# A hybrid stochastic-deterministic approach to explore multiple infection and evolution in HIV

**Jesse Kreger[a], Natalia L. Komarova[a] and Dominik Wodarz[b,a]**

a) Department of Mathematics, University of California Irvine, Irvine CA 92697

b) Department of Population Health and Disease Prevention Program in Public Health
Susan and Henry Samueli College of Health Sciences, University of California, Irvine CA 92697

**Abstract**

Multiple infection (when a single cell can become super-infected with multiple copies of virus) can effect evolutionary processes in HIV infection, such as the generation and spread of different mutations. Synaptic transmission (where multiple copies of virus can be transferred during a single cell-to-cell interaction) leads to an increase in infection multiplicity. Here, we analyze the effect of multiple infection on the evolution of an HIV infection using a hybrid stochastic-deterministic algorithm. This algorithm allows us to stochastically simulate a large number of cells and large number of cell subpopulations in a computationally efficient way. Specifically, we classify each cell subpopulation as "large" or "small" based on some size threshold (for which we provide an analytical lower bound), and then simulate the small populations stochastically, while simultaneously using the deterministic equations for the large populations. We first demonstrate to what extent deterministic models are incomplete in predicting mutant dynamics, and then use the hybrid method to study aspects of mutant evolution. Focusing on the mutant populations at the time of peak infection, we find that overall, multiple infection promotes the existence of mutant strains of virus, both in terms of faster generation and overall number. It also increases the variance in mutant numbers and results in a larger probability to observe rare strains, such as triple mutants. In the context of multiple infection, synaptic transmission is superior at generating mutants, as it increases the number of transcription events that occur. This effect is observed for neutral and advantageous mutants, even in the presence of interference. For disadvantageous mutants, purely synaptic or purely free virus transmission pathways result in an increased number of mutants, but in the presence of complementation, purely synaptic transmission maximizes mutant production and spread.

## 1 Introduction

Human Immunodeficiency Virus-1 (HIV-1) evolution is considered to be a driving factor of infection progression, which creates great potential for the virus to escape from immune system responses, form drug resistance, and overcome vaccine candidates [9, 16, 22]. HIV-1 has a relatively high mutation rate ($3 \times 10^{-5}$ per base pair) [21], which along with a quick turnover of the virus population [32, 10, 26], is thought to contribute to virus evolution *in vivo*.

As mutation is an important component of HIV-1 infection and evolution, mathematical models are a useful tool in order to study how virus mutation and evolution occurs *in vivo* [23, 34, 27, 7, 1], especially where experimental tracking and manipulation of mutant virus strains can be difficult. Mutant strains are key for the ability of an infection to escape from immune responses as well as to the evolution of drug resistance during anti-viral therapy, as viral variants can escape from drug effects as well as immune cell clones.

Since mutation is rare and random, the dynamics of infection contain important stochastic effects. However, since it is often not computationally practical to stochastically simulate infections at realistic population sizes, many mathematical models of virus infection turn to ordinary differential equation (ODEs) models. These models describe the overall effects of how populations interact

with each other, and thus can only describe the "average" generation and subsequent growth of mutant strains of the virus.

Multiple infection has been documented in HIV-1 infected patients [14], and multiple infection has been shown to be promoted by direct cell-to-cell transmission through virological synapses [19] [11, 3]. Multiple infection occurs when a single cell is infected with multiple (possibly genetically distinct) strains of virus. When a cell is infected with multiple copies of virus, complex social interactions can take place, such as viral complementation, viral interference, recombination, and others [20, 6, 30, 8, 2]. Synaptic transmission occurs when copies are transferred through virological synapses between cells. This typically results in multiple copies of virus being tranferred during one infection event. This has a large effect on the multiple infection of cells, and additionally is likely to result in the direct cotransmission of different virus strains from one cell to another [5]. This is in contrast to free virus transmission, which generally results in the infection of cells with only one copy of virus.

In this paper, in order to study the stochastic effects of multiple infection and synaptic transmission, we implement a hybrid stochastic-deterministic algorithm, which was introduced in [29]. This algorithm was developed for population dynamic and evolutionary biology systems, and allows us to stochastically simulate the trajectories of many mutant strains of the virus. We use an ODE mathematical model in order to include an arbitrary number of mutant strains, and then describe the hybrid algorithm used to simulate these systems, which is both computationally reasonable and includes important stochastic effects of mutant populations at low numbers. This hybrid algorithm is then used to analyze the effect of multiple infection, by modulating the maximum infection multiplicity from a minimum of one (single infection) to higher integers (multiple infection). We also study the effect of synaptic transmission on the existence and spread of different mutant populations throughout infection, by varying the percent of different virus transmission strategies (free virus and cell-to-cell). We also use the hybrid algorithm to compare differences between the completely deterministic mathematical model and the stochastic simulations of it, to demonstrate the impact of stochastic effects on basic viral infection *in vivo*.

## 2 Mathematical model description

In this section we derive a mathematical model in order to simulate infection with a large number of mutant strains.

### 2.1 Single virus strain model

We begin with a model for HIV-1 infection *in vivo*, which includes both free virus transmission and synaptic cell-to-cell transmission [17]. We assume that cells are sufficiently well-mixed, such that an ordinary differential equation (ODE) model is enough to capture the relevant dynamics. To include the possibility of multiple infection, we let $x_i(t)$ represent the number of cells infected with $i$ copies of the virus at time $t$, where $i$ takes on whole numbers between 0 (uninfected) and $N$ (maximum infection multiplicity). Descriptions of the model parameters can be found in Table 1.

With only one viral strain, the model is

$$\begin{aligned}
\dot{x}_0 &= \lambda - \beta Z x_0 - \gamma Z x_0 - d x_0, && (1)\\
\dot{x}_i &= \beta Z(x_{i-1} - x_i) + \gamma Z(-x_i) - a x_i, \text{ if } 0 < i < S && (2)\\
\dot{x}_i &= \beta Z(x_{i-1} - x_i) + \gamma Z(x_{i-S} - x_i) - a x_i, \text{ if } S \le i \le N - S && (3)\\
\dot{x}_i &= \beta Z(x_{i-1} - x_i) + \gamma Z(x_{i-S}) - a x_i, \text{ if } N - S < i < N && (4)\\
\dot{x}_N &= \beta Z(x_{N-1}) + \gamma Z(x_{N-S}) - a x_N. && (5)
\end{aligned}$$

We assume that $N$ is large enough to not result in a large amount of cells near the end of the cascade. In the above equations, cell free virus transmission happens at rate $\beta$, and synaptic transmission at rate $\gamma$. Mathematically the two processes are similar, except the number of viruses passed during an infection event is one for free virus transmission and $S$ for synaptic transmission. Therefore, the total number of infected cells is independent of transmission mode when the total rate of transmission, $\beta + \gamma = c$, is fixed. However, varying rates of free virus and synaptic transmission such that $\beta + \gamma = c$ will have a large effect on the multiplicity of infection (average number of virus copies in infected cells) and the relative fraction of different strains in infected cells.

Let us define the number of infected cells

$$Z(t) = \sum_{i=1}^{N} x_i(t), \tag{6}$$

which is the sum of all of the infected subpopulations. This model, and all subsequent models, have a virus free steady state and an infection steady state. The virus free steady state is

$$\begin{aligned}
x_0 &= \frac{\lambda}{d}, && (7)\\
Z &= 0. && (8)
\end{aligned}$$

The infection steady state is

$$\begin{aligned}
x_0 &= \frac{a}{\beta + \gamma}, && (9)\\
Z &= \frac{\lambda}{a} - \frac{d}{\beta + \gamma}. && (10)
\end{aligned}$$

The stability of these steady states depends on the basic reproduction ratio, $R_0 = \frac{\lambda(\beta+\gamma)}{ad}$. If $R_0 < 1$ the virus free steady state is stable and if $R_0 > 1$ then the infection steady state is stable. This is detailed and further explored in [17].

## 2.2 Including mutant strains

Mutation is an important component of HIV-1 infection. Therefore, we generalize the preceding ODE model to include an arbitray number of mutant strains. We assume that forward and back point mutations can occur independently at $k$ locations, all with rate $\mu$. This creates $2^k$ viral strains. In the case $k = 3$, we have the wild type (abc), three strains with a single point mutation (Abc, aBc, and abC), three strains with a double point mutation (ABc, AbB, and aBC), and a triple mutant strain (ABC). In the general case of $k$ mutations, we denote the wild type (abc...) as strain $\mathcal{S}_1$ and the $k$-mutant (ABC...) as the $\mathcal{S}_{2^k}$ strain. Single mutant through $k-1$ mutant strains are denoted by some $\mathcal{S}_i$ for $i = 2, 3, \ldots, 2^k - 1$. Similarly as in previous models, upon each infection event, mutation can happen at each location. Therefore, any strain can mutate into any other strain, and the probability of such an event is based on the number of point mutations in

| Notation | Description | Units (if applicable) |
|---|---|---|
| $\lambda$ | production rate of uninfected cells | days$^{-1}$ |
| $\beta$ | rate of free virus transmission | days$^{-1}$ |
| $\gamma$ | rate of synaptic cell-to-cell transmission | days$^{-1}$ |
| $d$ | death rate of uninfected cells | days$^{-1}$ |
| $a$ | death rate of infected cells | days$^{-1}$ |
| $x_0(t)$ | number of uninfected cells at time $t$ | NA |
| $x_i(t)$ | number of cells infected with $i$ copies of the virus at time $t$ | NA |
| $Z(t)$ | sum of all infected populations at time $t$, $Z(t) = \sum_{i=1}^{N} x_i(t)$ | NA |
| $Z_i(t)$ | sum of fraction of subpopulations infected with $i^{\text{th}}$ strain | N/A |
| $N$ | maximum infection multiplicity | N/A |
| $S$ | number of viruses transferred per synapse | N/A |
| $k$ | number of locations in which a point mutation can occur | N/A |
| $\mu$ | mutation rate | N/A |
| $h$ | ODE step size | N/A |
| $\mathcal{M}$ | hybrid algorithm size threshold | N/A |
| $F_i$ | fitness of the $i^{\text{th}}$ strain | N/A |

Table 1: Description of model parameters and units (if applicable).

which the two strains differ. The likelihood a strain mutates at $i \in \{0, 1, \ldots, k\}$ specified locations is $\mu^i(1-\mu)^{k-i}$. Define $D(\mathcal{S}_i, \mathcal{S}_j) \in \{0, 1, \ldots, k\}$ as the number of point mutations where strain $\mathcal{S}_i$ and $\mathcal{S}_j$ differ. The probability that strain $\mathcal{S}_i$ mutates into strain $\mathcal{S}_j$ (or vice versa) is then $\mu_{D(\mathcal{S}_i,\mathcal{S}_j)} := \mu^{D(\mathcal{S}_i,\mathcal{S}_j)}(1-\mu)^{k-D(\mathcal{S}_i,\mathcal{S}_j)}$.

