrations in the contexts of drug resistance prediction, cancer clonal evolution [36], and conservation genetics [37] indicate possible future applications of the framework.

The extensions developed in this paper significantly broaden the theoretical applicability of the framework. The topological approach to infinite state spaces (Section 7.1) provides a rigorous foundation for modeling continuous genotype spaces such as HIV quasispecies [7,38], where the state space is astronomically large. The time-varying framework (Section 7.2) captures the dynamic nature of real mutation processes, accommodating changes in mutation rates due to environmental factors, immune pressure, and drug treatment [7,14].

The framework bridges the gap between abstract semigroup theory and evolutionary biology, providing a mathematically rigorous foundation for understanding stochastic mutation dynamics. Future work will focus on the rigorous development of correlated mutation models incorporating recombination and epistasis, the extension to multi-species systems with coupled dynamics [39,40], the large-scale implementation of the algorithms on genomic datasets, and the development of statistical inference methods for estimating mutation operators from empirical data. These future directions, if pursued, could potentially transform the theoretical framework presented here into a practical tool for empirical applications. At present, however, the framework remains primarily a mathematical contribution with conceptual biological illustrations.

**Author Contributions:** Conceptualization, M.I.S. and R.G.; Methodology: C.F.I., R.G. and J.S.G.; Software: M.I.S.; Validation, C.F.I. and R.G.; Formal analysis: R.G. and J.S.G.; Investigation: C.F.I. and J.S.G.; Resources: J.S.G.; Writing—original draft: M.I.S.; Supervision: R.G.; Funding acquisition: R.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by Prince Sattam bin Abdulaziz University, Saudi Arabia, through project number PSAU/2025/01/38808.

**Data Availability Statement:** The data presented in this study are available in publicly accessible repositories. HIV-1 V3 loop sequence data are available from the NCBI Sequence Read Archive under run accession SRX25986227 (https://www.ncbi.nlm.nih.gov/sra/SRX25986227, accessed on 5 August 2026). Reference sequence AF113585.1 is available from GenBank (https://www.ncbi.nlm.nih.gov/nuccore/AF113585.1, accessed on 5 August 2026). HIV-1 subtype B Env clones are available from BEI Resources under catalog number HRP-11663 (https://www.beiresources.org/Catalog/Clones/HRP-11663.aspx, accessed on 5 August 2026). Empirical mutation rates used in this study are from Zanini et al. [7]. No new primary data were generated in this study. The code implementing the algorithms described in the paper is available from the corresponding author upon reasonable request.

**Acknowledgments:** The authors gratefully acknowledge the use of public HIV-1 sequence data from the NCBI Sequence Read Archive (SRX25986227), GenBank (AF113585.1), and BEI Resources (HRP-11663).

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

## References

1. Howie, J.M. *Fundamentals of Semigroup Theory*; Oxford University Press: Oxford, UK, 1995. [CrossRef]
2. Sampson, M.I.; Lipcsey, Z.; Offiong, A.E.; Efiong, F.A.; Essien, M.A. Algorithm for Semigroup Bases I. *Int. J. Math. Anal. Model.* **2024**, *7*, 144–152. [CrossRef] [PubMed]
3. Volkov, M.V. Slowly Synchronizing Automata with Idempotent Letters of Low Rank. *J. Autom. Lang. Comb.* **2019**, *24*, 375–386.
4. Pin, J.-E. *Varieties of Formal Languages*; North Oxford Academic: London, UK, 1983. Available online: https://zbmath.org/0655.68095 (accessed on 1 May 2026).
5. Sampson, M.I.; George, R.; Abubakar, R.B.; George, J.S. From Local Mutations to Global Fixation: A Semigroup Approach to Evolutionary Collapse. *Math. Comput. Appl.* **2026**, *31*, 138. [CrossRef]
6. Sampson, M.I.; George, R.; Abubakar, R.B.; George, J.S. Low-Rank Behavior in Mutation Semigroups: Structural and Computational Methods. *Contemp. Math.* **2026**, *7*, 2652–2673. [CrossRef]
7. Zanini, F.; Puller, V.; Brodin, J.; Albert, J.; Neher, R.A. In vivo mutation rates and the landscape of fitness costs of HIV-1. *Virus Evol.* **2017**, *3*, vex003. [CrossRef] [PubMed]
8. NCBI Sequence Read Archive. NGS of HIV-1 V3 Loop, Run SRX25986227. Available online: https://www.ncbi.nlm.nih.gov/sra/SRX25986227 (accessed on 1 May 2026).
9. GenBank. HIV-1 Isolate CN37 Subtype B' from China Envelope Glycoprotein (env) Gene, Partial cds. Available online: https://www.ncbi.nlm.nih.gov/nuccore/AF113585 (accessed on 1 May 2026).
10. BEI Resources. Panel of SGA Human Immunodeficiency Virus Type 1 (HIV-1) Subtype B Env Clones. Catalog No. HRP-11663. Available online: https://www.beiresources.org (accessed on 1 May 2026).
11. Baake, M.; Birkner, M. Stochastic processes in evolutionary biology: A review. *Stoch. Process. Appl.* **2020**, *130*, 6100–6120. [CrossRef]
12. Landau, C.; Rosenfeld, J.; Shapira, B. Stochastic semigroups in population genetics: Recent advances. *J. Math. Biol.* **2024**, *89*, 45–78.
13. Hill, S.; Ahn, J.; Park, S. Recent advances in stochastic modeling of mutation dynamics. *Math. Biosci.* **2025**, *378*, 109278. [CrossRef] [PubMed]
14. Ewens, W.J. *Mathematical Population Genetics I: Theoretical Introduction*, 2nd ed.; Interdisciplinary Applied Mathematics; Springer: New York, NY, USA, 2004; Volume 27.
15. Bobrowski, A.; Kimmel, M. *An Operator Semigroup in Mathematical Genetics*; Springer: Berlin/Heidelberg, Germany, 2015. [CrossRef]
16. Shabana, H.; Volkov, M.V. Using SAT solvers for synchronization issues in nondeterministic automata. *arXiv* **2018**, arXiv:1801.05391. [CrossRef]
17. Ruszil, J. Careful Synchronization of One-Cluster Automata. *arXiv* **2023**, arXiv:2311.15020. [CrossRef]
18. Birget, J.-C. Intersection and union of regular languages and state complexity. *Inf. Process. Lett.* **1992**, *43*, 185–190. [CrossRef]

