# Comparing mutational pathways to lopinavir resistance in HIV-1 subtypes B versus C

Susana Posada-Céspedes[1,2], Gert Van Zyl[3,4], Hesam Montazeri[5], Jack Kuipers[1,2], Soo-Yon Rhee[6], Roger Kouyos[7,8], Huldrych F. Günthard[7,8], and Niko Beerenwinkel[1,2,*]

**1** Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
**2** SIB Swiss Institute of Bioinformatics, Basel, Switzerland
**3** Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa
**4** National Health Laboratory Service, South Africa
**5** Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
**6** Stanford University, United States
**7** Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland
**8** Institute of Medical Virology, University of Zurich, Zurich, Switzerland

* niko.beerenwinkel@bsse.ethz.ch

## Abstract

Although combination antiretoviral therapies seem to be effective at controlling HIV-1 infections regardless of the viral subtype, there is increasing evidence for subtype-specific drug resistance mutations. The order and rates at which resistance mutations accumulate in different subtypes also remain poorly understood. Here, we present a methodology for the comparison of mutational pathways in different HIV-1 subtypes, based on Hidden Conjunctive Bayesian Networks (H-CBN), a probabilistic model for inferring mutational pathways from cross-sectional genotype data. We introduce a Monte Carlo sampling scheme for learning H-CBN models on a large number of resistance mutations and develop a statistical test to assess differences in the inferred mutational pathways between two groups. We apply this method to the temporal progression of mutations conferring resistance to the protease inhibitor lopinavir in a large cross-sectional data set of South African individuals living with HIV-1 subtype C, as well as a genotype data set of subtype B infections derived from the Stanford HIV Drug Resistance Database and the Swiss HIV Cohort Study. We find strong support for different initial mutational events in the protease, namely at residue 46 in subtype B and at residue 82 in subtype C. Our results also show that mutations can accumulate along various alternative paths within subtypes, as opposed to a unique total temporal ordering. Furthermore, the maximum likelihood mutational networks for subtypes B and C share only 7 edges (Jaccard distance 0.802) and imply many different evolutionary pathways. Beyond HIV drug resistance, the statistical methodology is applicable more generally for the comparison of inferred mutational pathways between any two groups.

# Author summary

There is a disparity in the distribution of infections by HIV-1 subtype in the world. Subtype B is predominant in America, Western Europe and Australia, and most therapeutic strategies are based on research and clinical studies on this subtype. However, non-B subtypes represent the majority of global HIV-1 infections; e.g., subtype C alone accounts for nearly half of all HIV-1 infections. We present a statistical framework enabling the comparison of patterns of accumulating mutations in different HIV-1 subtypes. Specifically, we study lopinavir resistance pathways in HIV-1 subtypes B versus C, but the methodology can be generally applied to compare the temporal ordering of genetic events in different subgroups.

# Introduction

The emergence of drug resistant viral strains, a process driven by the evolutionary escape dynamics of HIV-1, limits the success of antiretroviral therapies. Today, HIV-1 infections are clinically manageable by using two or more antiretroviral drugs together, a treatment strategy known as combination antiretroviral therapy (cART) [1]. However, when viral replication is inadequately suppressed, the virus may acquire mutations which confer resistance to cART, and the regimen must be replaced [2]. A better understanding of the underlying evolutionary process leading to resistance is believed to be crucial for predicting therapy outcome [3–5] and designing effective therapy sequences [6].

Mutational pathways of HIV-1 under the selective pressure of several antiretroviral drugs have been studied by sequencing the viral genome derived from patients over the course of treatment [7–12]. However, such longitudinal data are not available for most antiretroviral therapies. To leverage information from large cohorts and cross-sectional studies, different statistical models have been proposed to investigate mutational pathways leading to drug resistance. These approaches include several probabilistic graphical models, such as Markov processes [13]; a Markov model incorporating information from phylogenetic trees [4]; mutagenetic trees [3, 14]; Bayesian networks [15–19]; discrete and continuous-time Conjunctive Bayesian Networks (CBN) [20, 21]; and Suppes-Bayes Causal Networks (SBCN) [22]; as well as a Cox proportional-hazards model which is used to identify pairs of resistance mutations, in which one mutation alters the hazard of the other one [6].

Most of the aforementioned methods have been applied to study the accumulation of drug resistance mutations in HIV-1 subtype B infections. As an exception, Deforche *et al.* [15, 16] combined observations from various subtypes to investigate dependencies among resistance mutation and polymorphisms using Bayesian networks. The inferred network was used to explain the lower prevalence of protease mutation 30N in subtypes G and A as compared to subtype B through an interaction with the polymorphic locus 89L/M. Indeed, there is increasing evidence of differences in mutation profiles and evolutionary rates among subtypes [23–29], but implications on the order of accumulating mutations have not been systematically addressed. Here, we investigate the rate and order of accumulation of drug resistance mutations in different HIV-1 subtypes. Specifically, we compare mutational pathways to lopinavir resistance in HIV-1 subtypes B versus C. Although HIV-1 subtype B is the best studied and most prevalent subtype in Europe and North America, subtype C alone accounts for nearly half of all HIV infections worldwide [30]. It is therefore important to understand whether the evolution of drug resistance in subtype C proceeds in a similar fashion as for subtype B.

We use the Hidden Conjunctive Bayesian Network (H-CBN) [31], an extension of the continuous-time Conjunctive Bayesian Network (CT-CBN) accounting for noisy genotypes, to infer lopinavir mutational pathways in HIV-1 subtypes B and C. The CT-CBN model encodes constraints on the temporal ordering among mutations by assuming that the occurrence of genetic events can depend on the occurrence of predecessors. While in tree-based models the number of direct predecessors is constrained to be at most one, this assumption is relaxed in the CT-CBN model, where multiple predecessors are allowed. The partial order among mutations is inferred from observed viral genotypes. However, because genotyping is error-prone, two different error models have been proposed for the CT-CBN [21, 31].

First, Beerenwinkel *et al.* [21] used a mixture model with two components to distinguish signal from noise. This model has been applied to learn mutational pathways in HIV under different selective pressures. The data sets originally analyzed included at most nine resistance mutations, but Montazari *et al.* [32] presented a Monte Carlo expectation-maximization algorithm for parameter estimation of the mixture model with hundreds of mutations. The mixture error model has, yet, several limitations. Every genotype that violates the ordering constraints is assumed to occur with equal probability regardless of, e.g., the number of violations. Moreover, as the number of mutations increases the chance of obtaining an error-free genotype decreases rapidly. For instance, with a 1% per locus error rate and 64 mutations, we expect only around 53% of the genotypes to be correct. The mutation network is, however, inferred exclusively from the portion of the data assigned to the signal component of the mixture model, which can quickly result in a large portion of the data being discarded.

Second, in their H-CBN extension of the CT-CBN, Gerstung *et al.* [31] introduced latent variables to explicitly model the noisy observation process, which is parameterized by a per-locus error rate. In contrast to the mixture model, genetic events that apparently violate the ordering constraints can be explained by the latent variables, and the assumption that all violations are equally likely is relaxed. Moreover, instead of using only compatible genotypes to infer the maximum likelihood network as in the mixture model, the H-CBN uses all observed genotypes in a weighted fashion. Inference of the H-CBN model has been implemented via maximum likelihood estimation, but the time complexity of the likelihood computation is exponential in the number of mutations. In practice, computation quickly becomes impractical as the number of genetic events grows beyond 14 mutations.

Here, we take advantage of the improved error model of the H-CBN, but address its limitation regarding scalability in the number of mutations by employing an approximation scheme for the estimation of model parameters. We assess the performance of our method on simulated data and compare it to the original H-CBN method. Furthermore, we incorporate an adaptive simulated annealing algorithm to infer the maximum likelihood mutational network from the data, including different moves to explore the discrete space of networks. The resulting model and inference methods, called H-CBN2, are implemented as part of the MC-CBN R-package available at https://github.com/cbg-ethz/MC-CBN.

We use the H-CBN2 method to infer evolutionary pathways to lopinavir resistance in HIV-1 subtypes B and C confirming previous knowledge on frequently observed patterns of resistance-conferring mutations in response to lopinavir treatment [33, 34]. We also devise a statistical test to assess the similarity between two CBN models, which is available at https://github.com/cbg-ethz/H-CBN2-comparison-test. When applied to subtypes B versus C, we find significant differences in their mutational pathways.

## Methods

We first recapitulate the probabilistic graphical model underlying this work, the H-CBN. Second, we introduce a new parameter inference method for the H-CBN model, as well as an improved structure learning algorithm based on adaptive simulated annealing. Third, we develop a statistical test to assess structural differences between two CBN models.

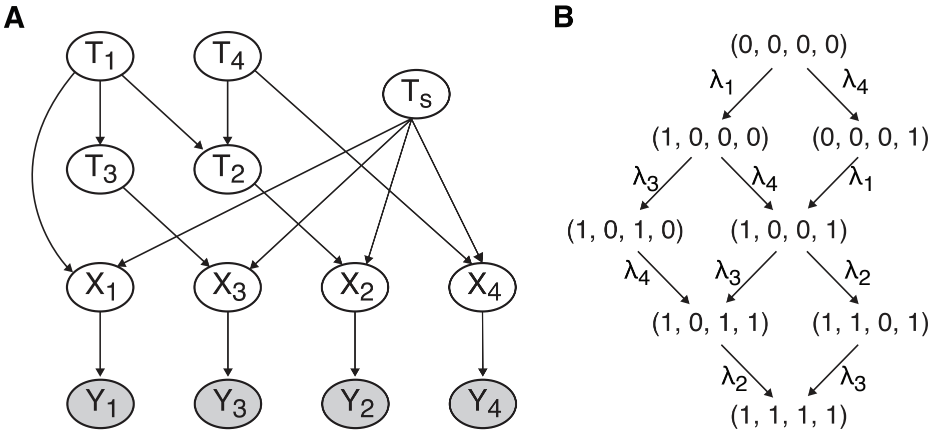
## Hidden Conjunctive Bayesian network

CBNs are probabilistic graphical models, in which a directed acyclic graph (DAG) represents the order in which genetic events may accumulate [20]. In the CT-CBN, the time between genetic events is modeled by independent exponential distributions [21]. The H-CBN extends the CT-CBN model by introducing hidden variables to model the error-prone observational process [31].

Formally, the CT-CBN is defined by a partially ordered set (poset) of genetic events, or mutations, and a rate for each mutation to occur. A poset $(P, \prec)$ consists of a set $P$ of size $p = |P|$ and a binary relation $\prec$. The relation $l \prec k$ indicates that mutation $l$ must take place before $k$. Further, a relation $l \prec k$ is a cover relation if $l \prec z \prec k$ implies $z = l$ or $z = k$. Drawing a directed edge from node $l$ to node $k$ for every cover relation $l \prec k$ yields a DAG which is transitively reduced and uniquely represents the poset (Fig 1A). It is therefore sufficient to consider transitively reduced DAGs only.

A genotype is a subset of genetic events of $P$, represented by a binary vector $x = (x_1, \ldots, x_p)$, where $x_j = 1$ indicates that mutation $j$ has occurred. A genotype $x$ is called compatible with the poset $P$ if $(x_l, x_k) \neq (0, 1)$ for all cover relations $l \prec k$. The collection of all genotypes compatible with $P$ is the space of all feasible mutational patterns, and it is denoted by $J(P)$ (Fig 1B).

**Fig 1. A partially ordered set of $p = 4$ mutations. A** The hidden variable $T_j$ is the waiting time to mutation $j$, and edges between mutation times encode the temporal ordering constraints among mutations. The hidden variable $X_j$ represents the true binary mutation status, whereas $Y_j$ denotes the corresponding error-prone binary observation. **B** The genotype lattice $J(P)$ consists of eight genotypes compatible with the poset. The lattice encodes all possible pathways from the wild-type (0,0,0,0) to the fully mutated type (1,1,1,1).



