Modeling SARS-CoV-2 virus mutations as a Markov Chain embedded Poisson Process

Maverick Lim[1], Ercan Kuruoglu[1], Victor Chan[1]

[1]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

Abstract

Accurate predictions of virus mutations allow us to be better equipped when dealing with the COVID-19 pandemic that has been ravaging the world for over a year. There is an ongoing race to keep producing efficacious vaccines in order to keep up with the mutation of new SARS-CoV-2 virus strains that threaten to render existing vaccines obsolete. This study seeks to provide an accurate prediction of SARS-CoV-2 virus mutations by modeling it as a Markov Chain embedded Poisson random process. It analyzes phylogenetic data from the GISAID global database, comprised of virus RNA sequences submitted from labs around the world.

Introduction & Motivation

There has been an alarming number of reports of "*breakthrough cases*" recently, whereby individuals are still becoming infected by the virus despite having been already vaccinated. Thus, it is imperative to predict when and where the next mutation would occur ahead of the actual realization. This keeps vaccine manufacturers one step ahead of the virus, enabling them to pre-emptivly adapt the vaccine production process, potentially saving countless lives in an outbreak of a dangerous strain.

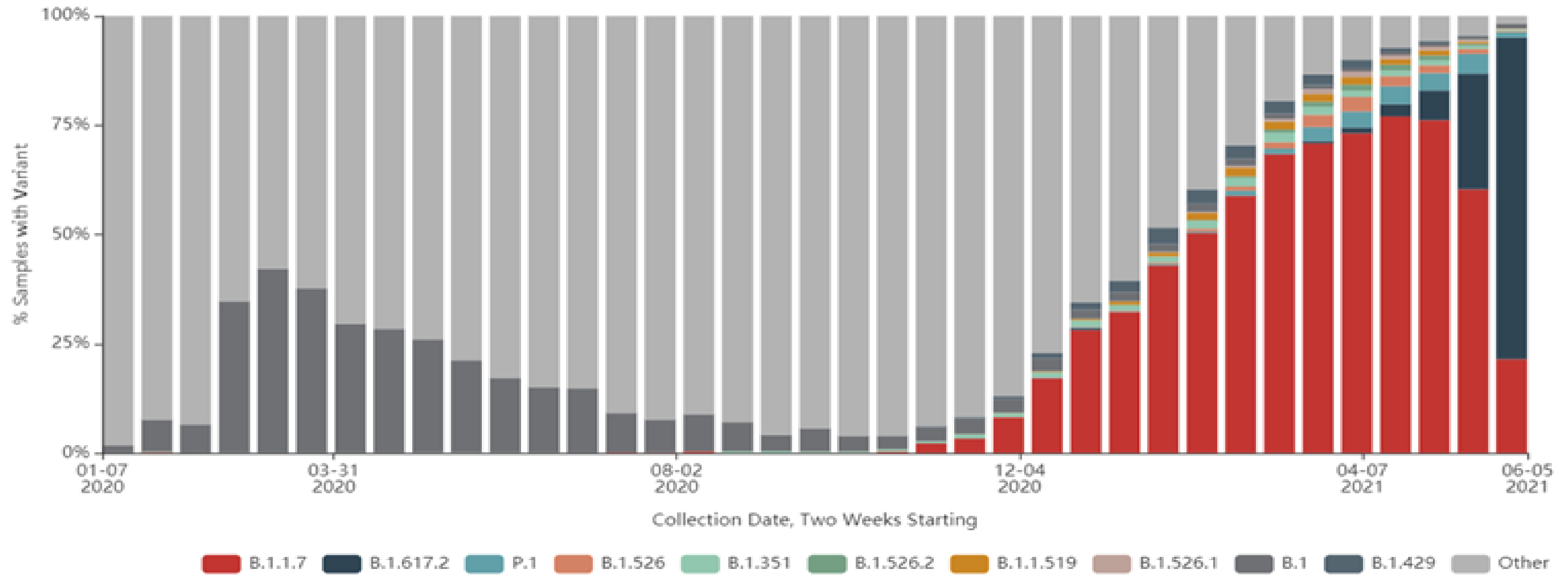

Fig 1: *Dominant SARS-CoV-2 strains sample distribution over time*

With constant emergence of new dominant strains (**Fig 1**), we require regular vaccine shots to boost our immunity, much like the seasonal flu. For this purpose, this study seeks to predict when and where the virus could mutate next, which can then be passed to virologists for analysis.

Our experiment focuses on 4 key variants of concern (VOCs) that have become a global concern due to higher transmissibility and virulence. These are namely the B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma) and B.1.617.2 (Delta) VOCs, constituting majority of present-day infections.