Let $x_{\mathbf{i}} = x_{s_1,s_2,\ldots,s_{2^k}}(t)$ be the number of cells that are infected with the respective number of viruses at time $t$. Let $Z_1(t),\ Z_2(t),\ \ldots,\ Z_{2^k}(t)$ be the sum of the fraction of all subpopulations infected with the relevant strain at time $t$, that is

$$Z_i(t) \quad = \quad \sum_{0<|\mathbf{i}|\leq N} \frac{s_i}{|\mathbf{i}|} x_{\mathbf{i}}(t) \text{ for } i = 1, 2, \ldots 2^k, \tag{11}$$

where we define $|\mathbf{i}| = s_1 + s_2 + \ldots + s_{2^k}$. To include mutation, when calculating each $Z_i$ we include the relevant mutation factors $\mu_{D(\mathcal{S}_i,\mathcal{S}_j)}$, which describe the likelihood of each strain mutating into each other. Examples of this are included in the Supplementary Information Section 1.1 (single mutants) and Section 1.2 (double mutants). Again we let $Z(t)$ be the sum of all infected subpopulations. We have that

$$Z(t) = \sum_{0<|\mathbf{i}|\leq N} x_{\mathbf{i}}(t) = Z_1(t) + Z_2(t) + \ldots + Z_{2^k}(t). \tag{12}$$

We have that $S \geq 1$ is the constant number of viral copies transferred per synaptic infection event. Therefore, for each synaptic infection event, we must choose $S$ viral copies from the infected cell to pass on to the target cell. Assume that the infecting cell is infected with $s_1, s_2, \ldots, s_{2^k}$ copies of the respective number of viruses, and that we transmit $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{2^k}$ copies of the respective strains. Synaptic transmission can result in infection by any combination of these strains, as long as $\hat{s}_1 + \hat{s}_2 + \ldots + \hat{s}_{2^k} = S$ is satisfied. The likelihood for each of these synaptic infection events is based on the probability to pick each strain for infection ($Z_i$ for $i = 1, 2, \ldots 2^k$). For each transmission event, we multiply the probability to transmit the respective number of each strain $Z_1^{\hat{s}_1} Z_2^{\hat{s}_2} \ldots Z_{2^k}^{\hat{s}_{2^k}}$ by the number of ways to transmit that number of each strain, which is the multinomial coefficient $\binom{S}{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{2^k}} = \frac{S!}{\hat{s}_1! \hat{s}_2! \ldots \hat{s}_{2^k}!}$. Therefore, for $\hat{s}_1 + \hat{s}_2 + \ldots + \hat{s}_{2^k} = S$, we define the synaptic infection

intensity as

$$G_{\hat{s}_1,\hat{s}_2,\ldots,\hat{s}_{2^k}} = \frac{S!}{\hat{s}_1!\hat{s}_2!\ldots\hat{s}_{2^k}!}\sum_{0<|\mathbf{i}|\leq N} Z_1^{\hat{s}_1}Z_2^{\hat{s}_2}\ldots Z_{2^k}^{\hat{s}_{2^k}}x_{i,j,k,l}. \tag{13}$$

The ODEs are then

$$\begin{aligned}
\dot{x}_{0,0,\ldots,0} &= \lambda - \beta Z x_{0,0,\ldots,0} - \gamma Z x_{0,0,\ldots,0} - d x_{0,0,\ldots,0}, && (14)\\
\dot{x}_{\mathbf{i}} &= \beta\Big[\sum_{i=1}^{2^k}\Big(Z_i x_{\ldots,s_{i-1},\ldots}\Big) - Z x_{\mathbf{i}}\Big] \\
&+ \gamma\Big[\sum_{\hat{s}_1+\hat{s}_2+\ldots+\hat{s}_{2^k}=S}\Big(x_{s_1-\hat{s}_1,s_2-\hat{s}_2,\ldots,s_{2^k}-\hat{s}_{2^k}}G_{\hat{s}_1,\hat{s}_2,\ldots,\hat{s}_{2^k}}\Big) - Z x_{\mathbf{i}}\Big] - a x_{\mathbf{i}}, \text{ for } |\mathbf{i}|>0 && (15)
\end{aligned}$$

where any term with a negative index is 0 and cells can be infected with a maximum of $N$ total copies of virus.

### 2.3 Including advantageous/disadvantageous mutants

There are many ways to model fitness differences between distinct viral strains. If all the strains are neutral, and a coinfected cell is chosen for an infection event, we assume that the probability to transmit a given virus strain is proportional to the fraction of the strain among all viruses in the cell. If different viral strains have different fitness, the difference is implemented during the infection event, which can correspond to different rates of reverse transcription. That is, an infecting strain is again chosen based on its fraction in the cell. A disadvantageous (advantageous) mutant would then have a lower (higher) rate of transmission upon infection. In this way, a strain's fitness is independent of whether or not it is contained within a coinfected cell. This method of modeling fitness is one choice among many, others are explored further in [17].

This can be implemented in a straightforward way, by modifying the previous models. Here, the mean fractions of each population are modified by a fitness parameter $F_i \in [0,\infty)$ for $i = 1, 2, \ldots, 2^k$. In the non-neutral case ($F_i \not\equiv 1$) we have that for instance Equation (11) becomes

$$Z_i(t) = F_i \sum_{0<|\mathbf{i}|\leq N} \frac{s_i}{|\mathbf{i}|}x_{\mathbf{i}}(t) \text{ for } i = 1, 2, \ldots 2^k. \tag{16}$$

Therefore, higher values of $F_i$ are correlated with higher fitness for the $i^{\text{th}}$ strain. Also note that when non-neutral fitness landscapes are considered, $Z(t)$ becomes the sum of all infected subpopulations weighted by their respective fitness and is no longer equal to just the sum of all infected subpopulations.

### 2.4 Model parameters

Based on experimental evidence, we estimate the kinetic parameters in the model. Recall from Section 2.1 that in the neutral case the basic reproduction ratio $R_0$ for these systems is $R_0 = \frac{\lambda(\beta+\gamma)}{ad}$, the infection steady state for infected cells is $\frac{\lambda}{a} - \frac{d}{(\beta+\gamma)}$, and the virus free steady state for uninfected cells is $\frac{\lambda}{d}$.

We have an infection steady state value of $3.1 \times 10^7$ infected cells [4]. Assuming that the death rate of infected cells is approximately 2.2 days [26], we assume that $a = 0.45$ days$^{-1}$. Further, we

set $R_0 = 8$ [28, 24]. Finally, we introduce parameter $\nu$ such that we have a total body population of $10^\nu$ lymphocytes at virus free steady state. Then we have the system of equations

$$3.1 \times 10^7 = \frac{\lambda}{a} - \frac{d}{(\beta+\gamma)}, \tag{17}$$

$$8 = \frac{\lambda(\beta+\gamma)}{ad}, \tag{18}$$

$$10^\nu = \frac{\lambda}{d}. \tag{19}$$

These equations uniquely imply that $\lambda \approx 1.59 \times 10^7$ days$^{-1}$, $d = \lambda \times 10^{-\nu}$ days$^{-1}$, and $\beta + \gamma = 3.6 \times 10^{-\nu}$ days$^{-1}$. We assume $S = 3$ and to get an average multiplicity of infection of around 3-4 [15, 12, 13, 31], we vary the rate of free virus transmission ($\beta$) and cell-to-cell transmission ($\gamma$). Widely varying estimates for average infection multiplicities have been published, and there is some uncertainty about that. Further, we choose $\nu = 9$, and therefore have $d \approx 0.0159$ days$^{-1}$. For initial conditions, we assume uninfected steady state and a single infected cell (infected with the wild type) [25], that is $x_{0,0,\ldots,0} = \frac{\lambda}{d}$ and $x_{1,0,\ldots,0} = 1$.

At these parameters, triple and further mutant strains are negligible at peak infection, that is on average less than half a cell is infected with these strains when the number of infected cells is maximized (around seven days into infection). After peak infection, each strain will tend toward selection-mutation balance. These patterns can be seen by running the model with triple mutation first to peak infection, and then many days past peak infection.

# 3 Hybrid algorithm

In this section we introduce the hybrid algorithm, which allows for stochastic simulations at large population sizes with many cell subpopulations.

## 3.1 Theoretical basis

Since our system includes mutations, which are rare, the dynamics will contain important stochastic effects. Because of the large population sizes and numerous cell subtypes, it is not computationally practical to simulate the entire system with a fully stochastic method such as the Gillespie algorithm. We therefore implement a hybrid stochastic-deterministic algorithm, introduced in [29]. This algorithm was developed specifically to describe biological population dynamics in evolutionary systems, and allows to simulate the trajectories of both large and small populations simultaneously.

The hybrid algorithm is based on the idea that if a cell population is sufficiently large, an ODE representation can provide a good approximation of most stochastic trajectories of the population. We can write the ODEs systems as a single vector equation $d\mathbf{V}/dt = \mathbf{F}(\mathbf{V})$, where $\mathbf{V}$ is a vector that contains all the cell subpopulations. Let $\mathcal{M} > 0$ be a given size threshold. We classify the cell population $x_{\mathbf{i}}$ as small at time $t$ if $x_{\mathbf{i}}(t) < \mathcal{M}$, or large otherwise. We simulate the small populations stochastically using the Gillespie algorithm and use the ODEs for the large populations. Further details of the hybrid method are given in the Supplementary Information Section 2.1 and 2.2.

## 3.2 Implementation

The size threshold $\mathcal{M}$ is a very important parameter in the hybrid algorithm. If $\mathcal{M} = 0$, then at each time point each non-zero population is classified as large and the hybrid algorithm is identical to the deterministic solution of the ODEs. If $\mathcal{M}$ is very large, that is larger than all populations

at all times, then the hybrid algorithm is the same as the completely stochastic Gillespie simulation of the model and can be extremely computationally inefficient. For intermediate $\mathcal{M} > 0$, the hybrid algorithm is computationally efficient and the averages over many hybrid simulations go from approximating the deterministic predictions to converging to the stochastic averages as $\mathcal{M}$ increases. Therefore, in order to efficiently approximate the completely stochastic implementation of the model, we need to choose a small intermediate $\mathcal{M}$ such that the results are close to the fully stochastic implementation.

We can achieve this by comparing the hybrid averages over many simulations to completely stochastic averages over many simulations for simplified models, such as assuming a constant large number of uninfected cells or using similar parameters that result in smaller and more computationally manageable population sizes. For these models, completely stochastic simulations can be carried out and allow us to determine what size threshold $\mathcal{M}$ is reasonable for the related models. Specifically, since the averages over many hybrid simulations start from the deterministic prediction ($\mathcal{M} = 0$) and converge to the completely stochastic average, similarly to [29] we i) set some difference threshold $\varepsilon > 0$, ii) test multiple size thresholds $\mathcal{M}$, and iii) choose the smallest $\mathcal{M}$ such that the hybrid average is within $\varepsilon$ of the completely stochastic average.

Table 2 contains approximate average computer simulation run times for the completely deterministic ODE system, the hybrid method, and the completely stochastic Gillespie algorithm. Each system is run for the single mutation, double mutation, and triple mutation models. All simulations include only free virus transmission with multiple infection ($N = 3$), represent established infections only (we ignore stochastic simulations in which the infection dies out), and are stopped once the infected cell population reaches $10^8$ cells. The times for the ODE and hybrid simulations also depend on the ODE deterministic step size, $h$ (here $h = 10^{-5}$). In general, the number of strains per model is $2^k$ and the number of equations (subpopulations) per model is $\binom{N+2^k}{2^k}$.