The waiting time to each mutation $j$ is represented by a random variable $T_j$. Their joint distribution is defined recursively as

$$T_j \sim Z_j + \max_{u \,\in\, \mathrm{pa}(j)} T_u, \quad Z_j \sim \mathsf{Exp}(\lambda_j), \tag{1}$$

where $\mathrm{pa}(j)$ denotes the set of parents of $j$ in the DAG, i.e., the set of mutations which precede mutation $j$. The random variable $Z_j$ is exponentially distributed with rate $\lambda_j$

and accounts for the time elapsed for generating and fixating mutation $j$, after its predecessors have occurred. The probability density of $T_j$ conditioned on the times to mutation of its parents $\mathrm{pa}(j)$ is defined as

$$f_T\left(t_j \,|\, (t_u)_{u \in \mathrm{pa}(j)}; \lambda_j\right) = \lambda_j \exp\left[-\lambda_j (t_j - \max_{u \in \mathrm{pa}(j)} t_u)\right] \mathbb{I}(t_j \geq \max_{u \in \mathrm{pa}(j)} t_u), \tag{2}$$

where the indicator function $\mathbb{I}$ encodes the ordering constraints of the poset. That is, the density function is zero if a mutation occurs before any of its predecessors.

The time at which every individual mutation emerges is generally unknown. Instead, patients are monitored with certain regularity, and oftentimes when the viral load increases, the virus population is sequenced. The sampling time is generally also unknown and typically differs among patients. To account for this uncertainty, an exponentially distributed random variable $T_s \sim \mathsf{Exp}(\lambda_s)$ is introduced. Hence, the observed data is censored, and a mutation $j$ occurred if and only if its waiting time $t_j$ was smaller than the sampling time $t_s$, i.e., $x_j = 1$ if $t_j < t_s$ and $x_j = 0$ otherwise. The model is not identifiable as long as the rate $\lambda_s$ is unknown. Therefore, unless known, this scaling factor is set to $\lambda_s = 1$ [21, 31].

There is another hidden process, namely the generation of viral genotype data. To account for false positives and false negatives, a variable $Y$ is introduced in the H-CBN model to denote the observed genotype, an error-bearing version of the true genotype $X$ [31] (Fig 1A). Assuming errors are independent and identically distributed across mutations, the probability of observing genotype $Y$ given the true underlying genotype $X$ is

$$\Pr(Y \,|\, X) = \epsilon^{d_H(X,Y)} (1-\epsilon)^{p - d_H(X,Y)}, \tag{3}$$

where $\epsilon$ is the per-locus error probability and $d_H$ is the Hamming distance.

The likelihood function for $N$ independent, observed genotypes $\mathcal{Y} = (Y^{(1)}, \ldots, Y^{(N)})$ is

$$\mathcal{L}_{\mathcal{Y}}(\lambda, \epsilon, P) = \prod_{i=1}^{N} \Pr\left(Y^{(i)}; \lambda, \epsilon, P\right), \tag{4}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)$ and the probability of each observed genotype $Y^{(i)}$ is

$$\Pr\left(Y^{(i)}; \lambda, \epsilon, P\right) = \sum_{X \in J(P)} \Pr\left(Y^{(i)} \,\middle|\, X\right) \Pr(X; \lambda, \epsilon, P). \tag{5}$$

The probability of the true genotype $X$ is defined in terms of the waiting times as discussed above,

$$\Pr(X; \lambda, \epsilon, P) = \int_{\mathbb{R}_{\geq 0}^{p+1}} \left(\prod_{j=1}^{p} f_Z(z_j; \lambda_j)\right) f_Z(z_s; \lambda_s)\, \mathbb{I}(z, z_s \vdash X)\, \mathrm{d}z, \tag{6}$$

where the indicator function $\mathbb{I}$ encodes mutation times that can give rise to genotype $X$, i.e., $z, z_s \vdash X$, $z_j = t_j - \max_{u \in \mathrm{pa}(j)} t_u$ $(j = 1, \ldots, p)$, and $f_Z(z_j; \lambda_j) = f_T\left(t_j \,|\, (t_u)_{u \in \mathrm{pa}(j)}; \lambda_j\right)$. Henceforth, we also set $z = (z_1, \ldots, z_p, z_s)$.

## Parameter estimation via Monte Carlo Expectation Maximization

Owing to censoring of mutation times and unobserved true genotypes, the Expectation Maximization algorithm (EM) has been previously used to obtain maximum likelihood estimates of model parameters $\epsilon$ and $\lambda_j$, $j = 1, \ldots, p$ [31]. To address the limitation on the scalability in the number of mutations, we develop a Monte Carlo Expectation Maximization algorithm (MCEM) to jointly estimate the error rate ($\epsilon$) and the conditional evolutionary rate parameters ($\lambda_j$, $j = 1, \ldots, p$) for a given poset $P$.

In the expectation step (E step) of the MCEM algorithm, we estimate the expected value of the complete-data log-likelihood $\ell_{\mathcal{X},\mathcal{Z},\mathcal{Y},}\left(\lambda, \epsilon\right)$ with respect to the current conditional distribution of the hidden data (i.e., the unobserved true genotypes $\mathcal{X} = (X^{(1)}, \ldots, X^{(N)})$ and mutation times $\mathcal{Z} = (Z^{(1)}, \ldots, Z^{(N)})$), given the observed genotypes $\mathcal{Y}$, as well as the current estimates of the parameters $\lambda^{(k)}$ and $\epsilon^{(k)}$

$$\begin{aligned}\mathbb{E}_{\mathcal{X},\mathcal{Z}|\left(\mathcal{Y},\lambda^{(k)},\epsilon^{(k)}\right)}\left[\ell_{\mathcal{Y},\mathcal{X},\mathcal{Z}}\left(\lambda,\epsilon\right)\right] &= \\ \sum_{x^{(1)}\,\in\, J(P)}\cdots\sum_{x^{(N)}\,\in\, J(P)}\int_{\mathbb{R}^{p+1}_{\geq 0}}\cdots\int_{\mathbb{R}^{p+1}_{\geq 0}}\prod_{i=1}^{N} f_{X,Z}&\left(x^{(i)}, z^{(i)}\,\middle|\, Y=y^{(i)};\lambda^{(k)},\epsilon^{(k)}\right) \\ &\ell_{\mathcal{Y},\mathcal{X},\mathcal{Z}}\left(\lambda,\epsilon\right)\, \mathrm{d}z^{(1)}\cdots \mathrm{d}z^{(N)},\end{aligned} \tag{7}$$

where $k$ denotes the current MCEM iteration. Assuming independent observations the complete-data log-likelihood is

$$\ell_{\mathcal{X},\mathcal{Z},\mathcal{Y},}\left(\lambda,\epsilon\right) = \sum_{i=1}^{N}\left[\log\Pr\left(Y = y^{(i)}\,|\,X = x^{(i)}\right) + \log f_Z\left(z^{(i)}\right)\right]. \tag{8}$$

Substituting Eq. (8) into Eq. (7) yields

$$\begin{aligned}\mathbb{E}_{\mathcal{X},\mathcal{Z}|\left(\mathcal{Y},\lambda^{(k)},\epsilon^{(k)}\right)}\left[\ell_{\mathcal{Y},\mathcal{X},\mathcal{Z}}\left(\lambda,\epsilon\right)\right] = \sum_{i=1}^{N}\sum_{x^{(i)}\,\in\, J(P)}\int_{\mathbb{R}^{p+1}_{\geq 0}} f_{X,Z}\left(x^{(i)}, z^{(i)}\,\middle|\, Y=y^{(i)};\lambda^{(k)},\epsilon^{(k)}\right) \\ \left[\log\Pr\left(Y = y^{(i)}\,|\,X = x^{(i)}\right) + \log f_Z\left(z^{(i)}\right)\right]\mathrm{d}z^{(i)}.\end{aligned} \tag{9}$$

According to Bayes' theorem,

$$f_{X,Z}\left(x, z\,\middle|\, Y = y;\lambda^{(k)},\epsilon^{(k)}\right) = \frac{\Pr\left(Y = y\,|\,X = x;\,\epsilon^{(k)}\right) f_Z(z;\,\lambda^{(k)})\mathbb{I}\left(z\vdash X\right)}{\Pr\left(Y = y;\lambda^{(k)},\epsilon^{(k)}\right)}. \tag{10}$$

We denote the numerator of Eq. (10) by $A^{(k)}\left(x, y, z\right)$ to obtain the expected complete-data log-likelihood

$$\begin{aligned}&\mathbb{E}_{\mathcal{X},\mathcal{Z}|\left(\mathcal{Y},\lambda^{(k)},\epsilon^{(k)}\right)}\left[\ell_{\mathcal{Y},\mathcal{X},\mathcal{Z}}\left(\lambda,\epsilon\right)\right] = \\ &\quad\sum_{i=1}^{N}\sum_{x^{(i)}\,\in\, J(P)}\int_{\mathbb{R}^{p+1}_{\geq 0}}\frac{A^{(k)}\left(x^{(i)}, y^{(i)}, z^{(i)}\right)}{\Pr\left(Y = y^{(i)};\lambda^{(k)},\epsilon^{(k)}\right)}\left[\log\Pr\left(Y = y^{(i)}\,|\,X = x^{(i)}\right) + \right. \\ &\qquad\qquad\left.\log f_Z\left(z^{(i)}\right)\right]\mathrm{d}z^{(i)}.\end{aligned} \tag{11}$$

For small H-CBN models, this integral has been computed by decomposing it into a sum of integrals over all possible maximal chains in the genotype lattice [21, 31]. However, the number of maximal chains is $p!$ in the worst case, where $p$ is the number of mutations. Moreover, the summation over all possible genotypes in $J(P)$ is bounded by the total number of unobserved true (binary) genotypes: $2^p$. For moderate to large numbers of mutations, the exact computation of the expected value thus becomes computationally infeasible. To overcome this limitation, we approximate the expected value (11) using importance sampling. The general idea is to generate $L$ samples of the unobserved true genotypes $x$ and the mutation times $z$ from a proposal distribution $Q(x, z)$. Then,

$$\begin{aligned}&\mathbb{E}_{\mathcal{X},\mathcal{Z}|\left(\mathcal{Y},\lambda^{(k)},\epsilon^{(k)}\right)}\left[\ell_{\mathcal{Y},\mathcal{X},\mathcal{Z}}\left(\lambda,\epsilon\right)\right] \approx \\ &\quad \frac{1}{L}\sum_{i=1}^{N}\sum_{l=1}^{L}\frac{1}{Q(x_l^{(i)}, z_l^{(i)})}\frac{A^{(k)}\left(x^{(i)}, y^{(i)}, z^{(i)}\right)}{\Pr\left(Y = y^{(i)}; \lambda^{(k)}, \epsilon^{(k)}\right)}\left[\log\Pr\left(Y = y^{(i)} \mid X = x^{(i)}\right) + \right. \\ &\qquad\qquad \left. \log f_Z\left(z^{(i)}\right)\right]. \qquad (12)\end{aligned}$$

Intuitively, we would like to draw samples from the important region, e.g., samples that are likely to have given rise to the observed data. We use two types of importance sampling schemes, which we refer to as the forward and backward sampling, and implement and compare several variations of them (see next subsections).