References

[1] Shu, Y., McCauley, J. (2021). GISAID EpiFlu™ Database. Sequence entries with complete collection date information shared via GISAID. https://www.gisaid.org/index.php?id=208

Methodology

The SARS-CoV-2 consists of approximately ~30,000 nucleotides (labeled as A, G, T, C respectively), each of which could undergo an independent and random mutation during replication (**Fig 2**).

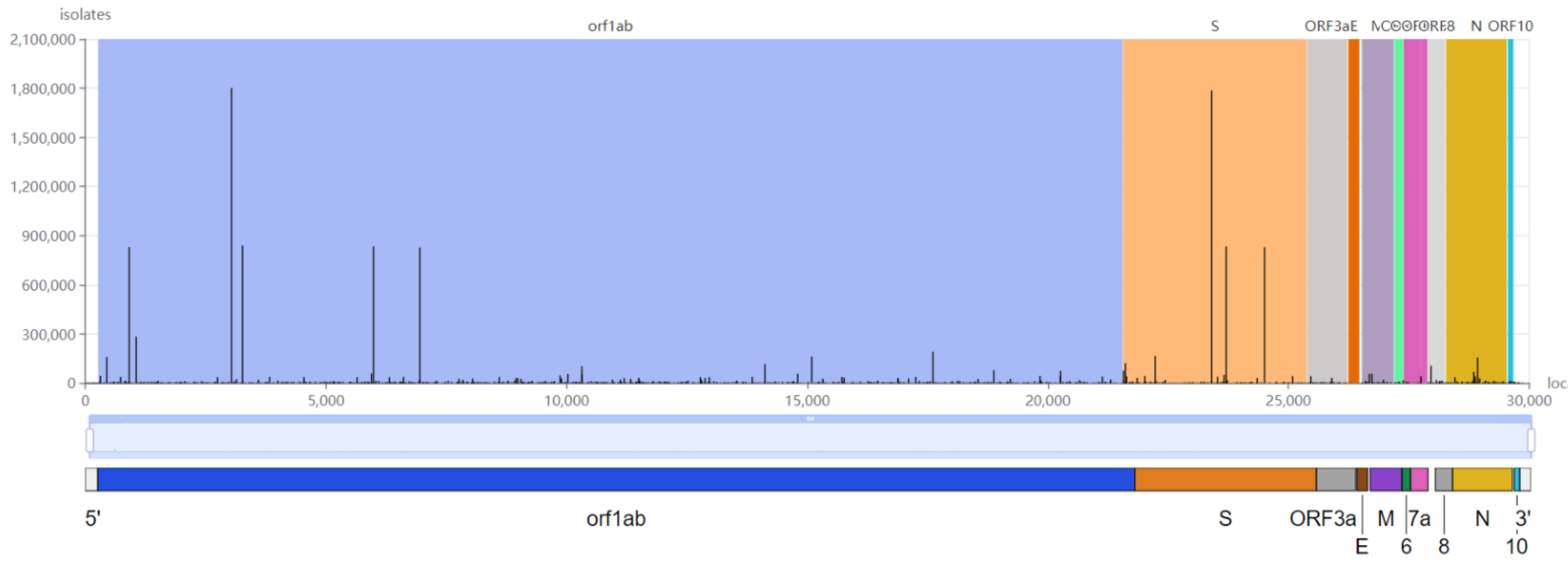

Fig 2: *Frequency of nucleotide mutations on SARS-CoV-2 genome; colours represent RNA components (eg. S=Spike, M=Membrane)*

Markov Chain embedded Poisson Process

Our method uses a Markov Chain Model to model the nucleotide substitution process, whereby each A, G, T, C are represented by four recurrent states which communicate with each other (**Fig 3**). This irreducible, positive-recurrent CTMC is embedded with a Poisson process wherein the transition rates of each nucleotide mutation is given by an arrival event with rate λ (**Fig 4**).