Because the parameters chosen for the simulations in Table 2 correspond to $R_0 = 8$, relatively small size threshold $\mathcal{M}$ gives a good approximation of the fully stochastic simulations. For a fixed infected cell population size, simulations with lower $R_0$ result in longer run times because it takes longer for the number of infected cells to reach the threshold. Larger maximum infection multiplicity $N$ also results in longer runs times as there are more subpopulations and equations.

| Model | single mutation<br>$k = 1$<br>2 strains, 10 equations | double mutation<br>$k = 2$<br>4 strains, 35 equations | triple mutation<br>$k = 3$<br>8 strains, 165 equations |
|---|---|---|---|
| Full ODEs | $< 1$ second | 4 seconds | 12 minutes |
| Hybrid, $\mathcal{M} = 10$ | $< 1$ second | 4 seconds | 13 minutes |
| Hybrid, $\mathcal{M} = 10^3$ | $< 1$ second | 4 seconds | 13 minutes |
| Hybrid, $\mathcal{M} = 10^5$ | $< 1$ second | 6 seconds | 15 minutes |
| Hybrid, $\mathcal{M} = 10^7$ | 1 minute | 7 minutes | 30 hours |
| Full Gillespie | 12 minutes | 100 minutes | 1 week |

Table 2: Approximate average run times for a single simulation for the completely deterministic ODE system, the hybrid method with different threshold values ($\mathcal{M}$), and the completely stochastic Gillespie algorithm (rows) Each system is run for the single mutation, double mutation, and triple mutation models (columns). In each system we assume all strains are neutral ($F_i = 1$ for all $i$). The other parameters are $N = 3,\ h = 10^{-5},\ \mu = 3 \times 10^{-5},\ \lambda = 1.59 \times 10^7,\ \beta = 3.60 \times 10^{-9},\ \gamma = 0,\ d = 0.016$, and $a = 0.45$.

### 3.3 Choosing a size threshold $\mathcal{M}$

In the case of virus dynamics, we have developed an analytical method for finding a lower bound on size threshold $\mathcal{M}$, which is based on the notion of $R_0$. This method does not depend on the number of mutations, infection multiplicity, fitness landscape, etc, and provides an analytic lower bound on $\mathcal{M}$. Depending on the details of the model, it is possible that a higher $\mathcal{M}$ is needed to get accurate descriptions of mutant dynamics. In general, we can always confirm that a chosen $\mathcal{M}$ is large enough using the $\varepsilon$ test described in the preceding section and in [29].

The basic reproduction ratio, $R_0$, is the average number of newly infected cells per single infected cell at the beginning of the infection. Therefore, infections with larger $R_0$ will lead to quicker and more successful growth of the overall virus population. For $R_0 > 1$, the dynamics of the infected subpopulations at the beginning of the infection are driven by two effects: i) largely deterministic exponential growth, and ii) stochastic fluctuation. Stochastic fluctuation is most important at initial infection, when the number of infected cells is low. As the number of infected cells grows, the dynamics transition to more deterministic exponential growth [25]. Furthermore, as $R_0$ increases, the amount of fluctuating in the infected cell subpopulations decreases and deterministic growth dominates, which accounts for decreased stochasticity in the system and better prediction from the ODEs. Therefore, as $R_0$ increases, lower hybrid size threshold $\mathcal{M}$ is needed in order to approximate the fully stochastic dynamics.

In the deterministic system, if $R_0 > 1$ then the infection will never go extinct. This is not the case in the stochastic setting, where even if $R_0$ is quite large, a single infected cell can die out before successfully infecting other cells. The rate at which infections stochastically go extinct, $\frac{1}{R_0}$, can be approximated from the deterministic ODEs (see Supplementary Information Section 2.3 for details).

The rate of stochastic extinction of the hybrid method should match the rate of stochastic extinction of the simulations in the fully stochastic case. We define a successful infection as an infection that reaches at least $K > 0$ cells. For $R_0 > 1$, the hybrid size threshold $\mathcal{M}$ and size of successful infection $K$ play similar roles in hybrid simulations. This is because once an infected cell subpopulation reaches $\mathcal{M}$ cells, the ODEs will take over and it will grow deterministically (i.e. the infection will never die out). This provides a nice lower bound for size threshold $\mathcal{M}$ for the hybrid method. Let $\delta > 0$ be some small difference threshold. We define the lower bound size threshold, $\hat{\mathcal{M}}$, as

$$\hat{\mathcal{M}} = \lceil \ln\left(1 + \frac{R_0 - 1}{\delta R_0}\right) / \ln R_0 \rceil, \tag{20}$$

where $\lceil . \rceil$ denotes the ceiling function. Details of this calculation are described in Supplementary Information Section 2.3.

Since increasing $R_0$ monotonically leads to a higher rate of successful infections, we see that increasing $R_0$ leads to decreasing $\hat{\mathcal{M}}$. This makes sense because as $R_0$ increases, the amount of fluctuating in the infected cell population decreases, resulting in better prediction from the deterministic ODEs. Therefore, for higher $R_0$, lower $\mathcal{M}$ is needed in the hybrid method in order to reasonably approximate the fully stochastic trajectories.

While this analytic method only provides a lower bound for the size threshold parameter for the hybrid method to approximate the completely stochastic dynamics, it also transforms the more abstract choice of a size threshold $\mathcal{M}$ into the more concrete choice of a difference threshold $\delta$ which explicitly measures the difference between the rate of stochastic extinction and the potentially much lower rate of extinction for hybrid simulations. This is especially important for studying statistics of the stochastic runs (such as the variation in the numbers of mutants).

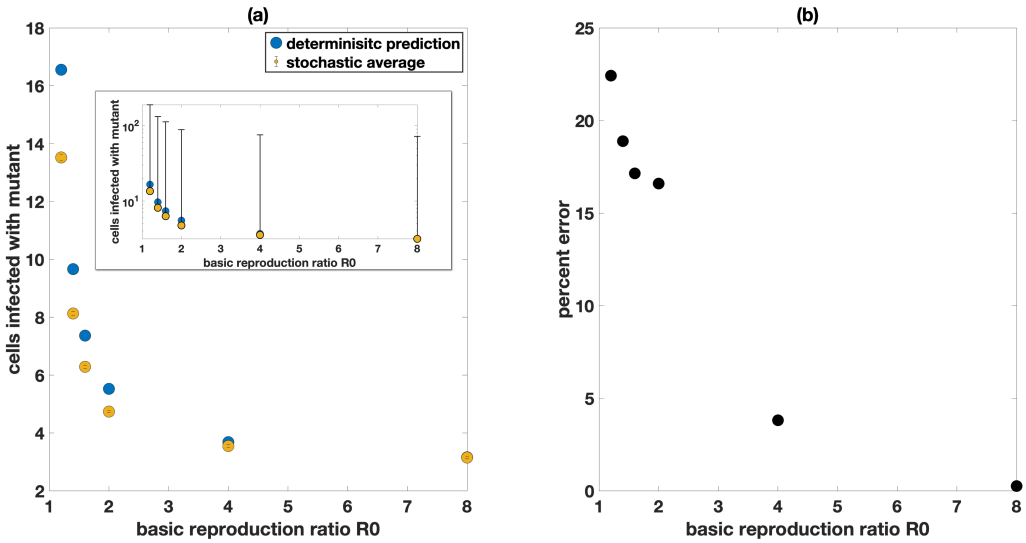


Figure 1: Comparison of the basic reproduction ratio $R_0$ and the accuracy of the deterministic prediction of the number of cells infected with the mutant virus at an infected cell population size of $10^4$ in the single mutation model. Simulations in which infections are not established are discarded when calculating the averages. The deterministic predictions are in blue and the stochastic hybrid simulations with $\mathcal{M} = 10^4$ (infected populations always treated completely stochastically) are in yellow. Each yellow stochastic dot represents the average taken over at least 2 million simulations. (a) Deterministic prediction versus stochastic averages with standard error bars (too small to be seen). The insert represents the deterministic prediction versus stochastic averages with standard deviation bars. (b) Percent error between the deterministic predictions and stochastic averages shown in (a). We see that increasing $R_0$ generally leads to a better prediction, both in terms of the difference between the deterministic prediction and the stochastic average and also in the percent difference between the two. We also see that the variance of the stochastic simulations decreases with increasing $R_0$. We have $R_0 = \frac{\lambda(\beta+\gamma)}{ad}$, and the parameters are $N = 3$, $\mu = 3 \times 10^{-5}$, $\lambda = 1.59 \times 10^7$, $\beta = 3.60 \times 10^{-9}$, $\gamma = 0$, and $d = 0.016$. The infected cell death rate $a$ is adjusted to achieve the required $R_0$. All strains are neutral.

# 4 Comparison of the ODE and hybrid models in the context of infection dynamics

## 4.1 The expected number of neutral mutants

We start by predicting the number of neutral mutant viruses. Since infections are always established in the deterministic setting for $R_0 > 1$, the ODEs do a better job of predicting stochastic averages over many runs if we restrict to only simulations in which a successful infection is established. If we calculate the averages of mutants over all stochastic runs, then the average for uninfected (infected) cells will be much larger (smaller) than the deterministic prediction. This is because when a stochastic extinction event occurs, the number of uninfected cells returns to the large $\frac{\lambda}{d}$ steady state and the number of infected cells is 0.

Even if we only take into account the stochastic runs, in which infections are successfully established and maintained for the relevant timespan, it is possible for the deterministic predictions and averages over many stochastic simulations to disagree significantly. As described in Section 3.3, for $R_0$ closer to 1, there is a higher degree of stochastic fluctuation, which is difficult for the ODE model to capture and predict. Accordingly, the variance of the stochastic simulations also

decreases with increasing $R_0$.

While HIV-1 infection is characterized by a relatively high value of $R_0 \approx 8$, bacterial and other infections can have values of $R_0$ very close to 1. For these systems, it is even more important to consider how stochasticity affects the system dynamics. Figure 1 compares the deterministic prediction versus the stochastic average for the number of cells infected with a mutant strain when the infected cell population has reached $10^4$ cells. The deterministic predictions are in blue and the stochastic averages are in yellow. We see that increasing $R_0$ leads to a better prediction, both in terms of the difference between the deterministic prediction and the stochastic average and also in the percent difference between the two. We also see that the variance of the stochastic simulations decreases with increasing $R_0$. For larger infected cell populations closer to peak infection, the absolute difference between the ODE predictions and stochastic averages are much larger, but the percent error is similar. For instance an infected cell population of $10^7$ gives similar results, with a relative error at $R_0 = 1.1$ of about 11%. Past peak infection, the wild type and mutant population will fluctuate around steady state values, which the ODEs correctly predict (and which can be determined analytically).

We now turn to studying the dynamics of single mutants that are advantageous or disadvantageous. These mutant strains are characterized by larger or smaller fitness values compared to the wild type viruses.

### 4.2 The expected number of advantageous/disadvantageous mutants

As in the neutral case, the ODEs do not give accurate predictions for the mean numbers of mutants, especially if $R_0$ is close to 1. Figure 2(a) compares the deterministic prediction versus the stochastic average (similarly to Figure 1) for the number of cells infected with an advantageous mutant strain when the infected cell population has reached $10^4$ cells. Here, the fitness of the wild type is 1 and the fitness of the advantageous mutants is 1.1. Figure 2(b) compares the deterministic prediction versus the stochastic average for the number of cells infected with a disadvantageous mutant strain, where the fitness of the wild type is 1 and the the fitness of the disadvantageous mutants is 0.9.