In the maximization step (M step), we are concerned with maximizing Eq. (11) with respect to the parameters $\epsilon$ and $\lambda_j$, $j = 1, \ldots, p$. The maximum likelihood (ML) estimate $\hat{\epsilon}$ of the error rate $\epsilon$ is found to be the conditional expectation of the sufficient statistic $d_H(X, Y)$ obtained in the E-step,

$$\hat{\epsilon}^{(k)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{x^{(i)} \in J(P)}\int_{\mathbb{R}_{\geq 0}^{p+1}} A^{(k)}\left(x^{(i)}, y^{(i)}, z^{(i)}\right)\frac{1}{p}d_H(x^{(i)}, y^{(i)})\mathrm{d}z^{(i)}}{\sum_{x^{(i)} \in J(P)}\int_{\mathbb{R}_{\geq 0}^{p+1}} A^{(k)}\left(x^{(i)}, y^{(i)}, z^{(i)}\right)\mathrm{d}z^{(i)}}. \qquad (13)$$

Similarly, the ML estimate for the rate parameters $\hat{\lambda}_j$ are,

$$\hat{\lambda}_j^{(k)} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{x^{(i)} \in J(P)}\int_{\mathbb{R}_{\geq 0}^{p+1}} A^{(k)}\left(x^{(i)}, y^{(i)}, z^{(i)}\right) z_j^{(i)}\mathrm{d}z^{(i)}}{\sum_{x^{(i)} \in J(P)}\int_{\mathbb{R}_{\geq 0}^{p+1}} A^{(k)}\left(x^{(i)}, y^{(i)}, z^{(i)}\right)\mathrm{d}z^{(i)}}\right]^{-1}. \qquad (14)$$

The estimates can also be written more succinctly as

$$\begin{aligned}\hat{\epsilon}^{(k)} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{X,Z|Y,\lambda^{(k)},\epsilon^{(k)}}\left[\frac{1}{p}d_H(x^{(i)}, y^{(i)})\right] \\ \left[\hat{\lambda}_j^{(k)}\right]^{-1} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{X,Z|Y,\lambda^{(k)},\epsilon^{(k)}}\left[z_j^{(i)}\right]\end{aligned}$$

### Forward sampling

Assume the rate parameters $\lambda$ and the poset $P$ are known. We generate a candidate error-free genotype $x$ by sampling the mutation and sampling times $z = (z_1, \ldots, z_p, t_s)$ from the corresponding exponential distributions as follows

$$z_j \sim \mathsf{Exp}(\lambda_j), \; j = 1, \ldots, p, \quad t_s \sim \mathsf{Exp}(\lambda_s).$$

To determine the waiting times $t = (t_1, \ldots, t_p)$ we set $t_j = z_j + \max_{u \in \mathrm{pa}(j)} t_u$. Whenever, the waiting time $t_j$ for mutation $j$ is smaller than the sampling time $t_s$, we record that the mutation $j$ has been observed. If we do this for every mutation $j$, we obtain a sample of an error-free genotype $x = (x_1, \ldots, x_p)$. We draw samples by traversing the DAG in topological order to ensure that we compute $t_u$ for all $u \in \mathrm{pa}(j)$ before visiting any dependent mutation $j$. Because we do not know the rate parameters

$\lambda$, nor the poset $P$, in each iteration of the MCEM algorithm, we use their current estimates, $\lambda^{(k)}$ and $P^{(k)}$.

For each observed genotype $y^{(i)}$, $i = 1, \ldots, N$, we draw $L$ samples using the forward sampling scheme described above. A sample is a tuple of waiting times and the corresponding error-free genotype. Because of the graph traversal and the loop over parents, the worst-case time-complexity of the forward sampling is $O(NLp^2)$. We note that the candidate hidden genotypes are generated without accounting for the observed data. Alternatively, we implement a second forward sampling scheme called *forward-pool*. In this case, for each iteration of the MCEM algorithm, we draw an initial pool of $K$ waiting times vectors ($t_j^{(l)}$, $j = 1, \ldots, p$), with $K \gg L$, and for each observed genotype, we choose a subset of $L$ samples according to their similarity to the observed genotype as explained below. For each of the waiting times samples, we first construct the error-free genotype $x^{(l)}$ and then draw $L$ genotypes, each with probability

$$q_l = \frac{\epsilon^{d_H(y^{(i)}, x^{(l)})}(1-\epsilon)^{p - d_H(y^{(i)}, x^{(l)})}}{\sum_{l=1}^{K} \epsilon^{d_H(y^{(i)}, x^{(l)})}(1-\epsilon)^{p - d_H(y^{(i)}, x^{(l)})}}. \tag{15}$$

### Backward sampling

For the backward sampling, we construct the sample of candidate error-free genotypes $x^{(l)}$, $l = 1, \ldots, L$, based on the observed genotype $y^{(i)}$ and then sample the mutation times as

$$z_j \sim \begin{cases} \mathsf{TExp}(\lambda_j,\, 0,\, t_j - \max_{u \,\in\, \mathrm{pa}(j)} t_u) & \text{if } x_j = 1 \\ \mathsf{Exp}(\lambda_j) & \text{otherwise,} \end{cases} \tag{16}$$

where $\mathsf{TExp}$ is a truncated exponential distribution. Montazeri *et al.* [32] have used Eq. (16) to generate mutation times only from the compatible genotypes while using a mixture error model. Here, we extend this approach to also include sampling of the hidden layer modeling the genotyping errors, which enables us to account for all the observations.

We implement three variations of backward sampling to construct the sample of candidate hidden (true) genotypes. For the first strategy, we generate the genotypes $x^{(l)}$ by enumerating all compatible genotypes within Hamming distance $k$ of the observed genotype $y^{(i)}$, typically with $k \leq 3$. We then draw $L$ waiting-time vectors for each candidate genotype according to Eq. (16). This sampling scheme is referred to as Hamming $k$-neighborhood sampling. In the second strategy, we sample candidate genotypes by altering individual mutations of the observed genotype using $p$ independent Bernoulli trials, one for each mutation $j = 1, \ldots, p$, with success probability equal to the current estimate of the error rate $\hat{\epsilon}^{(k)}$. We draw $L$ candidate genotypes some of which may be incompatible with the current poset $P^{(k)}$ and, thus, obtain a zero sampling weight; i.e., they do not contribute to the estimation of the model parameters. This sampling scheme is referred to as Bernoulli sampling. The third approach is a two-step scheme. First, we decide uniformly at random whether to (i) leave the genotype $y^{(i)}$ unperturbed, (ii) add, or (iii) remove a mutation. For (ii) and (iii), we draw a mutation from the set of mutations that can be added or removed, respectively. If we remove an event $j$, it is chosen with probability proportional to $\kappa_j = \frac{1}{\lambda_j} + \max_{l \,\in\, \mathrm{pa}(j)} \kappa_l$, which corresponds to a greedy approximation of the time to mutation assuming that the process is dominated by the slowest predecessor in each reverse breadth-first search generation. The rationale is to remove mutations from the genotype $y^{(i)}$ that are likely to occur at later times with higher probability. On the other hand, if we add an event, it is chosen with a probability which is inversely proportional to the probability of being removed. In this case, we add mutations that can arise faster with higher probability. In the second step of this scheme, we ensure the genotype is compatible with the current poset $P^{(k)}$ by adding or removing all incompatible mutations. This sampling scheme is referred to as the backward-add/remove (backward-AR) sampling.

### Evaluation of sampling schemes

We evaluate the accuracy of the different approximation schemes by computing the probability of a genotype $y^{(i)}$ and comparing it to the exact solution (Eq. 5). Since $\Pr(Y = y^{(i)})$ are the factors of the likelihood, we are assessing the accuracy of the likelihood computation. We approximate the probability of genotype $y^{(i)}$ by drawing $L$ samples from each of the proposal distributions,

$$\Pr(Y = y^{(i)}) \approx \frac{1}{L} \sum_{l=1}^{L} \frac{\Pr\left(Y = y^{(i)} | x_l^{(i)}\right) \Pr(x_l^{(i)})}{Q(x_l^{(i)}, z_l^{(i)})}. \qquad (17)$$

## Structure learning

Gerstung *et al.* [31] implemented a simulated annealing (SA) algorithm with a geometric annealing schedule to infer the network structure of the H-CBN model. However, as the size of the model increases, the poset search space increases rapidly (sequence A001035 in The On-Line Encyclopedia of Integer Sequences, https://oeis.org/A001035), and the standard SA algorithm is more prone to converge to local optima and to miss globally optimal or near-optimal solutions. Here, we incorporate an adaptive simulated annealing (ASA) algorithm [35] to improve the efficacy of the search. As in the standard SA algorithm [36], in each iteration, we propose an update $P'$ of the current poset $P^{(k)}$ and accept the new poset with probability

$$\min\left(1, \exp\left(\frac{-\left[\ell_Y\left(\hat{\lambda}^{(k)}, \hat{\epsilon}^{(k)}, P^{(k)}\right) - \ell_Y\left(\hat{\lambda}', \hat{\epsilon}', P'\right)\right]}{\Theta^{(k)}}\right)\right).$$

Conventionally, the temperature $\Theta$ is gradually reduced over iterations, initially allowing the system to explore a broad region of the search space, but ultimately moving exclusively towards solutions that improve the likelihood. In the ASA algorithm, the cooling schedule is adjusted according to the search progress, but following the same principle as before, i.e., gradually changing the temperature such that the system is able to converge [37, 38]. We have adopted the cooling schedule from Srivatsa et al. [39] as follows. For every interval consisting of $m$ consecutive iterations, we set the temperature $\Theta_m = \Theta_{m-1} \exp\left(\left(0.5 - a_{m-1}^c\right) a_r\right)$, where $a_{m-1}$ is the observed acceptance rate of the previous interval, $a_r$ is a custom adaptation rate, and $c = \frac{-\log(2)}{\log a_{\text{ideal}}}$ is a scaling factor accounting for deviations from an optimal acceptance rate. Following the previous work [39], the optimal acceptance rate is set to $a_{\text{ideal}} = 1/p$, where $p$ is the number of mutations. Moreover, the adaptation rate $a_r$ is an additional free parameter enabling to further control the abruptness of temperature changes.

The optimization includes proposing a neighboring poset, which ultimately defines how we explore the space of posets. To this end, we implement three move types: (i) add or remove an edge, (ii) add an element to or remove an element from the cover relations while preserving all the remaining ones, and (iii) swap node labels. When adding an element to the cover relation, or equivalently an edge in the DAG, we discard proposed networks which are not transitively reduced or contain cycles.

## Implementation

We collectively refer to the implementation of the methods described in the previous sections as H-CBN2. It consists of the importance sampling schemes for parameter inference and the adaptive simulated annealing algorithm for structure learning of the H-CBN model. The code has been integrated into the MC-CBN R-package. We used C++ with OpenMP and the Boost libraries to ensure computational efficiency. We also employed the Vector Statistics component of the Intel Math Kernel Library (MKL) to efficiently generate random numbers.

## Statistical test for the comparison of CBN models

To compare two CBN models, we compare the posets using the Jaccard distance. The Jaccard distance between two sets is the complement of their Jaccard index, which is obtained by dividing the cardinality of the intersection by the cardinality of the union of the two sets.

Based on this notion of distance, we develop a permutation test to assess whether two given posets differ significantly from each other. Given two CBN models (e.g., estimated separately for HIV-1 subtypes B and C), we compute the Jaccard distance between the posets, $d_J$. The test quantifies how likely it is to observe the distance $d_J$ under the null hypothesis of both data sets having been generated by the same underlying poset. The alternative is that the two data sets have been generated by two different posets.