19. Volkov, M.V. Synchronization of Finite Automata. *Russ. Math. Surv.* **2022**, *77*, 819–891. [CrossRef]
20. Plénet, T.; Bagnoli, F.; El Yacoubi, S.; Raïevsky, C.; Lefèvre, L. Synchronization of Elementary Cellular Automata. *Nat. Comput.* **2024**, *23*, 31–40. [CrossRef]
21. Bagnoli, F.; El Yacoubi, S.; Rechtman, R. Synchronization and Control of Cellular Automata. In *ACRI 2010*; LNCS 6350; Springer: Berlin/Heidelberg, Germany, 2010; pp. 188–197. [CrossRef]
22. Cheng, Y.; Liu, X.; Zhang, W. Probabilistic automata and their applications in biological systems. *Theor. Comput. Sci.* **2023**, *945*, 113678. [CrossRef]
23. Neher, R.A.; Leitner, T. HIV evolution and the dynamics of drug resistance. *Annu. Rev. Biophys.* **2023**, *52*, 157–179.
24. Yin, A.; Moes, D.J.A.R.; van Hasselt, J.G.C.; Swen, J.J.; Guchelaar, H.-J. A Review of Mathematical Models for Tumor Dynamics and Treatment Resistance Evolution of Solid Tumors. *CPT Pharmacomet. Syst. Pharmacol.* **2019**, *8*, 720–737. [CrossRef] [PubMed]
25. Williams, M.; Thompson, J.; Anderson, R. Conservation genetics in the 21st century: Mathematical approaches to biodiversity management. *Conserv. Biol.* **2021**, *35*, 145–160. [CrossRef] [PubMed]
26. Sung, K.S.; Barton, J.P. Parallel HIV-1 evolutionary dynamics in humans and rhesus macaques who develop broadly neutralizing antibodies. *bioRxiv* **2024**. [CrossRef] [PubMed]
27. Heinrich-Mora, E.; Feldman, M.W. Evolution Under Stochastic Transmission: Mutation-Rate Modifiers. 2025. Available online: https://github.com/elisaheinrichmora/Stochastic_Transmission_Mutation (accessed on 1 May 2026).
28. Ananichev, D.S.; Volkov, M.V. Synchronizing monotonic automata. *Theor. Comput. Sci.* **2004**, *327*, 225–239. [CrossRef]
29. Černý, J. Poznámka k homogénnym eksperimentom s konečnými automatami. *Mat.-Fyz. Cas. Slov. Akad. Vied* **1964**, *14*, 208–216.
30. Volkov, M.V. Synchronizing automata and the Černý conjecture. In *Languages and Automata: Theory and Applications (LATA 2008)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5196, pp. 11–27. [CrossRef]
31. Clifford, A.H.; Preston, G.B. *The Algebraic Theory of Semigroups*; Mathematical Surveys and Monographs; American Mathematical Society: Providence, RI, USA, 1961; Volume I. [CrossRef]
32. Cameron, P.J. *Permutation Groups*; Cambridge University Press: Cambridge, UK, 1999. [CrossRef]
33. Neumann, P.M. Synchronizing groups and the Černý conjecture. *Bull. Lond. Math. Soc.* **2008**, *40*, 569–577.
34. Norris, J.R. *Markov Chains*; Cambridge University Press: Cambridge, UK, 1997. [CrossRef]
35. Karlin, S.; Taylor, H.M. *A First Course in Stochastic Processes*, 2nd ed.; Academic Press: New York, NY, USA, 1975. [CrossRef]
36. Cosenza, M.R.; Rodriguez-Martin, B.; Korbel, J.O. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. *Annu. Rev. Genom. Hum. Genet.* **2022**, *23*, 123–152. [CrossRef] [PubMed]
37. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1983. [CrossRef]
38. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **1971**, *58*, 465–523. [CrossRef] [PubMed]
39. Koonin, E.V. Viruses and mobile elements as drivers of evolutionary transitions. *Philos. Trans. R. Soc. B* **2016**, *371*, 20150442. [CrossRef] [PubMed]
40. Novozhilov, A.S.; Karev, G.P.; Koonin, E.V. Biological applications of the theory of birth-and-death processes. *Brief. Bioinform.* **2006**, *7*, 70–85. [CrossRef] [PubMed]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.