In both cases (advantageous or disadvantageous) we again see that increasing $R_0$ leads to a better prediction, both in terms of the difference between the deterministic prediction and the stochastic average and also in the percent difference between the two. We also see that the variance and percent error between the ODEs and the stochastic simulations decreases with increasing $R_0$. For advantageous mutants, the percent error for $R_0 = 1.2$ is 1116% and for $R_0 = 8$ is 51%. For disadvantageous mutants, the percent error for $R_0 = 1.2$ is 18% and for $R_0 = 8$ is 7%.

In the advantageous case, we see that as in the neutral case, the deterministic prediction is an overestimate of the number of cells infected with the mutant. This is because low numbers of cells infected with the mutant can experience stochastic fluctuation before taking off and spreading rapidly. However, for the disadvantageous mutant we see that the deterministic prediction is an underestimate of the number of cells infected with the disadvantageous mutant. This is because the disadvantageous mutant can achieve fluctuating spikes of larger numbers before being out-competed by the wild type.

### 4.3 Probability distributions of mutant numbers

The deterministic setting tells us nothing about the distribution of mutant numbers. Many biological contexts, such as the evolutionary fate of drug resistant or immune escape mutants, require

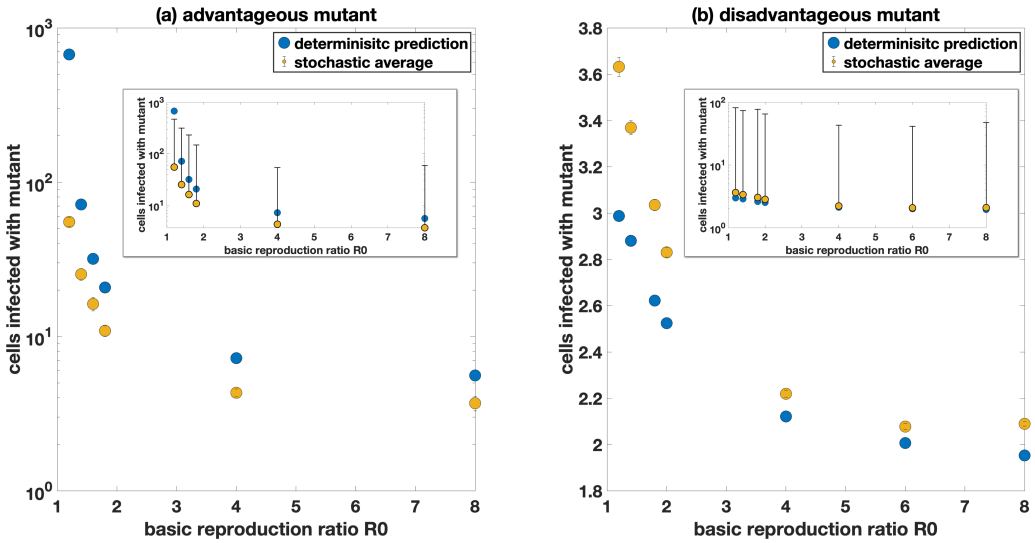


Figure 2: Comparison of deterministic and stochastic results for non-neutral mutant strains. (a) An advantageous mutant strain ($F_{\text{mutant}} = 1.1$). Each yellow stochastic dot represents the average taken over at least 11 thousand simulations. (b) A disadvantageous mutant strain ($F_{\text{mutant}} = 0.9$.). Each yellow stochastic dot represents the average taken over at least 3.5 million simulations. All other parameters are the same as in Figure 1.

information not only on the mean, but on the whole distribution. The hybrid method is especially useful here, because it allows for efficient computations at large population sizes. For relatively large $R_0$ values and large numbers of uninfected cells, the hybrid method produces very accurate results even for relatively small threshold values $\mathcal{M}$ (and thus requires little computational time). This is because with a large $R_0$ and large initial number of uninfected cells, the mutant dynamics depend largely on when the mutant is generated and grows to exceed a small threshold. Figure 3 shows an example of the distribution of the number of cells infected with the single and double mutant strain of the virus when the infected cell population reaches $10^6$ cells for hybrid simulations with size threshold $\mathcal{M} = 50$ and $\mathcal{M} = 500$. We find that for both single and double mutants, the probability distributions with $\mathcal{M} = 50$ and $\mathcal{M} = 500$ are statistically the same, and thus that $\mathcal{M} = 50$ is large enough to capture the stochastic dynamics.

Furthermore, the hybrid method also provides a good approximation of the stochastic distribution and variance for smaller $R_0 > 1$. Here again, the mutant dynamics depend largely on when the mutant is generated and grows to exceed a threshold. However, since there is more stochasticity, larger size threshold sizes $\mathcal{M}$ are needed. Figure 4 shows the distribution of the number of cells infected with the single mutant strain of the virus when the infected cell population reaches $10^6$ cells for hybrid simulations with varying size thresholds. Here we see larger size thresholds $\mathcal{M}$ are needed for a small basic reproduction ratio $R_0$. In summary, Figure 3 shows that for $R_0 = 8$ (less stochasticity) $\mathcal{M} = 50$ is large enough, but Figure 4 shows that for $R_0 = 1.5$ (more stocahsticity) $\mathcal{M} = 50$ is not large enough to capture the stocahstic dynamics, and larger size thresholds are needed.

### 4.4 The timing of mutant emergence

An important benchmark in the evolution of HIV infection is the emergence of mutant strains of the virus. Using the continuous deterministic ODE system, the best we can do to predict when

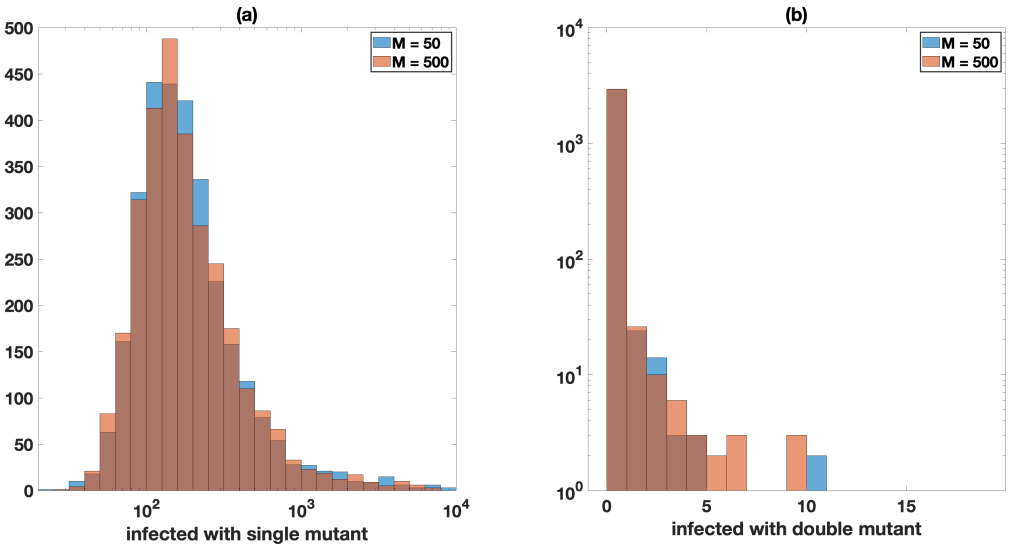


Figure 3: Distribution of the number of cells infected with the single and double mutant strain of the virus (double mutant model) when the infected cell population reaches $10^6$ cells for hybrid simulations with size threshold $\mathcal{M} = 50$ and $\mathcal{M} = 500$. Simulations in which infections are not established are discarded when calculating the averages. For each size threshold 3 thousand simulations were performed. (a) Number of cells infected with one of the single mutant strains. (b) Number of cells infected with the double mutant strain. The parameters are as in Figure 1 with $R_0 = 8$.For both single and double mutants, the probability distributions of the cell numbers infected with mutant viruses are statistically the same for $\mathcal{M} = 50$ and $\mathcal{M} = 500$ ($p > 0.1$ by Kolmogorov-Smirnov test).

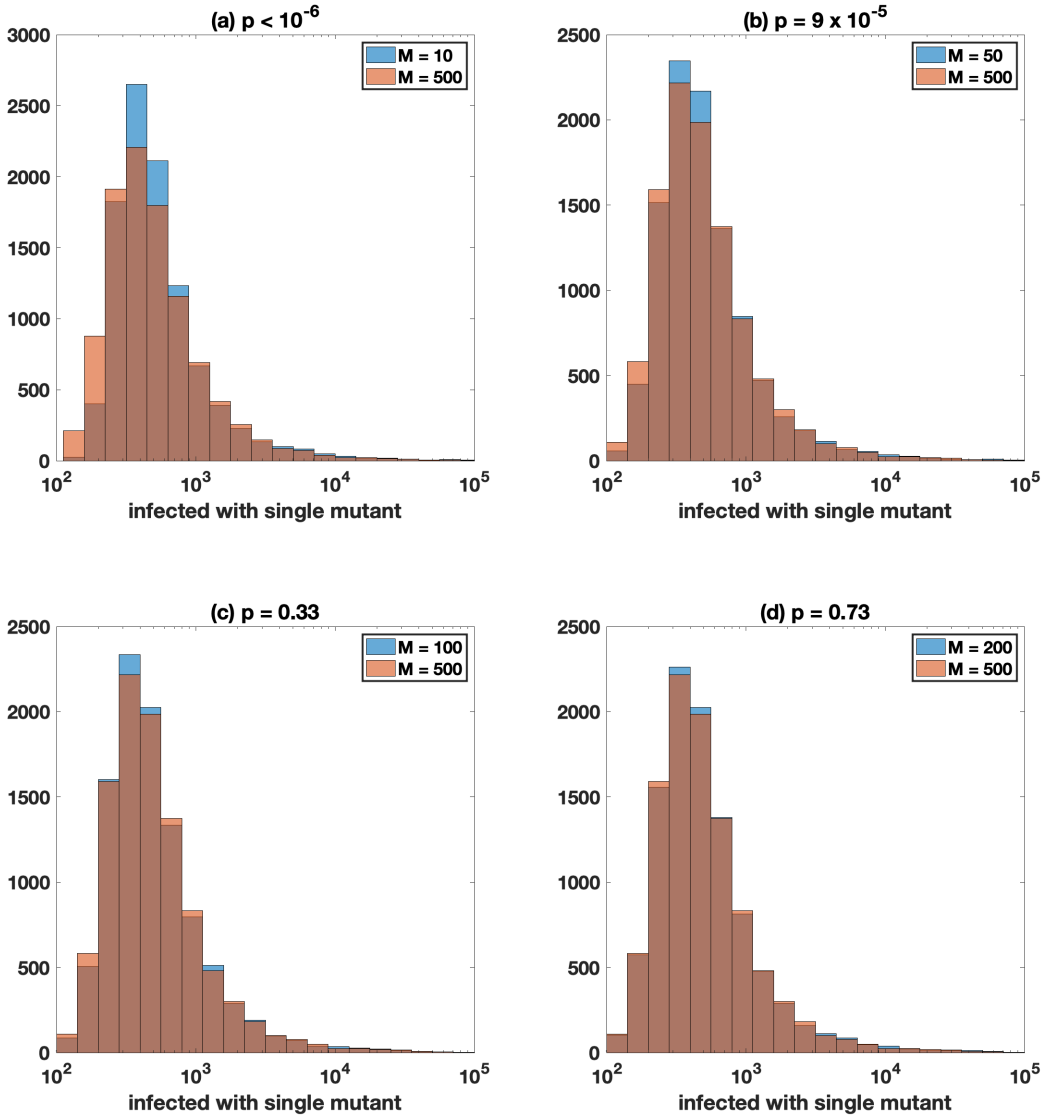


Figure 4: Distribution of the number of cells infected with the single mutant strain of the virus (single mutant model) when the infected cell population reaches $10^6$ cells for hybrid simulations. Simulations in which infections are not established are discarded when calculating the averages. Distributions for each size threshold represent 10 thousand simulations. (a) Size threshold $\mathcal{M} = 10$ and $\mathcal{M} = 500$. (b) Size threshold $\mathcal{M} = 50$ and $\mathcal{M} = 500$. (c) Size threshold $\mathcal{M} = 100$ and $\mathcal{M} = 500$. (d) Size threshold $\mathcal{M} = 200$ and $\mathcal{M} = 500$. Here $N = 1$ and the other parameters are as in Figure 1 with $R_0 = 1.5$. The $p$-value from Kolmogorov-Smirnov test is shown for each comparison.