We compute the distribution of the test statistic $D_J$ under the null hypothesis as follows. We combine all genotypes from the considered groups and randomly split the data into two disjoint sets $S_1$ and $S_2$ with $N_1$ and $N_2$ genotypes, respectively, where $N_1$ and $N_2$ are the sizes of the two original data sets. That is, we permute the group labels of the genotypes. Then we apply H-CBN2 to infer the poset for $S_1$ and $S_2$ separately and compute their Jaccard distance. We repeat this procedure $B$ times and construct the distribution of the test statistic $D_J$ under the null by aggregating the computed Jaccard distances (Fig 2). We assess how likely it is to observe a test statistic at least as extreme as $d_J$ under the null hypothesis by means of computing the associated p-value

$$\Pr(D_J > d_J \mid \mathcal{H}_0) = 1 - \hat{\mathrm{F}}(d_J) \quad , \tag{18}$$

where $\hat{\mathrm{F}}(d_J)$ is the empirical cumulative distribution function,

$$\hat{\mathrm{F}}(d_J) = \frac{1}{B} \sum_{S_1,\, S_2 \,\in \mathcal{S}_B} \mathbb{I}(d_J(S_1, S_2) \leq d_J), \tag{19}$$

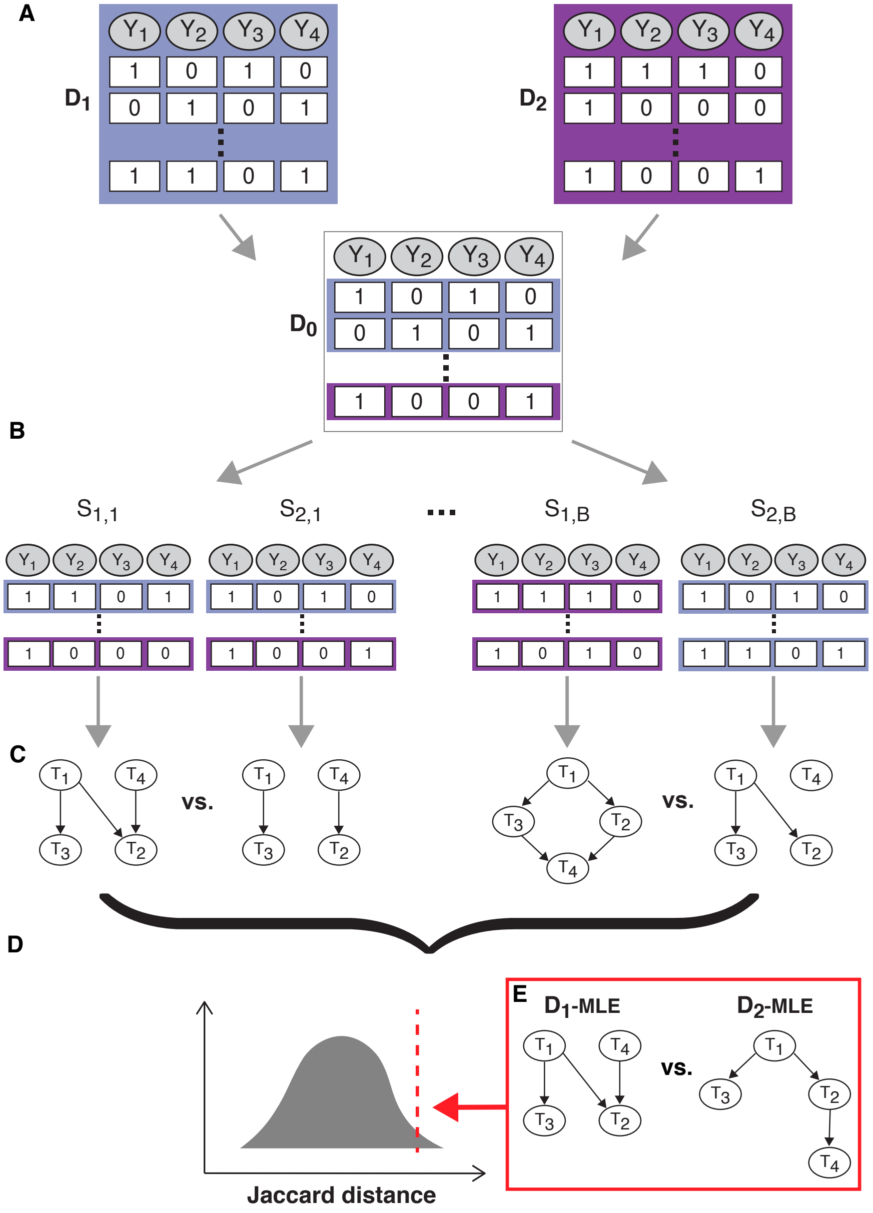
where $\mathcal{S}_B = \{(S_{1,j}, S_{2,j}) : j = 1, \ldots, B\}$. For the comparisons, we choose a significance level of 5% and perform $B = 50$ permutations of the group labels.

The code for performing this distance-based test is available at https://github.com/cbg-ethz/H-CBN2-comparison-test.git.

# Genotype data sets

We study lopinavir mutational pathways in three data sets: *(i)* a cohort of 1065 South African patients living with HIV-1 subtype C retrieved from the Stanford HIV Drug Resistance Database (HIVDB, File S1) [40, 41], *(ii)* a data set of 470 sequences of subtype B genotypes obtained from the HIVDB (File S2) and the Swiss HIV Cohort Study (SHCS) [42, 43], and *(iii)* a data set of 755 sequences of subtype C genotypes obtained from the HIVDB excluding genotypes from South African patients contained in the first data set (File S3).

**Fig 2. Schematic representation of the comparison of CBN models. A** Data sets $D_1$ and $D_2$ consist of $N_1$ and $N_2$ genotypes, respectively, and, in this example, $p = 4$ mutations. We combined both data sets into a single one $D_0$ with $N_1 + N_2$ genotypes. **B** We randomly split data set $D_0$ into data sets $S_1$ and $S_2$ and we do so $B$ times. **C** For each data set, we apply the H-CBN2 approach to learn the structure of the network and for each pair, $S_1$ and $S_2$, we compute the Jaccard distance. **D** The empirical distribution of the test statistic is computed by aggregating the distances between pairs $S_1$ and $S_2$. **E** We compare the inferred CBN posets from original data sets $D_1$ and $D_2$ by computing the Jaccard distance and assess its significance.



The HIVDB is a publicly available database that systematically aggregates data from published studies about HIV drug resistance. The SHCS is a nation-wide, prospective observational study covering approximately 75% of all treated patients in

Switzerland [43]. The SHCS has been approved by by the following ethical committees of all participating institutions: Kantonale Ethikkommission Bern; Ethikkommission beider Basel; comité d'éthique du département de médecine de Hôpitaux Universitaires de Genève; commission d'éthique de la recherche clinique, Lausanne; comitato etico cantonale, Bellinzona; Ethikkommission des Kanton St.Gallens; and Ethik-Kommission Zürich, all Switzerland. Written informed consent has been obtained from all participants.

# Results

We first evaluate and compare the different importance sampling schemes implemented in H-CBN2 for the scalable inference of H-CBN models in a simulation study. We then apply the best performing H-CBN2 approach to investigate the accumulation of lopinavir resistance-associated mutations in HIV-1 in a large South African cohort. Finally, we compare the results for HIV-1 subtype C to a data set of HIV-1 subtype B genotypes derived from lopinavir-treated patients and obtained from the HIVDB and the SHCS.

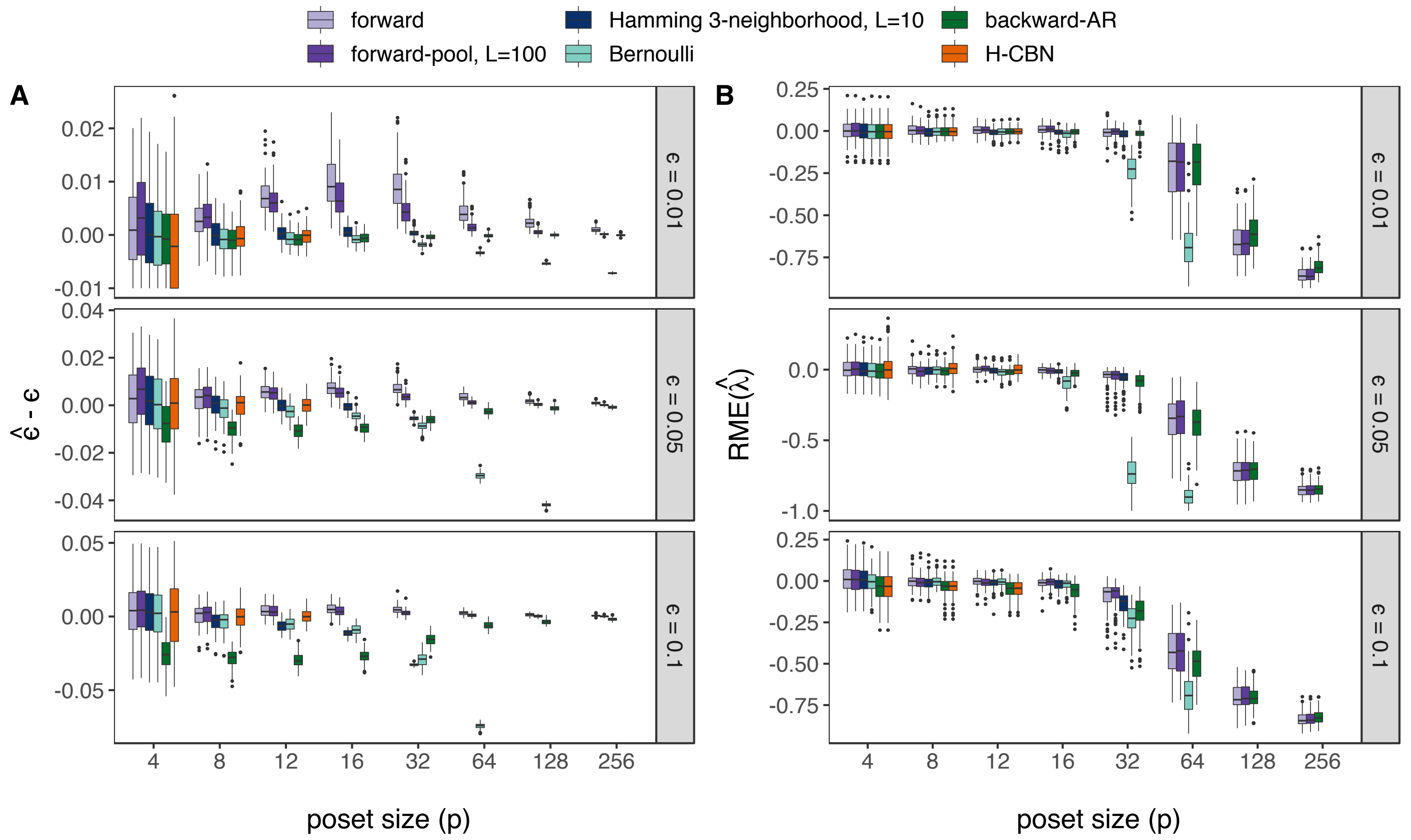
## Simulation study

We assess the quality of approximating the probability of a genotype $y$ (Eq. 17) by the H-CBN2 importance sampling schemes and compare it to the exact solution (Eq. 5). For a poset with $p = 16$ mutations, we vary the number of samples, $L$, drawn from the proposal distribution and find, as expected, that the accuracy of the approximation improves with $L$ (Fig S1–Fig S5). In most cases, we are able to accurately approximate the likelihood of the H-CBN model (Eq. 4) with $L = 1000$ (relative absolute error $\leq 0.02$, Table S1). For the forward-pool sampling, the relative absolute error of the log-likelihood depends almost exclusively on the the size of the initial pool $K$ for $L \geq 10$. Similarly, the approximation accuracy of the Hamming $k$-neighbourhood sampling is primarily determined by the extent $k$ of the considered neighborhood.

### Assessment of importance sampling schemes on simulated data

We assess accuracy and runtime performance of parameter estimation using the H-CBN2 sampling schemes on simulated data sets and compare it to the H-CBN model [31]. We simulate data sets with three different error rates ($\epsilon = 0.01$, 0.05, and 0.10) and various numbers of mutations ($p = 4$, 8, 12, 16, 32, 64, 128, and 256). For each error rate and number of mutations, we generate 100 data sets with different rate parameters and different transitively reduced DAGs. We draw the rate parameters $\lambda_j$ uniformly at random from the interval $\left[\frac{\lambda_s}{3}, 3\lambda_s\right]$ ($\lambda_s = 1$). We set the number of genotypes to $N = \min(50\,p, 1000)$, where the upper limit is motivated by the number of genotypes available in our application, i.e., comparing mutational pathways in different HIV-1 subtypes under lopinavir treatment. We use a fixed number of samples drawn from the proposal distribution for each of the sampling schemes (Fig 3). We draw $L = 10$ samples for the Hamming 3-neighborhood sampling, $L = 100$ samples for the forward-pool sampling, and $L = 1000$ for the other sampling schemes. These choices are motivated by the preceding results on the quality of the log-likelihood approximation via importance sampling. Then, we evaluate the performance of the H-CBN2 sampling schemes based on the deviation from the true value of the estimated error rate $\hat{\epsilon}$ and rate parameters $\hat{\lambda}$. To summarize results for all the different rates $\hat{\lambda}_j$, we compute the relative (median) error, which is given by $\frac{\text{median}(\hat{\lambda}-\lambda)}{\text{median}(\lambda)}$. Generally, we observe that for a known poset $P$, the estimation of the error rate and the rate parameters is accurate for small- and medium-sized posets (of up to about $p = 32$ mutations) under the evaluated conditions in terms of $N$ and $L$. For posets with up to 12 mutations, we can compare results to the H-CBN model, finding that most schemes perform as well as the H-CBN.