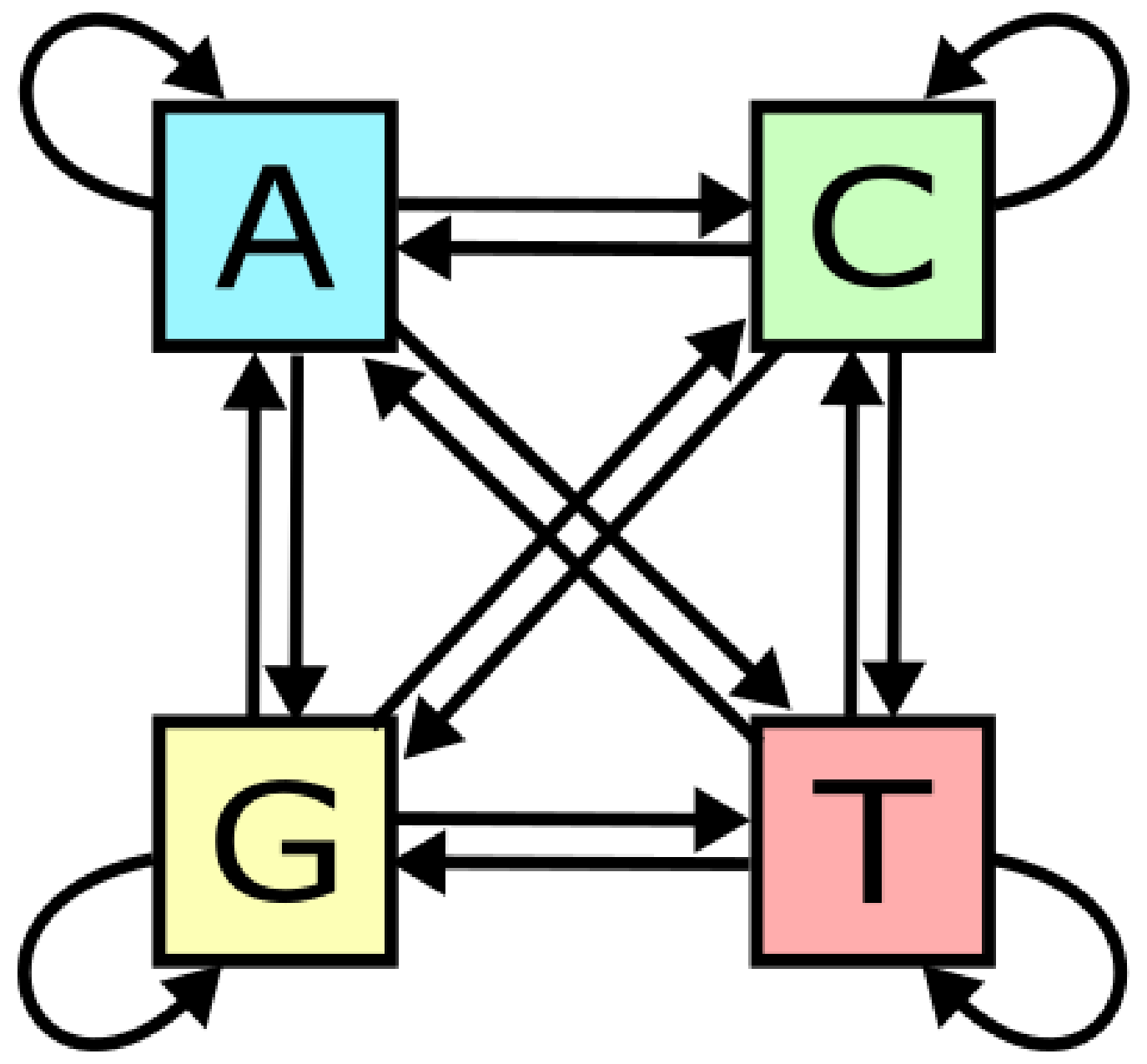

Fig 3: *State transition diagram*

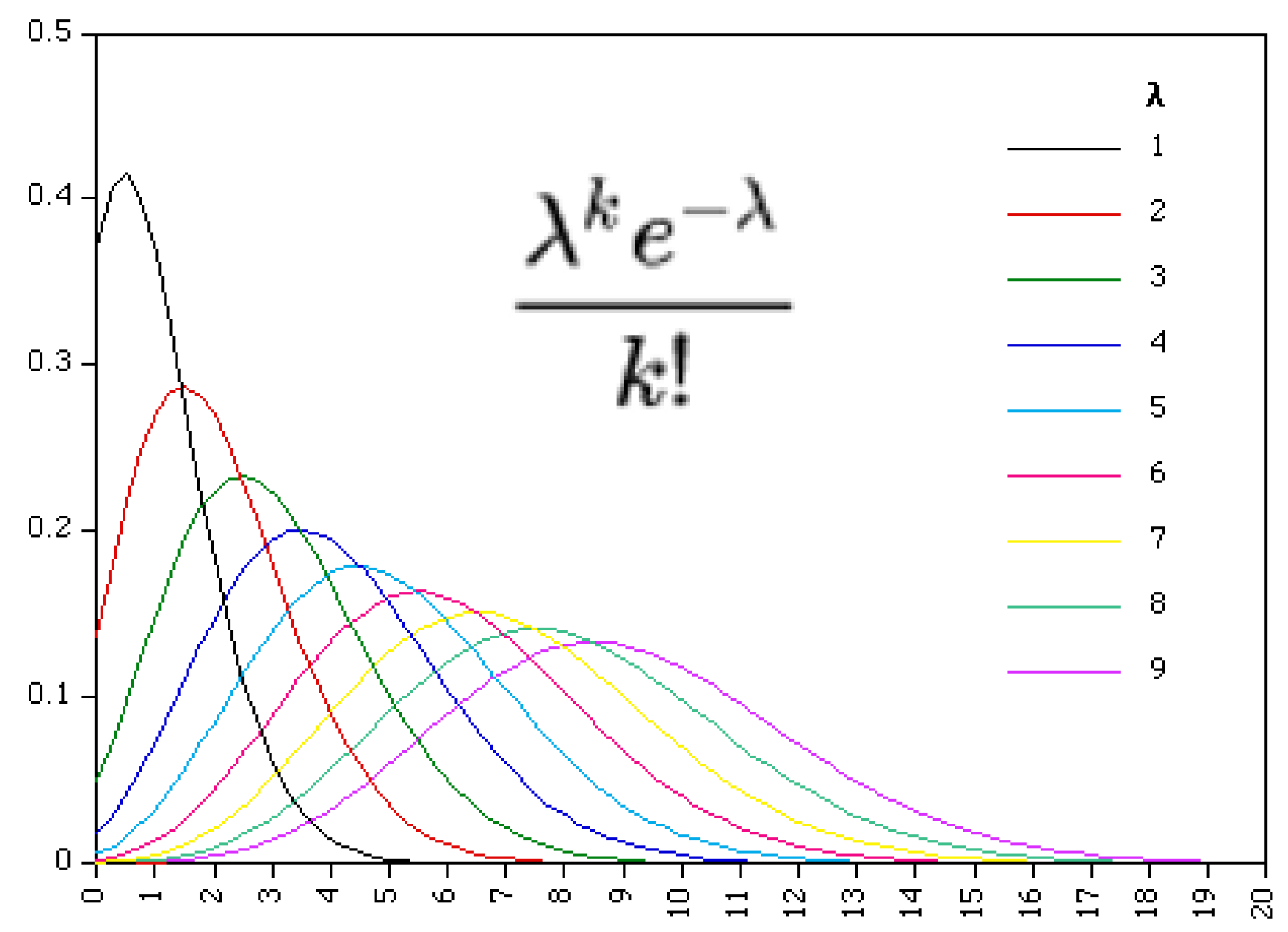

Fig 4: *Poisson processes with rate λ*

GISAID EpiFlu™ Database

This research accesses a registered GISAID account for SARS-CoV-2 genome sequences as the data source for training the model (**Fig 5**). GISAID[1] is a public data-sharing platform that aggregates virus genome submissions from around the world.