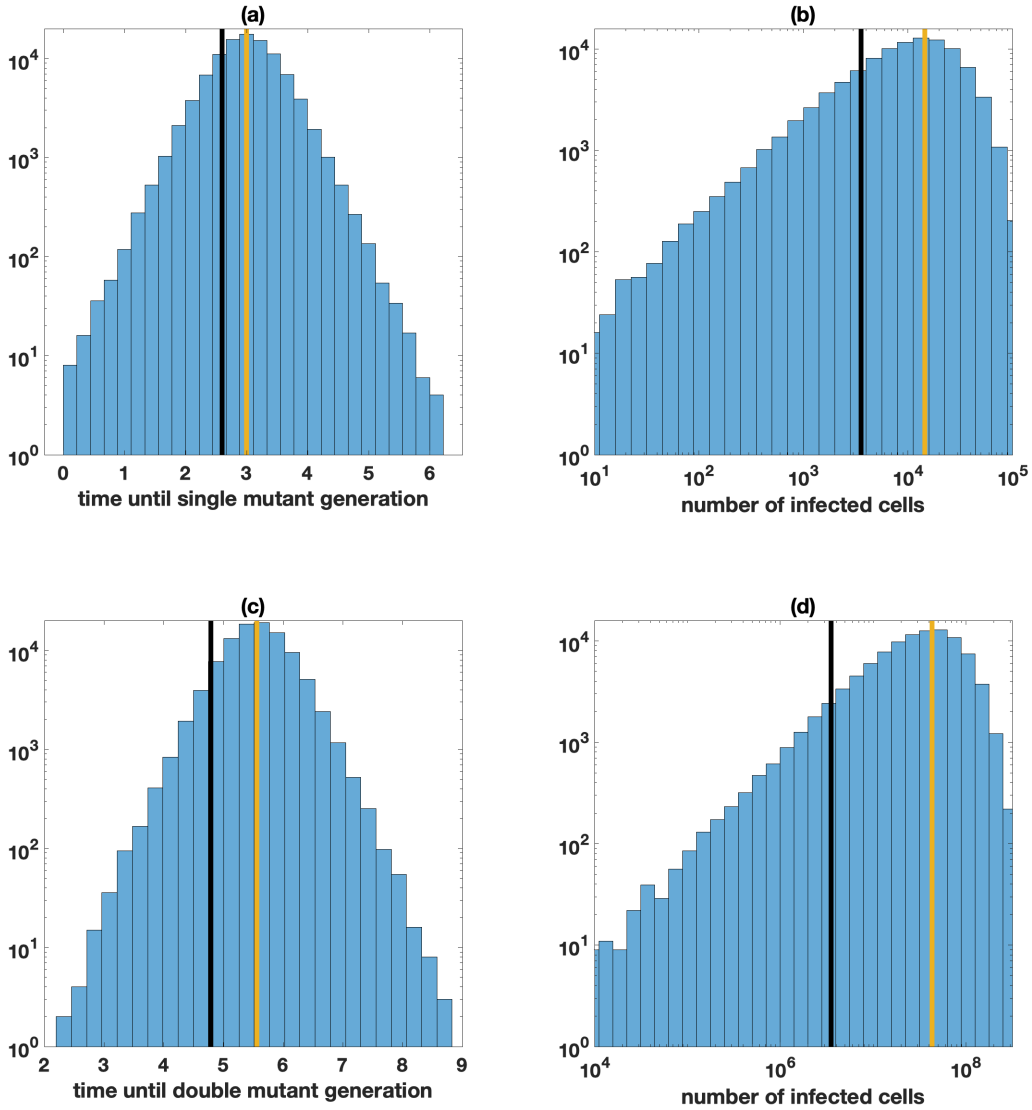


Figure 5: Distribution of generation time of first single/double mutant virus in the double mutation model. Simulations in which infections are not established are discarded when calculating the averages. The deterministic prediction is denoted with the black vertical line and the hybrid average is denoted with the yellow vertical line. Distributions represent 100 thousand hybrid simulations with size threshold $\mathcal{M} = 50$. (a) Time until either single mutant generation. The deterministic prediction that the first single mutant virus (of both strains) will be generated is around 2.6 days, whereas in the stochastic case it is around 3 days. (b) Number of infected cells at first single mutant generation. The deterministic prediction is that the number of infected cells is around $3.5 \times 10^3$, whereas in the stochastic case it is around $1.4 \times 10^4$. (c) Time until double mutant generation. The deterministic prediction that the first double mutant virus will be generated is around 4.8 days, whereas in the stochastic case it is around 5.6 days. (d) Number of infected cells at first double mutant generation. The deterministic prediction is that the number of infected cells is around $3.5 \times 10^6$, whereas in the stochastic case it is around $4.3 \times 10^7$. The parameters are as in Figure 1 with $R_0 = 8$.

the first single/double mutant virus will emerge is by finding where the number of single/double mutant viruses first reaches 1. However, with the hybrid model, which is stochastic and discrete for small populations, we can get a full distribution of when we expect the first single/double mutant virus to appear, both in terms of time and infected cell population size, which could be important for treatment. An example of such a distribution can be seen in Figure 5. Here, the deterministic model underestimates the mean time of single mutant generation by 13%, and underestimates the mean number of infected cells at first single mutant generation by 75%. Similarly, the deterministic model underestimates the mean time of double mutant generation by 14%, and underestimates the mean number of infected cells at first double mutant generation by 94%. In general, the ODE error will become larger as more mutations are introduced, because of the increasing importance of stochastic generation for more mutational steps.

# 5 Impact of multiple infection and synaptic transmission on mutant evolution

Mutant evolution is impacted by multiple infection [33]. Here, we use the hybrid algorithm to simulate stochastic infections, to determine to what extent multiple infection makes a difference under different evolutionary scenarios. To this end, we compare simulations where only single infection is taken into account (maximum infection multiplicity $N = 1$) and simulations containing multiple infection (maximum infection multiplicity $N > 1$). Multiple infection, and coinfection (when a cell is infected with distinct virus strains), is especially important because interactions between distinct mutant strains can occur within a cell (this includes the phenomenon of viral recombination, which goes beyond the scope of this paper [18]). In the context of multiple infection, synaptic transmission (where multiple copies of virus are transferred per transmission event) is also a factor. As stated before, the spread of mutant virus depends on two factors i) stochastic generation of the mutant strain, and ii) the subsequent spread of the mutant strain once it has been generated (which is more stochastic for small population sizes, and more deterministic for larger populations sizes).

## 5.1 Neutral mutants

We first study the dynamics of neutral mutants, and then consider more complex fitness landscapes.

### 5.1.1 Effect of multiple infection in the absence of synaptic transmission

First, we assume that only free virus transmission takes place. In this model, for $R_0 = 8$, most cells do not become infected with more than 11 copies of virus by the time of peak infection (about $6.2 \times 10^8$ infected cells). Therefore, for the multiple infection case we set maximum infection multiplicity $N = 11$. Figure 6 shows the distributions of cells infected with single and double mutants for simulations with single and multiple infection, close to peak infection. We record the number of mutants at $6 \times 10^8$ infected cells, as it is close to the peak and most stochastic simulations will reach this threshold. Here, we see that when multiple infection is allowed ($N = 11$), there is a larger number of both single and double mutant viruses at peak infection. The number of single mutants is higher because of two factors:

(i) When multiple infection is included, there are more target cells the single mutant strains can infect. When limited to single infection, the wild type initially spreads, limiting the number of cells in which the mutants, which are generated later, can infect (as only uninfected cells can be infected).

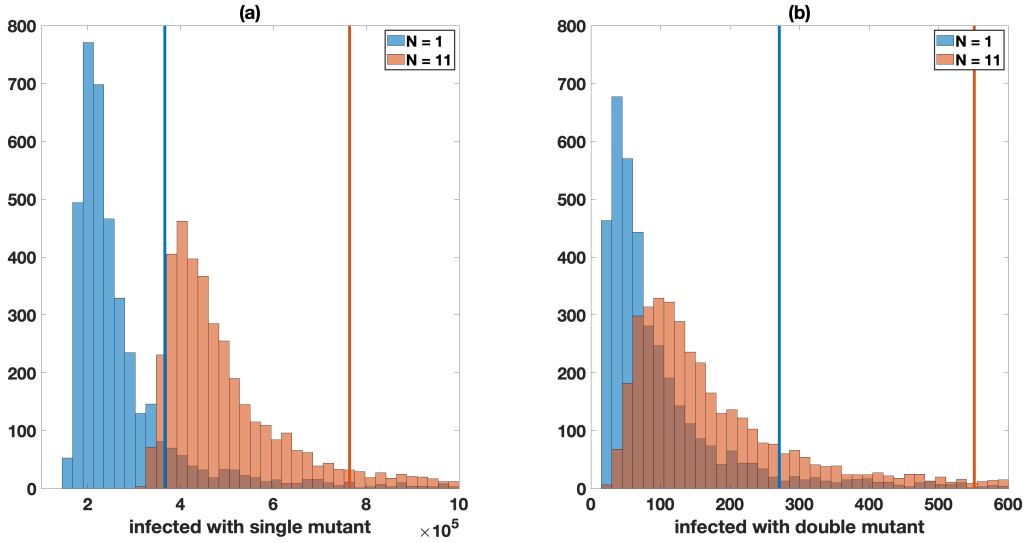


Figure 6: Neutral mutant evolution in the absence of synaptic transmission, comparing simulations with single infection only ($N = 1$, blue) and in the presence of multiple infection ($N = 11$, red). The mean values are shown by vertical lines. (a) Number of cells infected with one of the single mutant strains. The average for single infection is approximately $3.7 \times 10^5$ and for multiple infection is approximately $7.6 \times 10^5$. The T test between the two distributions gives a $p$-value less than $10^{-6}$. (b) Number of cells infected with the double mutant strain. The average for single infection is approximately 271 and for multiple infection is approximately 551. The T test between the log of the two distributions gives a $p$-value less than $10^{-6}$. Distributions represent 4 thousand hybrid simulations with size threshold $\mathcal{M} = 50$. Simulations in which infections are not established (or in the rare case a simulation does not reach the infected size threshold) are discarded. Simulations are stopped when the infected cell population is close to peak infection ($6 \times 10^8$ cells). The other parameters are as in Figure 1 with $R_0 = 8$.