**Fig 3. Assessment of parameter estimation for various numbers of mutations, error rates, and posets. A** Box plots of the difference between true ($\epsilon$) and estimated ($\hat{\epsilon}$) error rate (y-axis) for 100 simulated data sets for each of the evaluated model sizes (x-axis). **B** Box plots of the relative median error (RME; y-axis) of the estimated rate parameters $\hat{\lambda}$. Different colors indicate different importance sampling schemes. The sample size is $N = \min(50\,p, 1000)$ and the number of samples drawn from the proposal distribution is set to $L = 1000$ unless specified otherwise. We run 100 iterations of the Monte Carlo EM algorithm.



When estimating the error rate $\epsilon$, we observe that the forward sampling schemes tend to overestimate small error rates, while the backward sampling schemes tend to underestimate high error rates. In all cases, the variance of the estimated error rate decreases as the number of mutations increases and, in most cases, $\hat{\epsilon}$ converges to the true value. Recalling that the error rate is defined per locus, with an increase in the number of mutations and in the number of genotypes, we have more power to estimate $\epsilon$. As an exception, the estimates obtained by the Bernoulli sampling deteriorate as the number of mutations increases. This is because the fraction of incompatible genotypes increases with the number of mutations and it becomes less likely to sample candidate genotypes with non-zero weight (i.e., compatible with the poset). In fact, the Bernoulli sampling scheme failed to provide any samples for the data sets with 5% error rate and 256 mutations, as well as for the data sets with 10% error rate and 128 and 256 mutations.

By contrast, the relative error in estimating the rate parameters $\lambda$ increases with the number of mutations. This is likely due to the bounded number of genotypes to 1000 for data sets with more than 32 mutations. For this particular constellation of sample size $N$ and number $L$ of samples drawn from the proposal distribution, for posets with more than 32 mutations the sampling schemes fall short of accurately estimating the rate parameters. We also evaluate the Hamming $k$-neighborhood sampling for different $k$ (Fig S6 E). We observe that as the number of mutations increases, we need to expand the neighborhood based on the Hamming distance for accurate estimation of the model parameters. However, the run time increases substantially and becomes a limiting factor for larger posets (Fig S7). Once again, the performance of the Bernoulli sampling scheme declines as the number of mutations increases, and for data sets with more than 64 mutations, the relative median error for the rate parameters is outside the range displayed in Figure 3.

We also assess the run time performance of various sampling schemes implemented in H-CBN2 and compare to H-CBN (Fig S7). Each run corresponds to 100 iterations of the MCEM or EM algorithm. We observe that H-CBN is faster than any of the H-CBN2 sampling schemes for posets with up to 6 mutations. Nonetheless, in most cases, we find an almost linear relationship between the number of mutations and the run time of the H-CBN2 sampling schemes, whereas the H-CBN run time grows exponentially with the number of mutations and is outperformed by H-CBN2 for $p \gtrsim 10$. We also observe that, for larger posets, the forward-pool sampling is slower than the standard forward sampling, because the size of the pool increases with the number of mutations; we set $K = pL$ to assure accurate parameter estimates (Fig S2). As the number of mutations increases, the computation time of the Hamming distance becomes the limiting factor (Eq. 15).

The forward sampling and the backward-AR sampling perform equally well in terms of accuracy of the estimated model parameters for small- and medium-sized posets, even when the number $L$ of samples drawn from the proposal distribution is set to 100 for the backward-AR sampling (Fig S6 G-H). The run times of these sampling schemes with $L = 1000$ and $L = 100$, respectively, are also similar. The forward and the backward-AR sampling schemes thus enable performant parameter estimation for posets with more than 14 mutations and up to about 32 mutations. Since we do not observe any advantage in using the backward-AR sampling over the forward sampling, we choose the latter for all further analyzes presented in this work.

## Assessment of the simulated annealing algorithm on simulated data

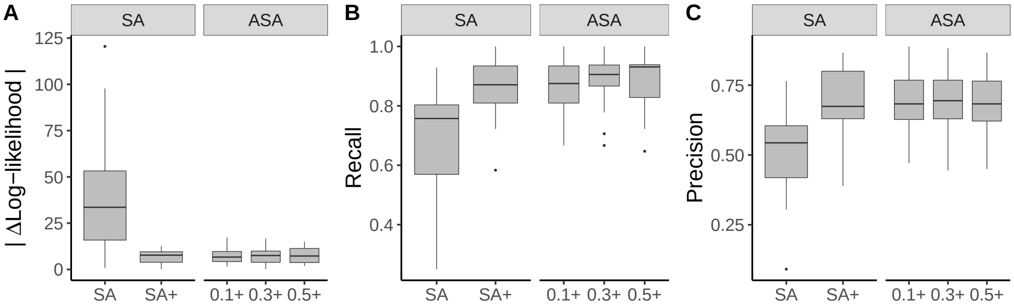
So far, we assumed that the poset $P$ is known. In the following, we evaluate the performance of the H-CBN2 structure learning algorithm, which, in addition to adding or removing an edge, includes new moves to propose candidate posets, as well as an ASA schedule. We employ a similar approach as before: (i) draw a transitive reduced DAG and parameters at random, (ii) generate a data set from the joint probability distribution of the model, and (iii) infer the network structure in addition to the model parameters.

We first compare the accuracy of the estimated model parameters when the poset $P$ is also learned. We do not observe any manifest difference in the absolute error between the true and the estimated error rate (Fig S8 A), but the relative absolute error of the rate parameters is marginally larger when the poset is learned in addition to the model parameters, as well as the absolute error of the log-likelihood (Fig S8 B-C).

Next, we compare different SA strategies for structure learning. We observe a notable improvement in the log-likelihood of the reconstructed network after including the additional new moves (SA+) compared to a simulated annealing algorithm (SA) with only addition and removal of edges (Fig 4A). Incorporating, in addition, an adaptive annealing schedule yields similar performance to SA+. Similarly, the error in estimation of the model parameters also decreases mostly upon including the new moves

(Fig S9 A-B). We also compute the recall and precision of reconstructing the elements of the cover relation and find a clear improvement of SA+ over SA (Fig 4B-C).

**Fig 4. Evaluation of the adaptive simulated annealing algorithm on simulated data. A** Absolute error in the estimation of the log-likelihood. **B** Fraction of the correctly identified elements of the cover relation over all the expected ones (recall). **C** Fraction of the correctly identified elements of the cover relation over all the recovered ones (precision). Box plots show results for 20 different transitively reduced networks and simulated data sets with 16 mutations and an error rate of 5%. We use the forward sampling scheme with $L = 1000$ samples drawn from the proposal distribution. We fix the ideal acceptance rate to $1/p = 0.0625$, and run 25,000 iterations of the simulated annealing algorithm. The initial temperature is set to $\Theta_0 = 50$ for all runs, and for adaptive simulated annealing, three adaptation rates are evaluated ($a_r = 0.1, 0.3, 0.5$). SA: simulated annealing, ASA: adaptive simulated annealing, +: with additional new moves.



Finally, we investigate the influence of the annealing schedule hyper-parameters, such as the initial temperature and the adaptation rate (Fig S8). In general, the performance of the ASA algorithm is not critically influenced by the choice of the annealing hyper-parameters. Moreover, the ASA algorithm is neither better nor worse than the SA+ algorithm, at least for the test cases with $p = 16$ mutations. Nevertheless, the ASA algorithm has the conceptual advantage of adjusting the temperature adaptively according to the system behaviour rather than using a fixed schedule and thus may be more reliable across unknown likelihood landscapes.

## Comparison of drug resistance-associated mutational pathways in different HIV-1 subtypes under lopinavir treatment

We analyze viral genotypes from a cohort of 1065 South African patients living with HIV-1 subtype C retrieved from the HIVDB. These patients were treated with lopinavir boosted with low-dosed ritonavir. We select a subset of 21 major protease inhibitor (PI) resistance mutations associated with lopinavir resistance and 15 non-polymorphic accessory mutations according to the HIVDB [41] (Fig S10). We follow the convention of reporting mutations relative to the amino acid sequence of the HIV-1 subtype B reference strain HXB2. Among the selected loci, HIV-1 subtype C sequences typically differ at residue 89 from the subtype B reference strain: instead of a leucine (L) a methionine (M) is frequently observed [44]. This naturally occurring polymorphism is found in 77.09% of the patient in our cohort.

We find 15 out of the 21 major PI mutations and 13 out of 15 non-polymorphic accessory mutations in the South African cohort. The remaining unobserved mutations include PI mutations G48M, I54A/L/M/T, and V82T, and non-polymorphic mutations L24F and A71I. We exclude polymorphisms commonly found in wild type subtype C viruses as they likely correspond to baseline mutations, but some of these are highly prevalent in the cohort—for instance, I93L (97.18%), M36I (86.95%), and K20R (34.18%). Among the 1065 genotypes, 911 are wild type for the selected loci and the maximum number of co-occurring mutations is eight.

In addition, we analyze two genotype-treatment data sets from the HIVDB corresponding to HIV-1 subtype B and C genotypes. For the latter, we exclude genotypes from South Africa that constitute the data set described above. All patients in these data sets were treated with lopinavir or lopinavir and ritonavir but not with another protease inhibitor. The data sets include 298 and 775 sequences of subtype B and C genotypes, respectively. Additionally, we consider 172 HIV-1 subtype B sequences of the SHCS derived from patients treated with lopinavir as the first PI. We jointly analyse all 470 subtype B genotypes to mitigate the small sample size.

We use H-CBN2 for analyzing and comparing the accumulation of resistance mutations in HIV-1 subtype B and subtype C under the selective pressure of lopinavir. We employ the forward sampling scheme to learn the partial order among mutations. The robustness of the network estimation is investigated by using 100 bootstrap samples and the consensus networks are shown in Figs 5A and 6 ($p = 20$ and $p = 18$, respectively). In the South African cohort (subtype C sequences), we identify a mutation at residue 82 in the protease as an early event. The initial substitution is likely to be V82A, as it is predominantly observed in the data set (Fig S10). After this initial event, we find strong support for mutations at residues 10, 33, 46, 54 and 76 (Fig 5A). For subtype B, we find strong support for a mutation at residue 46 as an initial event (Fig 6). The inferred posets can explain previously observed mutation patterns, such as M46I+I54V alone or in combinations with L76V or V82A in subtype B [33], as well as M46I+I54V+V82A and L10F+M46I+I54V+L76V+V82A in subtype C [34].

At first glance, the subtype-specific H-CBN2 posets appear to be different. However, they also share many features. We find that they have 5 cover relations in common, namely, I54V $\prec$ L24I, I54V $\prec$ F53L, I54V $\prec$ G73S, I54V $\prec$ T74P, and I54V $\prec$ L89V. In addition, in both posets mutation at residue 82 precedes G73S and T74P, and mutation at residue 46 precedes K43T, F53L, T74P, and L89V either in a direct manner or through an intermediary event.

To assess whether the two H-CBN2 posets are significantly different beyond reconstruction uncertainty, we have developed a customized statistical test based on the Jaccard distance between the posets. The distance between the maximum likelihood posets (Fig S11 and Fig S12) is 0.802. To assess the significance of this result, we compare it to the empirical distribution of pairwise distances computed between reconstructed networks after randomly permuting the group labels (Fig 7). At a significance level of 5%, we reject the null hypothesis that the data sets stem from the same underlying poset (p-value $< 0.02$, Fig 7B), for $p = 18$ mutations. Similarly, we reject the null hypothesis while comparing subtype-specific CBN models for HIV-1 subtype B and C with $p = 11$ mutations (p-value $= 0.04$, Fig 7A). The smaller data sets are obtained by discarding mutations with marginal counts less or equal 5 in either of the two data sets.

As a negative control, we also compare the two H-CBN2 models for subtype C inferred from the South African cohort versus the remaining subtype C genotypes from the HIVDB (Fig 5). The consensus posets share 16 cover relations, namely, L10FR $\prec$ K43T, L10FR $\prec$ F53L, K20MT $\prec$ F53L, L33F $\prec$ T74P, M46ILV $\prec$ K43T, M46ILV $\prec$ F53L, M46ILV $\prec$ T74P, I54V/L $\prec$ F53L, I54V/L $\prec$ T74P, I54V/L $\prec$ L89V, V82AMF/CS $\prec$ L10FR, V82AMF/CS $\prec$ M46ILV, V82AMF/CS $\prec$ I54V/L, V82AMF/CS $\prec$ Q58E, V82AMF/CS $\prec$ L76V, and V82AMF/CS $\prec$ I84V. Moreover, in both posets mutation at residue 10 precedes T74P and mutation at residue 82 precedes L24I, L33F, K43T, F53L, G73S, T74P, L89V, and L90M. We also employed the aforementioned statistical test to compare posets with different number of mutations,

**Fig 5. Consensus posets for lopinavir resistance for two different HIV-1 subtype C data sets.** Shown are the consensus poset for **A** the South African cohort and **B** for the remaining HIV-1 subtype C sequences retrieved from the HIVDB. Nodes in the network correspond to amino acid changes in the HIV-1 protease, where mutations at the same locus are grouped together in one event. Only edges with a bootstrap support greater than 0.7 are shown and the edge thickness indicates the bootstrap support. Nodes with white background show residues with at least one major PI mutation.

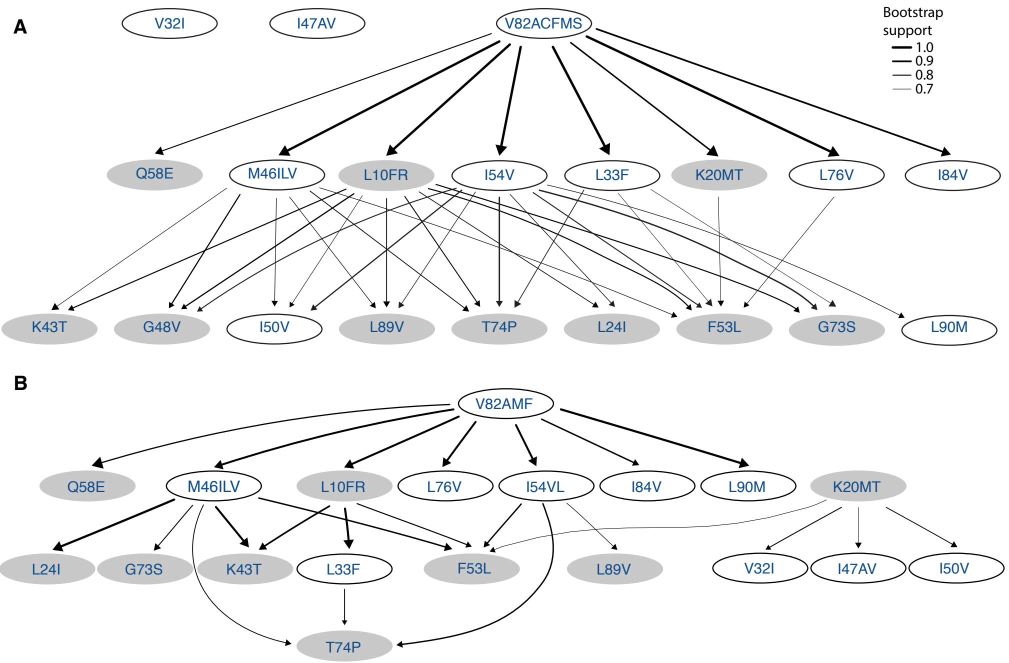
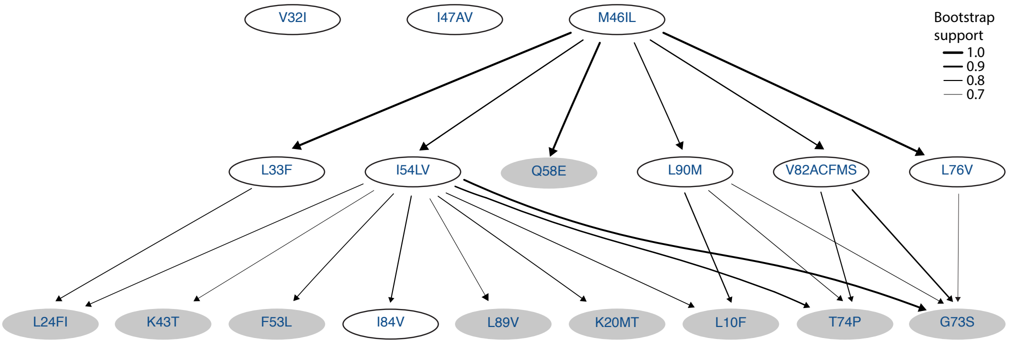
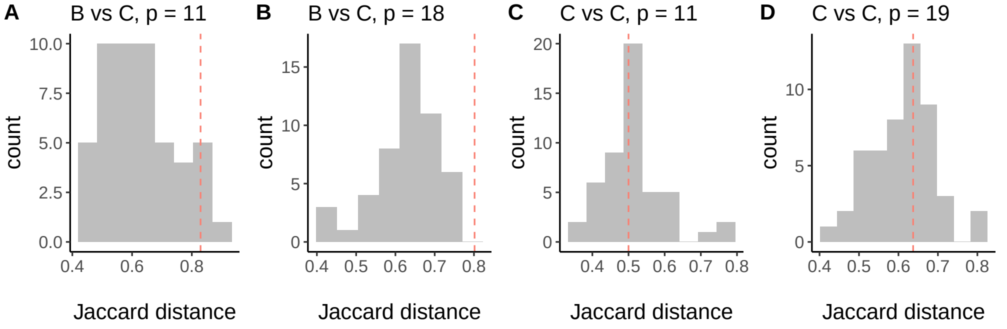


**Fig 6. Consensus poset for the accumulation of mutations in HIV-1 subtype B under lopinavir treatment.** The underlying data set contains 470 genotypes retrieved from the HIVDB and SHCS. Nodes in the network correspond to amino acid changes in the HIV-1 protease, and mutations at the same locus are grouped together. Edge labels indicate the bootstrap support, and we show only edges with a bootstrap support greater or equal to 0.7.



namely $p = 19$ and $p = 11$ mutations. The larger poset size corresponds to all the mutated loci common in both data sets and the threshold on the marginal mutation counts for constructing the smaller data sets is set to 8 mutations. The Jaccard distance between these two H-CBN2 models is 0.637 and 0.5 for posets with $p = 19$ and $p = 11$ mutations, respectively. There is no evidence supporting that the posets learned from different data sets but the same subtype C are different (p-values 0.42 and 0.66, respectively; Fig 7C-D).

**Fig 7. Empirical null distribution of pairwise Jaccard distances estimated by permuting group labels.** Displayed are the histograms of Jaccard distances for the comparison of subtypes B and C for H-CBN2 posets with **A** 11 mutations and **B** 18 mutations, as well as the histograms of Jaccard distances for the comparison of two data sets for subtype C for H-CBN2 posets with **C** 11 mutations and **D** 19 mutations. Vertical dotted lines indicate the distance between the CBNs obtained from the observed data.



# Discussion

We have presented the H-CBN2 model and inference methods which are based on Monte Carlo sampling and enable us to consider a larger number of mutations. In simulation studies, we demonstrated that this method can be used to accurately estimate model parameters for up to about 32 mutations. For larger numbers of mutations, the sample sizes used in this work are insufficient to obtain accurate parameter estimates. To learn the graph, we proposed an extension of the simulated annealing algorithm, including additional move types that allow exploring the space of posets more efficiently. We validated the structure learning algorithm for 16 mutations which aligns with the numbers of mutations relevant for our application to HIV-1. Structure learning is, however, a hard problem and further improving the efficiency of this step might be worthwhile addressing in future research.

Even though there are descriptive analyses of subtype-specific PI mutation profiles [23, 33, 34, 44], to our knowledge, this study is the first comparative analysis of pathways of accumulating mutations over time in different HIV-1 subtypes. In addition to a more systematic approach to investigating mutation patterns, the number of observations in our study is greater than in any of the previous studies, which ranged from 88 to 165 patients. We applied the H-CBN2 approach to learn the partial temporal ordering of resistance mutations in HIV-1 subtypes B and C under the selective pressure of lopinavir. Our results indicate that despite some similarities, for the considered numbers of mutations, the resistance pathways differ significantly between the two subtypes. Moreover, we compared H-CBN2 posets for subtype C inferred from two independent data sets as a validation of the distance-based test and the outcome aligns with the expectation that there exists a single underlying poset explaining both data sets better than two distinct posets.

In our analysis, we included major PI mutations associated with lopinavir resistance and non-polymorphic accessory mutations. Although some polymorphisms, in combination with PI resistance mutations, are associated with an increase in viral fitness [45], these are also highly prevalent in treatment-naïve patients, especially in non-B subtypes [46–48]. Therefore, despite observing polymorphisms with relatively high prevalence, we did not include these mutations in our study. We also found more than one PI-associated mutation in only about 14% and 16% of the patients in the South African cohort (subtype C) and the subtype B data set, respectively. The absence of resistance mutations in the protease gene has been repeatedly observed at virological failure, even in the absence of reverse transcriptase inhibitors [33, 49–52]. In addition to poor adherence to treatment [53, 54], there may be other reasons for observing a low percentage of patients harboring PI resistance mutations, and some of them are listed below. First, the genetic barrier to lopinavir resistance appears to be high. Barber *et al.* [33] have suggested that PI resistance mutations are more likely to accumulate under prolonged virological failure. Second, there is increasing evidence that mutations in the *gag* gene play a role in decreasing susceptibility to protease inhibitors by, e.g., inhibiting the proteolytic cleavages necessary for protein maturation [19, 55, 56]. Virions with immature particles may not adequately complete cell entry or reverse transcription [56]. Third, resistance mutations may exist in the intra-host virus population at frequency below the detection threshold. Mutations are typically detected by Sanger sequencing-based methods, while next-generation sequencing methods could improve upon the sensitivity of detecting mutations [57].

A limitation of the methodology is the aggregation of different amino acid substitutions at the same locus as single events. This is required because when observing a specific substitution, we do not know which other mutations at that locus led to the current state. One potential way to overcome this limitation is to incorporate additional data sources, such as time-series data per patient, single-genome amplification data, or even resort to next-generation sequencing data. However, these data are not generally available in public databases, nor are they part of routine diagnostic tests. If data were available, such hidden states could be accounted for in an extended model. Nonetheless, with the number of genotypes available in the current application, we may not be able to include more mutations without decreasing the accuracy of the parameter estimates of the H-CBN model.

The comparison of the cross-sectional data sets is challenging due to the existence of several confounders. First, the data are gathered from various sources, which entails potential differences in HIV surveillance and clinical monitoring protocols. Moreover, observations come from distinct geographical locations, which implies, e.g., differences in socio-demographic aspects and health-care standards. Lastly, therapeutic strategies tend to differ between developed and developing countries, and there is a limited number of observations of various subtypes undergoing the same therapy. In the present study, the number of observations in the subtype B data set is approximately half of the observations available for subtype C. Such an imbalance poses additional challenges for the CBN comparisons. The spread of the empirical distribution of Jaccard distances might be wider for imbalanced data sets, which could result in an apparent increase in false negatives. But rather than a shortcoming of the distance-based test, small sample size generally lead to reduced accuracy of the parameter estimates, including the network structure.