(ii) Multiple infection results in faster generation of the mutant strains (relative to the total number of infected cells), because of the increased number of successful infection events in the presence of multiple infection.

These effects of multiple infection contribute to increased mutant virus spread during early infection. In particular, near the peak infection, we observe that in the presence of multiple infection, the number of cells infected with a mutant virus is about 2.4 higher than that under single infection only; the standard deviation is also larger by a factor of about 4.2, figure 6(a). Both of these factors also apply to the double mutant strain, which also benefits from larger numbers of single mutants at peak infection, as higher numbers of single mutants will lead to higher numbers of double mutants through mutation. Under multiple infection, the mean number of cells infected with double-hit mutants is abut 2.0 times higher than that under single infection only, and the standard deviation is about 3.2 times higher, figure 6(b).

Finally, we tested the prevalence of triple mutants, by comparing simulations with $N = 1$ and $N = 3$ (not shown in the figure). While with or without multiple infection, we expect a significant number of single and double mutants at peak infection, the number of triple mutants is very low (less than 0.2 on average). The probability to have at least one cell infected by a triple-mutant is 3.3% under multiple infection, which is about 1.7 times higher than that under single infection (1.9%) (this result is significant with $p = 2.5 \times 10^{-3}$ by the z-test, with $2.7 \times 10^4$ runs under single infection and $1.9 \times 10^3$ runs under multiple infection).

Multiple infection can also have complex effects that can hinder mutant virus spread. In the absence of mutation, if for instance an infecting cell is singly (or only) infected with the mutant, then it is guaranteed to pass on the mutant strain. However, if an infecting cells is infected with a single copy of the mutant but many copies of the wild type, then the mutant strain is unlikely to spread. In the case of only free virus transmission, this is not much of a factor as the multiplicity of infection at peak infection is relatively small (1 when $N = 1$ and about than 1.9 when $N = 11$).

#### 5.1.2 Effect of multiple infection in the presence of synaptic transmission

Next, we study the role of synaptic transmission in the evolutionary dynamics of infection. Again, we compare simulations with single infection only ($N = 1$), and simulations that allow multiple infection, where we now include synaptic cell-to-cell transmission.

During cell-to-cell transmission, we assume that $S \geq 1$ viruses are transferred to the target cell per successful infection event. We allow simultaneous transfer of all possible non-negative integer combinations of the different strains such that the total number of virus copies transferred is $S$. During an infection event, upon selection of each virus, the viral copy has a probability to mutate. This greatly increases the ability of synaptic transmission to generate mutant strains of virus. To see this, consider infection from a cell that is infected only by the wild type. During a free virus transmission event, there is only one chance for mutation into a mutant strain. However, during a synaptic transmission event, there are $S$ chances for mutation into a mutant strain.

Including synaptic transmission results in increased multiplicity of infection. Therefore, we need to adjust the maximum infection multiplicity used in the model, to avoid spurious accumulation of cell numbers at the end of the infection cascade. For our model, for $R_0 = 8$, $S = 3$, and 100% synaptic transmission, it is rare for a cell to become infected with more than 25 copies of virus by the time of peak infection. Therefore, we set $N = 26$ for all simulations that include synaptic transmission.

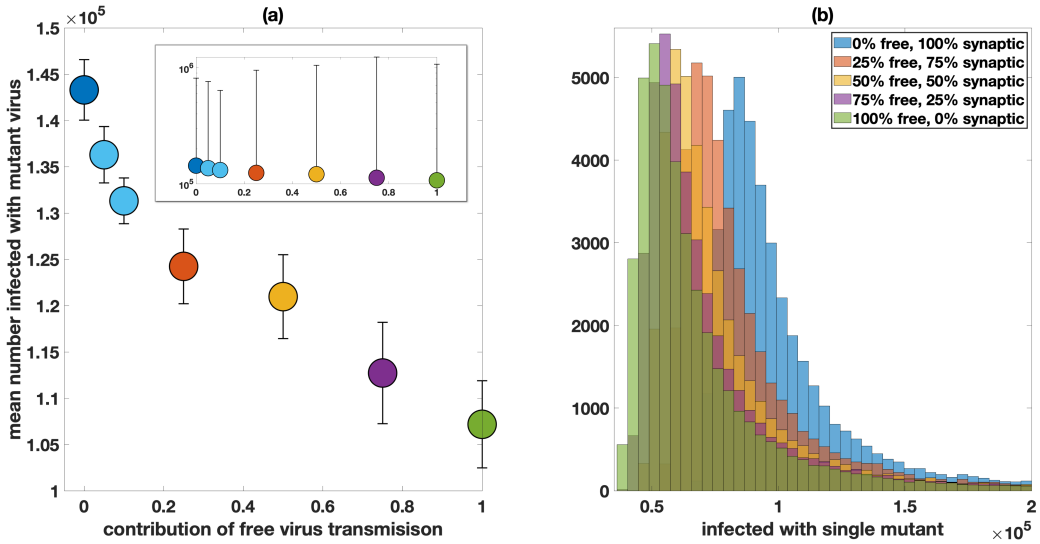


Figure 7: Neutral mutant evolution in the presence of synaptic transmission, comparing simulations with different combinations of free virus and synaptic transmission (single mutations only). (a) Number of cells infected with the mutant strain under different transmission strategies. The horizontal axis is the percent contribution of free virus transmission for the five cases. Standard error bars are shown, and standard deviation bars are shown in the inset. (b) Histograms for the number of cells infected with a single mutant for the different strategies, representing 42 thousand hybrid simulations with size threshold $\mathcal{M} = 50$. Simulations are stopped when the infected cell population is close to peak infection ($1.59 \times 10^8$ cells). Here $S = 3$, $N = 26$, $\beta + \gamma = c = 3.6 \times 10^{-9}$, and the other parameters are as in Figure 1 with $R_0 = 8$.

We use varying rates of free virus transmission ($\beta$) and cell-to-cell transmission ($\gamma$) such that $\beta + \gamma = c$. In this way, we can compare the effect of synaptic transmission at different levels of contribution, while preserving $R_0$, the steady state values, and the overall rate of transmission. We will refer to different values of $\gamma$ (under a constant total transmission rate, $c$) as the different transmission strategies. Figure 7 shows the distributions for multiple transmission strategies for the single mutation model, close to peak infection. Here, we see that when synaptic transmission is included, the average number of mutants is maximized with only synaptic transmission. This is because synaptic transmission is better for generating mutants earlier and more often (here $S = 3$). As we decrease the amount of synaptic transmission, we greatly reduce the generation of mutant strains of virus, which results in a smaller number of cells infected with the mutant strain at peak infection. Not surprisingly, in the presence of synaptic transmission, the result of Section 5.1.1 remains valid: simulations with multiple infection correspond to larger numbers of mutants compared to simulations with single infection only.

## 5.2 Disadvantageous/advantageous mutants, complementation and interference

In this section, we examine the effect of multiple infection and synaptic transmission on mutant strains with varying fitness. We first consider the dynamics of non-neutral (disadvantageous and advantageous) single mutants. Then we consider scenarios where the mutant strain is disadvantageous when it is alone in a cell, but when a cell is coinfected with both the wild type strain and the mutant strain then both strains are neutral (complementation). Finally, we also consider the scenario of interference, when a single mutant (alone in a cell) is advantageous, but this advantage is removed once the cell is coinfected with a wild type virus.

### 5.2.1 Effect of multiple infection in the absence of synaptic transmission

We first assume only free virus transmission. Figure 8 shows the histograms of numbers of cells infected with disadvantageous mutants in the presence of complementation (a) and advantageous mutants in the presence of interference (b). Simulations with and without multiple infection were compared. As before, the number of cells was recorded near peak infection, and the maximum multiplicity parameter was for $N = 1$ for simulations with single infection only, and $N = 11$ for the model including multiple infection.

Again, we see that when multiple infection is included, there is a larger number of cells infected with the mutant virus at peak infection. We observe a 2.1-fold increase in the mean number of cells infected with the mutant, and a 3.3-fold increase in the standard deviation when multiple infection is included. In the case of disadvantageous mutations, when multiple infection is included, the effect of complementation comes into play by decreasing the disadvantage of the mutants (it does not in the single infection case as cells cannot be coinfected). Complementation helps disadvantageous mutant spread, as the disadvantageous mutant becomes neutral in coinfected cells.

In the case of advantageous mutants, we observe a 1.9-fold increase in the mean number of cells infected with the mutant, and a 1.5-fold increase in the standard deviation when multiple infection is included. The effect of interference here is the opposite, as it hinders advantageous mutant spread, because the advantageous mutant loses its fitness advantage in coinfected cells. However, the benefits of multiple infection to mutant spread outweighs this effect because the multiplicity of infection is still relatively small just before peak infection (about 1.9 copies of virus per infected cell on average), and therefore multiple infection still results in larger number of cells infected with the mutant virus.

### 5.2.2 Effect of multiple infection in the presence of synaptic transmission

As before, with multiple infection we can include synaptic transmission. Figures S2 and S4 show the number of cells infected with a disadvantageous/advantageous mutant near peak infection, and Figures S3 and S5 show the number of cells infected with a disadvantageous mutant with complementation/advantageous mutant with interference near peak infection. Furthermore, Figure 9 shows the mean number of mutants for the different transmission strategies and all fitness landscapes.

We first see that in all fitness landscapes considered, only synaptic transmission is the best for mutant evolution. This is again because synaptic transmission is superior at generating mutants, regardless of fitness landscape. We also see that at peak infection there are more advantageous mutants (without and with interference), followed by neutral mutants, followed by disadvantageous mutants (with and without complementation), because of fitness differences.

**Disadvantageous mutants.** For the disadvantageous mutant landscapes (Figures S2, S3, and 9(b-c)), we see that the mean number of cells infected with the mutant does not decrease monotonically with increasing contribution of free virus transmission. The reason is that only free virus is better at spreading the mutant, once it is generated, as free virus transmission can result in cells singly infected with the mutant, rather than coinfected with multiple copies of the wild type. This effect is most important for disadvantageous mutants, which have the most difficulty spreading after generation.