In summary, the inferred CBN models provide insights into the evolution of drug resistance in HIV-1 subtype C infections and enable comparisons with other subtypes, as demonstrated for subtype B. Moreover, the methods proposed in this work can be generally applied to investigate subtype-associated differences pertaining to HIV-1 drug resistance.

# Supporting information

**Fig S1 Assessment of the forward sampling.** Probability of the observed genotype estimated by using the forward sampling scheme $\widetilde{\Pr}(Y = y)$ (y-axis, Eq. 17) vs. the exact solution $\Pr(Y = y)$ (x-axis). The data set consists of $N = 800$ genotypes with $p = 16$ mutations. Results are obtained by drawing **A** $L = 10$, **B** $L = 100$, and **C** $L = 1000$ samples from the proposal distribution.

**Fig S2 Assessment of the forward-pool sampling.** Probability of the observed genotype estimated by using the forward-pool sampling scheme $\widetilde{\Pr}(Y = y)$ (y-axis, Eq. 17) vs. the exact solution $\Pr(Y = y)$ (x-axis). The data set consists of $N = 800$ genotypes with $p = 16$ mutations. First, we evaluate the impact of the size of the initial pool on the accuracy of the approximations. We show results for pools consisting of **A**, **D** $K = 200$, **B**, **E** $K = 800$, and **C**, **F** $K = 1600$ samples, while the number of samples drawn from the proposal distribution is set to either **A-C** $L = 10$ or **D-F** $L = 100$. Next, we evaluate the impact of the number of samples drawn from the proposal distribution (**G** $L = 10$; **H** $L = 100$; **I** $L = 1000$), while the size of the initial pool is kept constant at $K = 2000$ samples. We observe that the accuracy of the computation improves primarily as the size of the initial pool increases. By default, the size of the initial pool of waiting times is set to $K = p \times L$.

**Fig S3 Assessment of the Hamming-*k*-neighborhood sampling.** Probability of the observed genotype estimated by using the Hamming $k$-neighbourhood sampling scheme $\widetilde{\Pr}(Y = y)$ (y-axis, Eq. 17) vs. the exact solution $\Pr(Y = y)$ (x-axis). The data set consists of $N = 800$ genotypes with $p = 16$ mutations. Results are shown for **A-C** a neighborhood including the leading and the first-order terms $(k = 1)$, **D-F** a neighborhood including the leading, the first-order, and the second-order terms $(k = 2)$, and **G-I** a neighborhood including the leading, the first-order, the second-order, and the third-order terms $(k = 3)$. In this case, the value of $L$ indicates the number of waiting time vectors sampled per genotype in the neighborhood.

**Fig S4 Assessment of the Bernoulli sampling.** Probability of the observed genotype estimated by using the Bernoulli sampling scheme $\widetilde{\Pr}(Y = y)$ (y-axis, Eq. 17) vs. the exact solution $\Pr(Y = y)$ (x-axis). The data set consists of $N = 800$ genotypes with $p = 16$ mutations. Results are obtained by drawing **A** $L = 10$, **B** $L = 100$, and **C** $L = 1000$ samples from the proposal distribution. In the lower panel, we show the number of samples compatible with the poset $L_{\text{compatible}}$ per genotype for **D** $L = 10$, **E** $L = 100$, and **F** $L = 1000$ samples drawn from the proposal distribution.

**Fig S5 Assessment of the backward-AR sampling.** Probability of the observed genotype estimated by using the backward-AR sampling scheme $\widetilde{\Pr}(Y = y)$ (y-axis, Eq. 17) vs. the exact solution $\Pr(Y = y)$ (x-axis). The data set consists of $N = 800$ genotypes with $p = 16$ mutations. Results are obtained by drawing **A** $L = 10$, **B** $L = 100$, and **C** $L = 1000$ samples from the proposal distribution.

**Fig S6 Assessment of the parameter estimation for various numbers of mutations and various numbers of samples drawn from the proposal distribution.** Box plots of the absolute error in estimating the error rate $\hat{\epsilon}$ for the true poset $P$ by using **A** the forward sampling, **D** the Hamming $k$-neighborhood sampling, and **G** the Bernoulli or the backward-AR sampling. Box plots of the relative median absolute error (RMAE) for the estimated rate parameters $\hat{\lambda}$ by using **B** the forward sampling, **E** the Hamming $k$-neighborhood sampling, and **H** the Bernoulli or the

backward-AR sampling. The relative (median) absolute error is given by $\frac{\text{median}(|\hat{\lambda}-\lambda|)}{\text{median}(\lambda)}$. Average run times over simulated data sets for **C** the forward sampling, **F** the Hamming $k$-neighborhood sampling, and **I** the Bernoulli or the backward-AR sampling. Results correspond to 100 simulated data sets for each of the number of mutations and 100 iterations of the MCEM algorithm. The number of samples drawn from the proposal distribution is $L = 10, 100, 1000$, as shown in the corresponding legend. The sample size is $N = \min(50\,p, 1000)$ and the true error rate is $\epsilon = 0.05$.

**Fig S7 Average run time of the MCEM step (y-axis, logarithmic scale) using various sampling schemes for different poset sizes (x-axis, logarithmic scale).** We also show the run times of the H-CBN method for posets with up to 14 mutations. The benchmark is conducted on 100 different data sets per poset size, and the number of EM iterations is set to 100. The blue dotted line corresponds to linear scaling, whereas the red line corresponds to quadratic scaling. We conduct the benchmark on two 12-core Intel Xeon E5-2680 v3 processors (2.5 GHz)

**Fig S8 Evaluation of the simulated annealing algorithm for various initial temperatures ($\Theta_0 = 10$, 30, 50) and adaptation rates ($a_r = 0.1$, 0.3, 0.5).** We show box plots corresponding to 20 different transitively reduced DAGs with 16 mutations and for an error rate of 5%. Gray box plots correspond to results of the MCEM algorithm for the true poset. We use the forward sampling scheme with $L = 1000$ samples. For learning the poset, we fix the ideal acceptance rate to $1/p = 0.0625$ and run 25000 simulated annealing iterations. The relative (median) absolute error (RMAE) is given by $\frac{\text{median}(|\hat{\lambda}-\lambda|)}{\text{median}(\lambda)}$. P: true poset, SA: simulated annealing, +: with additional new moves.

**Fig S9 Evaluation of the adaptive simulated annealing algorithm on simulated data. A** Absolute error in estimating the error rate parameter $\hat{\epsilon}$. **B** Relative median absolute error (RMAE) of the estimated rate parameters $\hat{\lambda}$. The relative (median) absolute error is given by $\frac{\text{median}(|\hat{\lambda}-\lambda|)}{\text{median}(\lambda)}$. **C** Jaccard distance computed on the cover relation sets for the true and estimated poset. We show box plots corresponding to 20 different transitively reduced DAGs for simulated data sets with 16 mutations and an error rate of 5%. We use the forward sampling scheme with $L = 1000$ samples drawn from the proposal distribution. We fix the ideal acceptance rate to $1/p = 0.0625$ and run 25000 iterations of the simulated annealing algorithm. The initial temperature is set to $\Theta_0 = 50$ for all runs and for the adaptive simulated annealing three adaptation rates are evaluated ($a_r = 0.1, 0.3, 0.5$). SA: simulated annealing, ASA: adaptive simulated annealing, +: with additional new moves.

**Fig S10 Marginal mutation frequencies observed in HIV-1 subtype B and subtype C populations under lopinavir treatment. A** Data collected from 1065 patients from South Africa (HIV-1 subtype C). **B** HIV-1 subtype C genotypes retrieved from the HIVDB, excluding genotypes in data set A. **C** HIV-1 subtype B genotypes retrieved from the HIVDB. **D** Data obtained from the SHCS corresponding to subtype B genotypes. Major protease inhibitor resistance mutations are shown in black.

**Fig S11 Maximum likelihood poset for lopinavir resistance in the South African cohort.** Nodes in the network correspond to amino acid changes in the protease, with mutations at the same locus grouped together. Mutations G48V and I50V are excluded for the comparison of H-CBN2 models, as they are not observed in the subtype B data set.

**Fig S12 Maximum likelihood poset for lopinavir resistance in HIV-1 subtype B.** Nodes in the network correspond to amino acid changes in the protease, with mutations at the same locus grouped together. Data sources: the HIVDB and the SHCS.

**Table S1 Relative error in approximating the log-likelihood via importance sampling.** The relative error is computed by dividing the absolute error by the absolute value of the true log-likelihood.

**File S1 Patient identifiers of the South African cohort as retrieved from the HIVDB.**

**File S2 Patient identifiers of HIV-1 subtype B genotype sequences retrieved from the HIVDB.**

**File S3 Patient identifiers of HIV-1 subtype C genotype sequences retrieved from the HIVDB and excluding genotypes from South African cohort.**

## Acknowledgments

The authors thank Lisa Lamberti for critical reading, Marek Pikulski for valuable discussions, and Huyen Nguyen for providing data sets from the SHCS. The authors thank the patients who participate in the Swiss HIV Cohort Study (SHCS); the physicians and study nurses, for excellent patient care; the resistance laboratories for high-quality genotyping drug resistance testing; SmartGene (Zug, Switzerland) for technical support; Alexandra Scherrer, Susanne Wild, Anna Traytel from the SHCS data center for data management; and Danièle Perraudin and Mirjam Minichiello for administration. The members of the SHCS include the following: Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C (Chairman of the Mother and Child Substudy), Scherrer AU (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

## Financial support

The SHCS and it's drug resistance database was supported by the Swiss National Science Foundation (grant number 179 571 to H. F. G.) and was financed within the framework of the SHCS, supported by the Swiss National Science Foundation (grant number 148522), by the SHCS Research Foundation and by the Yvonne Jacob Foundation (to H. F. G.)

## Competing Interests

H. F. G. has received unrestricted research grants from Gilead Sciences and Roche, fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences, Viiv and Merck; and grants from SystemsX, and the National Institutes of Health.

## Author Contributions

**Conceptualization:** Susana Posada-Céspedes, Gert Van Zyl, Hesam Montazeri and Niko Beerenwinkel
**Data curation:** Soo-Yon Rhee
**Formal Analysis:** Susana Posada-Céspedes
**Funding Acquisition:** Roger Kouyos, Huldrych F. Günthard and Niko Beerenwinkel
**Methodology:** Susana Posada-Céspedes, Hesam Montazeri, Jack Kuipers and Niko Beerenwinkel
**Resources and data acquisition:** Roger Kouyos and Huldrych F. Günthard
**Software:** Susana Posada-Céspedes
**Supervision:** Niko Beerenwinkel
**Writing – Original Draft Preparation:** Susana Posada-Céspedes and Niko Beerenwinkel

## References

1. Saag MS, Benson CA, Gandhi RT, Hoy JF, Landovitz RJ, Mugavero MJ, et al. Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2018 recommendations of the International Antiviral Society-USA Panel. JAMA. 2018;320(4):379–396. doi:10.1001/jama.2018.8431.

2. Günthard HF, Calvez V, Paredes R, Pillay D, Shafer RW, Wensing AM, et al. Human Immunodeficiency Virus drug resistance: 2018 recommendations of the International Antiviral Society-USA Panel. Clin Infect Dis. 2019;68(2):177–187. doi:10.1093/cid/ciy463.

3. Beerenwinkel N, Däumer M, Sing T, Rahnenführer J, Lengauer T, Selbig J, et al. Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. J Infect Dis. 2005;191(11):1953. doi:10.1086/430005.

4. Buendia P, Cadwallader B, DeGruttola V. A phylogenetic and Markov model approach for the reconstruction of mutational pathways of drug resistance. Bioinformatics. 2009;25(19). doi:10.1093/bioinformatics/btp466.