Since many cells infected with the mutant are coinfected with the wild type as well in the presence of synaptic transmission, in the scenario with complementation (Figures S3 and 9(c)),

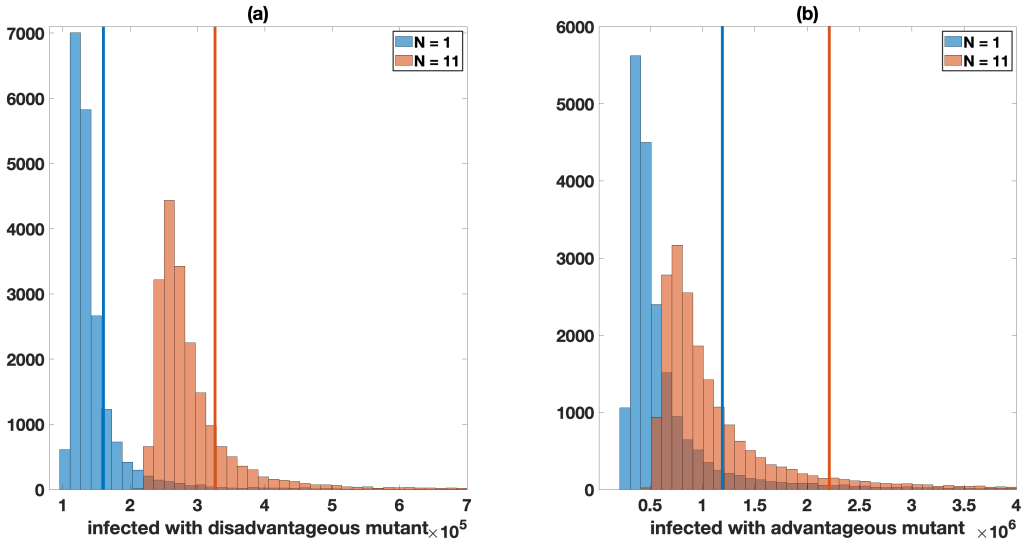


Figure 8: Non-neutral (disadvantageous and advantageous) mutant evolution in the presence of complementation and interference, under free virus transmission only. The histograms show the numbers of cells infected with a mutant virus in the absence and in the presence of multiple infection, and vertical bars represent the mean numbers. (a) Disadvantageous mutant with complementation; here $F_{\text{mutant}} = 0.9$. The average under single infection only is approximately $1.6 \times 10^5$ and for multiple infection is approximately $3.3 \times 10^5$. The T test between the two distributions gives a $p$-value less than $10^{-6}$. (b) Advantageous mutant with interference; here $F_{\text{mutant}} = 1.1$. The average for single infection is approximately $1.2 \times 10^6$ and for multiple infection is approximately $2.2 \times 10^6$. The T test between the two distributions gives a $p$-value less than $10^{-6}$. Histograms represent 20 thousand hybrid simulations. Simulations in which infections are not established (or in the rare case a simulation does not reach the infected size threshold) are discarded. Simulations are stopped when the infected cell population is close to peak infection ($6 \times 10^8$ cells). The other parameters are as in Figure 1 with $R_0 = 8$.

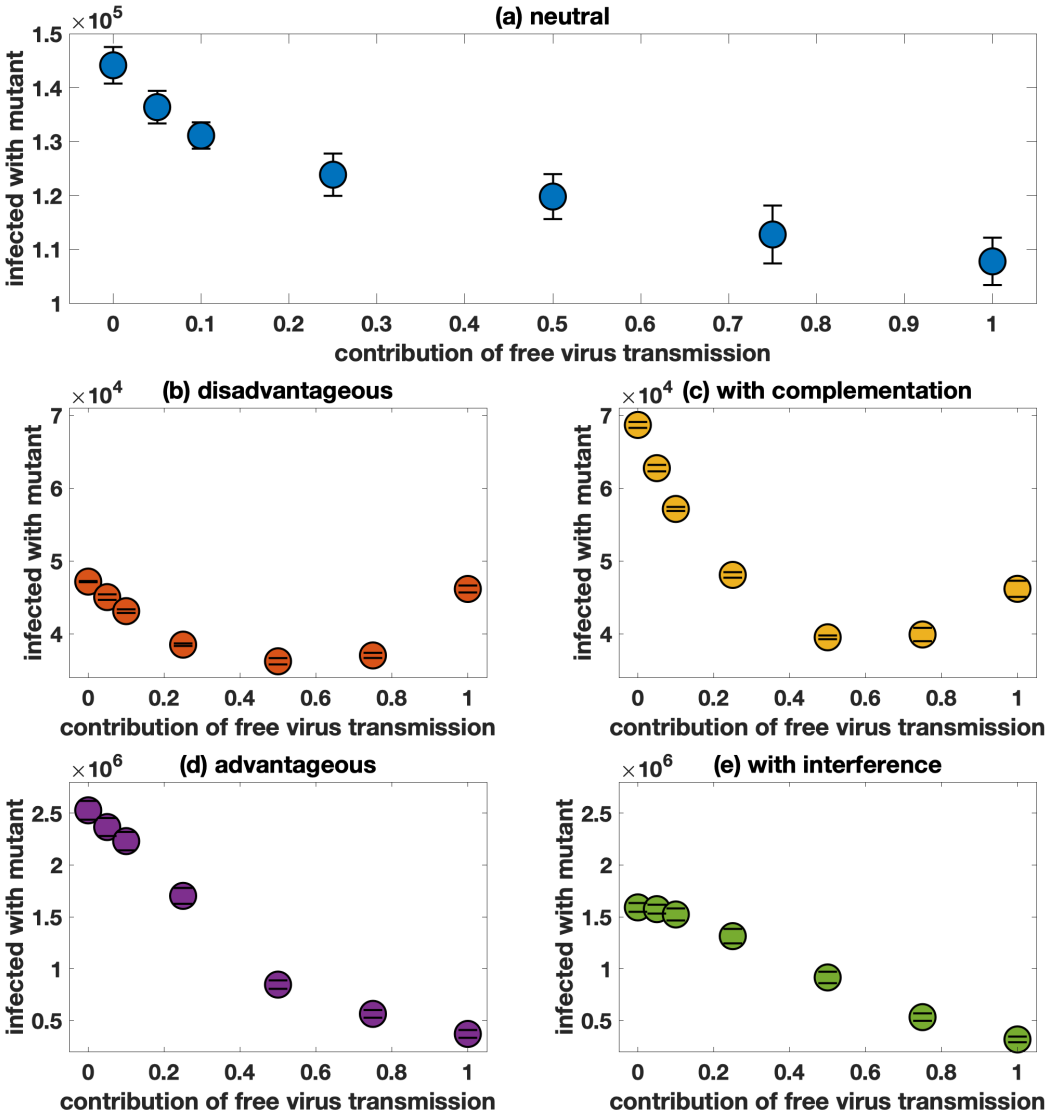


Figure 9: Mutant evolution with synaptic transmission. Comparison of the mean number of cells infected with the single mutant virus for different fitness landscapes. In all panels, the horizontal axis is the contribution of free virus transmission and the vertical axis is the mean number of cells infected with the mutant virus. Standard error bars for each point are shown. For all simulations, $\mathcal{M} = 50$, and simulations are stopped when the infected cell population is close to peak infection ($1.59 \times 10^8$ cells). (a) Neutral fitness landscape. Each point represents the average over at least 42 thousand simulations. (b) Disadvantageous fitness landscape ($F_{\text{mutant}} = 0.9$). Each point represents the average over at least 2 thousand simulations. (c) Disadvantageous fitness landscape ($F_{\text{mutant}} = 0.9$) with complementation. Each point represents the average over at least 2 thousand simulations. (d) Advantageous fitness landscape ($F_{\text{mutant}} = 1.1$). Each point represents the average over at least 10 thousand simulations. (e) Advantageous fitness landscape ($F_{\text{mutant}} = 1.1$) with interference. Each point represents the average over at least 10 thousand simulations. All other parameters are the same as in Figure 7.

we see that synaptic transmission becomes more powerful. This is because with complementation, the disadvantageous mutant gains fitness in coinfected cells, and larger contributions of synaptic transmission are correlated with higher number of coinfected cells. Therefore synaptic transmission is beneficial for complementation because it increases the multiplicity of infection and rate of coinfection (so mutant viruses are more likely to be with wild type virus in a cell). For only free virus transmission, complementation has a very small effect because not a lot of cells are coinfected with both the wild type and mutant.

**Advantageous mutants.** For the advantageous mutant landscapes (Figures S4, S5, and 9(d-e)), initial mutant generation is very important, since advantageous mutants will spread efficiently after generation. Synaptic transmission is better than free virus at generating mutants, so again we see that only (or mostly) synaptic transmission is optimal, and that the average number of cells infected with the mutant strain decrease monotonically with increasing contribution of free virus transmission.

However, since many cells infected with the mutant are coinfected with the wild type as well with high contributions of synaptic transmission, in the scenario with interference (Figures S5 and 9(e)), we see that synaptic transmission hinders mutants spread with interference compared to the non-interference case. This is because with interference, the mutant virus loses fitness in cells coinfected with the wild type. For only free virus transmission, interference has a very small effect as there are few cells that are coinfected with both the wild type and mutant.

# 6 Discussion

In this study, we developed a hybrid stochastic-deterministic framework to simulate *in vivo* HIV dynamics in order to assess the effect of multiple infection and synaptic cell-to-cell transmission on virus evolution. We analyzed the effect of multiple infection and synpatic transmission for different fitness landscapes, including disadvantageous mutants with complementation and advantageous mutants with interference. Furthermore, we also describe and characterize the differences between the stochastic simulations and deterministic predictions, which is an important but understudied phenomena that has large implications in mathematical modeling of biological processes and virus dynamics, as many classical models of virus dynamics only make use of ODEs.

Stochastic effects of virus dynamics are especially important in the context of rare mutation, because small cell populations can experience large amounts of fluctuation. However, large cell numbers and multiple subpopulations make fully stochastic methods such as the Gillespie method difficult and computationally intractable. Therefore, we adapted a hybrid stochastic-deterministic algorithm [29] which treats small populations stochastically and large populations deterministically in order to feasibly simulate evolutionary infection dynamics of this complex system. This algorithm has large potential for applications to other viral dynamics models as well as other models in the fields of population dynamics and evolutionary biology, where large population sizes also make fully stochastic simulations or agent-based models difficult.

An important parameter in the algorithm is the size threshold $\mathcal{M}$, which determines which populations are treated stochastically and and which ones can be considered large enough to be treated deterministically by the algorithm. Larger size thresholds capture more stochastic effects, but come at a large potential computational cost. A priori estimations of the size threshold are difficult, but here we introduce an analytical lower bound on the size threshold, using the fact that the rate of stochastic extinction using the hybrid algorithm should be the same as the rate

of stochastic extinction in a completely stochastic setting. This is an important step toward being able to determine a reasonable size threshold a priori.

We started by using the hybrid algorithm to compare the population averages over many stochastic runs with deterministic predictions. Focusing on the prediction of the number of cells infected with mutant virus, we found that for neutral and advantageous mutants, ODEs overpredict the number of infected cells, and for disadvantageous mutants (with the disadvantage above a threshold) ODEs underpredict the number of infected cells. The relative error depends on system parameters but generally decreases with $R_0$. Further, ODEs could be used to estimate the timing of mutant generation (by scaling the population variable to correspond to cell numbers and finding the time when the population of cells infected by the mutant virus reaches unity). We find that such ODE estimates underestimate the time of single- and double-mutant generation compared to the stochastic model.

Our study demonstrates that multiple infection is an important part of HIV infection. In the presence of multiple infection, higher number of mutants are expected. The reason is as follows. When multiple infection is included, there are more target cells in which to infect, which also results in faster generation of mutants relative to the total number of infected cells. We showed that at the peak infection, the number of cells infected with single- or double-mutant viruses is about twice as high in the simulations with multiple infections, compared to simulation with single infections only.

We also show the effect of synaptic transmission on the dynamics of an infection near peak infection. In general, the respective contributions of the two virus transmission pathways depend on the following trade-off: synaptic transmission is more effective at mutant generation, and free virus transmission and lower multiplicities are better at facilitating mutant spread, as the mutant viruses do not have to compete with the wild type strain (which is initially well-established) within a cell to spread. We investigated a spectrum of different transmission strategies ranging from purely synaptic to purely free-virus transmission, to estimate the number of cells infected with mutant viruses at infection peak. We found that in the case of neutral and advantageous viruses, purely synaptic transmission results in the largest number of cells infected with the mutant virus. This effect remains even in the case when the advantage of mutants is countered by interference (when coinfection with a wild type virus reverses the advantage conferred by the mutation, making the mutant virus neutral.) For disadvantageous mutants, we observe that both purely synaptic and purely free-virus transmission result in an increased number of cells infected by the mutant virus (the relative numbers depend on specific parameters). In the presence of complementation, when co-infection of a disadvantageous mutant with a wild type virus removes the disadvantage, we observe an increase in the mutant number for purely synaptic transmission.

While the hybrid method is capable of handling a large number of subpopulations, the number of "reactions" included in the stochastic part of the algorithm increases with (i) the number of different virus strains, (ii) the maximum multiplicity $N$, and (iii) the number of virus transferred per synapse $S$. If these parameters are too large, the number of reactions for the Gillespie algorithm can becomes too high to be computationally feasible (even if only small populations are handled stochastically). In general, the number of strains per model is $2^k$ and the number of differential equations (subpopulations) per model is $\binom{N+2^k}{2^k}$. The events in the model are uninfected cell generation (only one reaction), cell death ($\binom{N+2^k}{2^k}$ reactions), faithful infection, and infection with mutation. When there is only free virus transmission, the number of infection events is the number of strains multiplied by the number of subpopulations eligible to be infected, but when synaptic transmission is included, there are many more infection events, which is correlated with the number of ways to partition $S$ into $2^k$ non-negative integers that sum to $S$. When the number

of reactions is on the order of $10^4$, each simulation becomes very computationally expensive, which happens, for example, if we consider triple mutants in the presence of synaptic transmission.

Another limitation of our study is that synaptic transmission is non-spatial, which makes it relatively more powerful for promoting mutant generation. In the spatial case, synaptic transmission is restricted to spread to neighboring cells only, and so it is less dominant [18]. *In vivo* data from humanized mice indicate that synaptic transmission can result in spatial clusters of infected cells [19]. One way to incorporate this in future work is to extend our current model to include a system of spatial subdivided regions of cells where synaptic cell-to-cell transmission can only occur between cells within the same region, but free virus transmission can occur between all regions. Each region would be simulated stochastically with our hybrid stochastic-deterministic framework.

## References


[1] Christian L Althaus and Rob J De Boer. Dynamics of immune escape during hiv/siv infection. *PLoS Comput Biol*, 4(7):e1000103, 2008.

[2] Christelle Brégnard, Gregory Pacini, Olivier Danos, and Stéphane Basmaciogullari. Suboptimal provirus expression explains apparent nonrandom cell coinfection with hiv-1. *Journal of virology*, 86(16):8810–8820, 2012.

[3] Ping Chen, Wolfgang Hübner, Matthew A Spinelli, and Benjamin K Chen. Predominant mode of human immunodeficiency virus transfer between t cells is mediated by sustained env-dependent neutralization-resistant virological synapses. *Journal of virology*, 81(22):12582–12595, 2007.

[4] Tae-Wook Chun, Lucy Carruth, Diana Finzi, Xuefei Shen, Joseph A DiGiuseppe, Harry Taylor, Monika Hermankova, Karen Chadwick, Joseph Margolick, Thomas C Quinn, et al. Quantification of latent tissue reservoirs and total body viral load in hiv-1 infection. *Nature*, 387(6629):183–188, 1997.

[5] Armando Del Portillo, Joseph Tripodi, Vesna Najfeld, Dominik Wodarz, David N Levy, and Benjamin K Chen. Multiploid inheritance of hiv-1 during cell-to-cell infection. *Journal of virology*, pages JVI–00231, 2011.

[6] Christophe Fraser. Hiv recombination: what is the impact on antiretroviral therapy? *Journal of the Royal Society Interface*, 2(5):489–503, 2005.

[7] Helen R Fryer, John Frater, Anna Duda, Mick G Roberts, Rodney E Phillips, Angela R McLean, SPARTAC Trial Investigators, et al. Modelling the evolution and spread of hiv immune escape mutants. *PLoS Pathog*, 6(11):e1001196, 2010.

[8] Huub C Gelderblom, Dimitrios N Vatakis, Sean A Burke, Steven D Lawrie, Gregory C Bristol, and David N Levy. Viral complementation allows hiv-1 replication without integration. *Retrovirology*, 5(1):1–18, 2008.

[9] Vanessa M Hirsch. Evolution of the fittest ends in tragedy. *Nature medicine*, 5(5):488, 1999.

[10] David D Ho, Avidan U Neumann, Alan S Perelson, Wen Chen, John M Leonard, and Martin Markowitz. Rapid turnover of plasma virions and cd4 lymphocytes in hiv-1 infection. *Nature*, 373(6510):123, 1995.

[11] Wolfgang Hübner, Gregory P McNerney, Ping Chen, Benjamin M Dale, Ronald E Gordon, Frank YS Chuang, Xiao-Dong Li, David M Asmuth, Thomas Huser, and Benjamin K Chen. Quantitative 3d video microscopy of hiv transfer across t cell virological synapses. *Science*, 323(5922):1743–1747, 2009.

[12] Lina Josefsson, Martin S King, Barbro Makitalo, Johan Brännström, Wei Shao, Frank Maldarelli, Mary F Kearney, Wei-Shau Hu, Jianbo Chen, Hans Gaines, et al. Majority of CD4+ t cells from peripheral blood of HIV-1–infected individuals contain only one HIV DNA molecule. *Proceedings of the National Academy of Sciences*, 108(27):11199–11204, 2011.

[13] Lina Josefsson, Sarah Palmer, Nuno R Faria, Philippe Lemey, Joseph Casazza, David Ambrozak, Mary Kearney, Wei Shao, Shyamasundaran Kottilil, Michael Sneller, et al. Single cell analysis of lymph node tissue from HIV-1 infected patients reveals that the majority of CD4+ T-cells contain one HIV-1 DNA molecule. *PLoS pathogens*, 9(6):e1003432, 2013.

[14] Andreas Jung, Reinhard Maier, Jean-Pierre Vartanian, Gennady Bocharov, Volker Jung, Ulrike Fischer, Eckart Meese, Simon Wain-Hobson, and Andreas Meyerhans. Multiply infected spleen cells in hiv patients. *Nature*, 418(6894):144–144, 2002.

[15] Andreas Jung, Reinhard Maier, Jean-Pierre Vartanian, Gennady Bocharov, Volker Jung, Ulrike Fischer, Eckart Meese, Simon Wain-Hobson, and Andreas Meyerhans. Recombination: Multiply infected spleen cells in hiv patients. *Nature*, 418(6894):144, 2002.

[16] Jason T Kimata, LaRene Kuller, David B Anderson, Peter Dailey, and Julie Overbaugh. Emerging cytopathic and antigenic simian immunodeficiency virus variants influence aids progression. *Nature medicine*, 5(5):535, 1999.

[17] Natalia L Komarova and Dominik Wodarz. Virus dynamics in the presence of synaptic transmission. *Mathematical biosciences*, 242(2):161–171, 2013.

[18] Jesse Kreger, Natalia L Komarova, and Dominik Wodarz. Effect of synaptic cell-to-cell tranmission and recombination on the evolution of double mutants in HIV. *Journal of the Royal Society Interface*, 17(164):104–115, 3 2020.

[19] Kenneth M Law, Natalia L Komarova, Alice W Yewdall, Rebecca K Lee, Olga L Herrera, Dominik Wodarz, and Benjamin K Chen. In vivo hiv-1 cell-to-cell transmission promotes multicopy micro-compartmentalized infection. *Cell reports*, 15(12):2771–2783, 2016.

[20] David N Levy, Grace M Aldrovandi, Olaf Kutsch, and George M Shaw. Dynamics of hiv-1 recombination in its natural target cells. *Proceedings of the National Academy of Sciences*, 101(12):4204–4209, 2004.

[21] Louis M Mansky and Howard M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[22] Meriet Mikhail, Bin Wang, Philippe Lemey, Brenda Beckthold, Anne-Mieke Vandamme, M John Gill, and Nitin K Saksena. Role of viral evolutionary rate in hiv-1 disease progression in a linked cohort. *Retrovirology*, 2(1):41, 2005.

[23] Martin A Nowak, Roy M Anderson, Maarten C Boerlijst, Sebastian Bonhoeffer, Robert M May, Andrew J McMichael, Steven M Wolinsky, Kevin J Kunstman, Jeffrey T Safrit, Richard A Koup, et al. Hiv-1 evolution and disease progression. *Science*, pages 1008–1011, 1996.

[24] Martin A Nowak, Alun L Lloyd, Gabriela M Vasquez, Theresa A Wiltrout, Linda M Wahl, Norbert Bischofberger, Jon Williams, Audrey Kinter, Anthony S Fauci, Vanessa M Hirsch, et al. Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection. *Journal of virology*, 71(10):7518–7525, 1997.

[25] John E. Pearson, Paul Krapivsky, and Alan S. Perelson. Stochastic theory of early viral infection: Continuous versus burst production of virions. *PLOS Computational Biology*, 7(2):1–17, 02 2011.

[26] Alan S Perelson, Avidan U Neumann, Martin Markowitz, John M Leonard, and David D Ho. Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, 1996.

[27] Ruy M Ribeiro and Sebastian Bonhoeffer. Production of resistant hiv mutants during antiretroviral therapy. *Proceedings of the National Academy of Sciences*, 97(14):7681–7686, 2000.

[28] Ruy M Ribeiro, Li Qin, Leslie L Chavez, Dongfeng Li, Steven G Self, and Alan S Perelson. Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. *Journal of virology*, 84(12):6096–6102, 2010.

[29] Ignacio A. Rodriguez-Brenes, Natalia L. Komarova, and Dominik Wodarz. The role of telomere shortening in carcinogenesis: A hybrid stochastic-deterministic approach. *Journal of Theoretical Biology*, 460:144 – 152, 2019.

[30] Daniel Shriner, Allen G Rodrigo, David C Nickle, and James I Mullins. Pervasive genomic recombination of hiv-1 in vivo. *Genetics*, 167(4):1573–1583, 2004.

[31] Alex Sigal, Jocelyn T Kim, Alejandro B Balazs, Erez Dekel, Avi Mayo, Ron Milo, and David Baltimore. Cell-to-cell spread of HIV permits ongoing replication despite antiretroviral therapy. *Nature*, 477(7362):95, 2011.

[32] Xiping Wei, Sajal K Ghosh, Maria E Taylor, Victoria A Johnson, Emilio A Emini, Paul Deutsch, Jeffrey D Lifson, Sebastian Bonhoeffer, Martin A Nowak, Beatrice H Hahn, et al. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510):117, 1995.

[33] Dominik Wodarz, David N Levy, and Natalia L Komarova. Multiple infection of cells changes the dynamics of basic viral evolutionary processes. *Evolution letters*, 3(1):104–115, 12 2018.

[34] Pia Abel zur Wiesch, Roger Kouyos, Jan Engelstädter, Roland R Regoes, and Sebastian Bonhoeffer. Population biological principles of drug-resistance evolution in infectious diseases. *The Lancet infectious diseases*, 11(3):236–247, 2011.