5. Mbisa JL, Gupta RK, Kabamba D, Mulenga V, Kalumbi M, Chintu C, et al. The evolution of HIV-1 reverse transcriptase in route to acquisition of Q151M multi-drug resistance is complex and involves mutations in multiple domains. Retrovirology. 2011;8(31). doi:10.1186/1742-4690-8-31.

6. Lawyer G, Altmann A, Thielen A, Zazzi M, Sönnerborg A, Lengauer T. HIV-1 mutational pathways under multidrug therapy. AIDS Res Ther. 2011;8(26). doi:10.1186/1742-6405-8-26.

7. Larder BA, Kemp SD. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). Science. 1989;246(4934):1155–1158.

8. Boucher CAB, O'Sullivan E, Mulder JW, Ramautarsing C, Kellam P, Darby G, et al. Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency Virus-positive subjects. J Infect Dis. 1992;165(1):105. doi:10.1093/infdis/165.1.105.

9. Condra J, Schleif W, Blahy O, Gabryelski LJ, Graham DJ, Quintero JC, et al. *In vivo* emergence of HIV-1 variants resistant to multiple protease inhibitors. Nature. 1995;374:569–571. doi:10.1038/374569a0.

10. Molla A, Korneyeva M, Gao Q, Vasavanonda S, Schipper PJ, Mo HM, et al. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nat Med. 1996;2:760–766. doi:10.1038/nm0796-760.

11. Beerenwinkel N, Drton M. A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. Biostatistics. 2006;8(1):53–71. doi:10.1093/biostatistics/kxj033.

12. Yang WL, Kouyos RD, Scherrer AU, Böni J, Shah C, Yerly S, et al. Assessing efficacy of different nucleos(t)ide backbones in NNRTI-containing regimens in the Swiss HIV Cohort Study. J Antimicrob Chemoth. 2015;70(12):3323–3331. doi:10.1093/jac/dkv257.

13. Foulkes AS, DeGruttola V. Characterizing the progression of viral mutations over time. J Am Stat Assoc. 2003;98(464):859–867. doi:10.1198/016214503000000792.

14. Beerenwinkel N, Rahnenführer J, Däumer D, Hoffmann D, Kaiser J, Selbig J, et al. Learning multiple evolutionary pathways from cross-sectional data. J Comput Biol. 2005;12(6):584–598.

15. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, et al. Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance. Bioinformatics. 2006;22(24):2975–2979. doi:10.1093/bioinformatics/btl508.

16. Deforche K, Camacho RJ, Grossman Z, Soares MA, Van Laethem K, Katzenstein DA, et al. Bayesian network analyses of resistance pathways against efavirenz and nevirapine. AIDS. 2008;22(16):2107–2115. doi:10.1097/QAD.0b013e32830fe940.

17. Zhang J, Hou T, Wang W, Liu JS. Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. Proc Natl Acad Sci USA. 2010;107(4):1321–1326. doi:10.1073/pnas.0907304107.

18. Hernandez-Leal P, Rios-Flores A, Ávila Rios S, Reyes-Terán G, Gonzalez JA, Fiedler-Cameras L, et al. Discovering human immunodeficiency virus mutational pathways using temporal Bayesian networks. Artif Intell Med. 2013;57(3):185 – 195. doi:https://doi.org/10.1016/j.artmed.2013.01.005.

19. Marie V, Gordon M. Gag-protease coevolution shapes the outcome of lopinavir-inclusive treatment regimens in chronically infected HIV-1 subtype C patients. Bioinformatics. 2019;35(18):3219–3223. doi:10.1093/bioinformatics/btz076.

20. Beerenwinkel N, Eriksson N, Sturmfels B. Conjuctive Bayesian networks. Bernoulli. 2007;13(4):893–909. doi:10.3150/07-BEJ6133.

21. Beerenwinkel N, Sullivant S. Markov models for accumulating mutations. Biometrika. 2009;96(3):645. doi:10.1093/biomet/asp023.

22. Ramazzotti D, Graudenzi A, Caravagna G, Antoniotti M. Modeling cumulative biological phenomena with suppes-Bayes causal networks. Evolutionary Bioinformatics. 2018;14:1176934318785167. doi:10.1177/1176934318785167.

23. Ariyoshi K, Matsuda M, Miura H, Tateishi S, Yamada K, Sugiura W. Patterns of point mutations associated with antiretroviral drug treatment failure in CRF01_AE (subtype E) infection differ from subtype B infection. J Acquir Immune Defic Syndr. 2003;33(3):335–342.

24. Wainberg MA. HIV-1 subtype distribution and the problem of drug resistance. AIDS. 2004;18:S63–68.

25. Kantor R, Katzenstein DA, Efron B, Carvalho AP, Wynhoven B, Cane P, et al. Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. PLoS Med. 2005;2(4):e112. doi:10.1371/journal.pmed.0020112.

26. Kantor R. Impact of HIV-1 pol diversity on drug resistance and its clinical implications. Curr Opin Infect Dis. 2006;19:594–606. doi:10.1097/QCO.0b013e3280109122.

27. Kosakovsky Pond SL, Smith DM. Are all subtypes created equal? The effectiveness of antiretroviral therapy against non-subtype B HIV-1. Clin Infect Dis. 2009;48(9):1306–1309. doi:10.1086/598503.

28. Martinez-Cajas JL, Pai NP, Klein MB, Wainberg MA. Differences in resistance mutations among HIV-1 non-subtype B infections: a systematic review of evidence (1996-2008). J Int AIDS Soc. 2009;12(11). doi:10.1186/1758-2652-12-11.

29. Han YS, Mesplède T, Wainberg MA. Differences among HIV-1 subtypes in drug resistance against integrase inhibitors. Infect Genet Evol. 2016;46:286 – 291. doi:10.1016/j.meegid.2016.06.047.

30. Hemelaar J. The origin and diversity of the HIV-1 pandemic. Trends Mol Med. 2012;18:182–192.

31. Gerstung M, Baudis M, Moch H, Beerenwinkel N. Quantifying cancer progression with conjunctive Bayesian networks. Bioinformatics. 2009;25(21):2809–2815. doi:10.1093/bioinformatics/btp505.

32. Montazeri M, Kuipers J, Kouyos R, Böni J, Yerly S, Klimkait T, et al. Large-scale inference of conjunctive Bayesian networks. Bioinformatics. 2016;32:i727–i735. doi:10.1093/bioinformatics/btw459.

33. Barber TJ, Harrison L, Asboe D, Williams I, Kirk S, Gilson R, et al. Frequency and patterns of protease gene resistance mutations in HIV-infected patients treated with lopinavir/ritonavir as their first protease inhibitor. J Antimicrob Chemother. 2012;67(4):995–1000. doi:10.1093/jac/dkr569.

34. Grossman Z, Schapiro JM, Levy I, Elbirt D, Chowers M, Riesenberg, et al. Comparable long-term efficacy of Lopinavir/Ritonavir and similar drug-resistance profiles in different HIV-1 subtypes. PLoS One. 2014;1:e86239. doi:10.1371/journal.pone.0086239.

35. Ingber L. Simulated annealing: practice versus theory. Mathl Comput Modelling. 1993;18(11):29–57.

36. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671–680. doi:10.1126/science.220.4598.671.

37. Ingber L. Adaptive simulated annealing (ASA): lessons learned. Control and Cybernetics. 1996;25(1):33–54.

38. Chen S, Luk BL. Adaptive simulated annealing for optimization in signal processing applications. Signal Processing. 1999;79(1):117 – 128. doi:https://doi.org/10.1016/S0165-1684(99)00084-5.

39. Srivatsa S, Kuipers J, Schmich F, Eicher S, Emmenlauer M, Dehio C, et al. Improved pathway reconstruction from RNA interference screens by exploiting off-target effects. Bioinformatics. 2018;34(13):i519–i527. doi:10.1093/bioinformatics/bty240.

40. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, W SR. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. 2003;31(1):298–303. doi:10.1093/nar/gkg100.

41. Shafer RW. Rationale and uses of a public HIV drug-resistance database. J Infect Dis. 2006;194:Suppl 1 S51–S58. doi:10.1086/505356.

42. von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Battegay M, et al. Emergence of HIV-1 drug resistance in previously untreated patients initiating combination antiretroviral treatment: a comparison of different regimen types. Arch Intern Med. 2007;167(16):1782–1790. doi:10.1001/archinte.167.16.1782.

43. Swiss HIV Cohort Study, Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Günthard HF, et al. Cohort profile: the Swiss HIV Cohort study. Int J Epidemiol. 2010;39(5):1179–1189. doi:10.1093/ije/dyp321.

44. Champenois K, Deuffic-Burban S, Cotte L, André P, Choisy P, Ajana F, et al. Natural polymorphisms in HIV-1 protease: impact on effectiveness of a first-line lopinavir-containing antiretroviral therapy regimen. J Med Virol. 2008;80:1871–1879.

45. Nijhuis M, Schuurman R, de Jong D, Erickson J, Gustchina E, Albert J, et al. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. AIDS. 1999;13(17):2349–2359.

46. Pieniazek D, Rayfield M, Hu DJ, Nkengasong J, Wiktor SZ, Downing R, et al. Protease sequences from HIV-1 group M subtypes A–H reveal distinct amino acid mutation patterns associated with protease resistance in protease inhibitor-naive individuals worldwide. AIDS. 2000;14(11):1489–1495.

47. Vergne L, Peeters M, Mpoudi-Ngole E, Bourgeois A, Liegeois F, Toure-Kane C, et al. Genetic diversity of protease and reverse transcriptase sequences in non-subtype-B human immunodeficiency virus type 1 strains: evidence of many minor drug resistance mutations in treatment-naive patients. J Clin Microbiol. 2000;38(11):3919–3925.

48. Shafer RW, Schapiro JM. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. AIDS rev. 2008;10(2):67–84.

49. Parkin NT, Chappey C, Petropoulos CJ. Improving lopinavir genotype algorithm through phenotype correlations: novel mutation patterns and amprenavir cross-resistance. AIDS. 2003;17(7):955–961. doi:10.1097/00002030-200305020-00003.

50. Kempf DJ, King MS, Bernstein B, Cernohous P, Bauer E, Moseley J, et al. Incidence of resistance in a double-blind study comparing lopinavir/ritonavir plus stavudine and lamivudine to nelfinavir plus stavudine and lamivudine. J Infect Dis. 2004;189(1):51–60. doi:10.1086/380509.

51. Sahali S, Chaix ML, Delfraissy JF, Ghosn J. Ritonavir-boosted protease inhibitor monotherapy for the treatment of HIV-1 infection. AIDS Rev. 2008;10(1):4–14.

52. Scherrer AU, Böni J, Yerly S, Klimkait T, Aubert V, Furrer H, et al. Long-lasting protection of activity of nucleoside reverse transcriptase inhibitors and protease inhibitors (PIs) by boosted PI containing regimens. PLoS One. 2012;7(11):e50307. doi:10.1371/journal.pone.0050307.

53. Rosenbloom DI, Hill AL, Rabi SA, Siliciano RF, Nowak MA. Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. Nat Med. 2012;18(9):1378–1385. doi:10.1038/nm.2892.

54. von Wyl V, Klimkait T, Yerly S, Nicca D, Furrer H, Cavassini M, et al. Adherence as a predictor of the development of class-specific resistance mutations: the Swiss HIV Cohort Study. PLoS One. 2013;8(10):e77691. doi:10.1371/journal.pone.0077691.

55. Knops E, Kemper I, Schülter E, Pfister H, Kaiser R, Verheyen J. The evolution of protease mutation 76V is associated with protease mutation 46I and gag mutation 431V. AIDS. 2010;24(5):779–781. doi:10.1097/QAD.0b013e328336784d.

56. Rabi SA, Laird GM, Durand CM, Laskey S, Shan L, Bailey JR, et al. Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. J Clin Invest. 2013;123(9):3848–3860. doi:doi.org/10.1172/JCI67399.

57. Parikh UM, McCormick K, van Zyl G, Mellors JW. Future technologies for monitoring HIV drug resistance and cure. Curr Opin HIV AIDS. 2017;12(2):182–189. doi:10.1097/COH.0000000000000344.