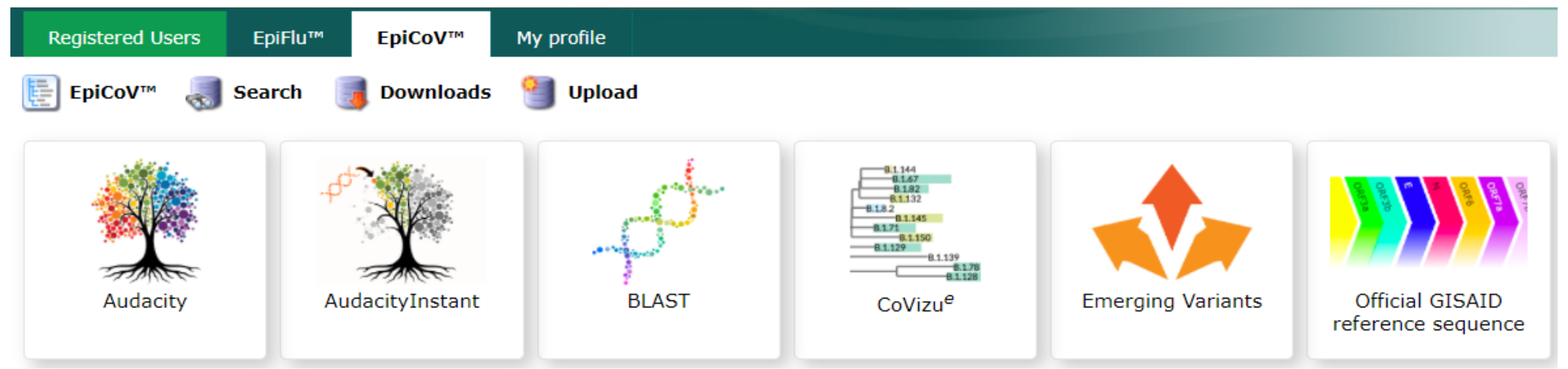

Fig 5: *GISAID registered user access dashboard*

Experimental Results

The Markov transition rates are plotted in a stacked bar chart with the respective mutation rate (**Fig 6**). Simulating the Markov Chain as a time series, we observe convergence to a limiting distribution (**Fig 7**).

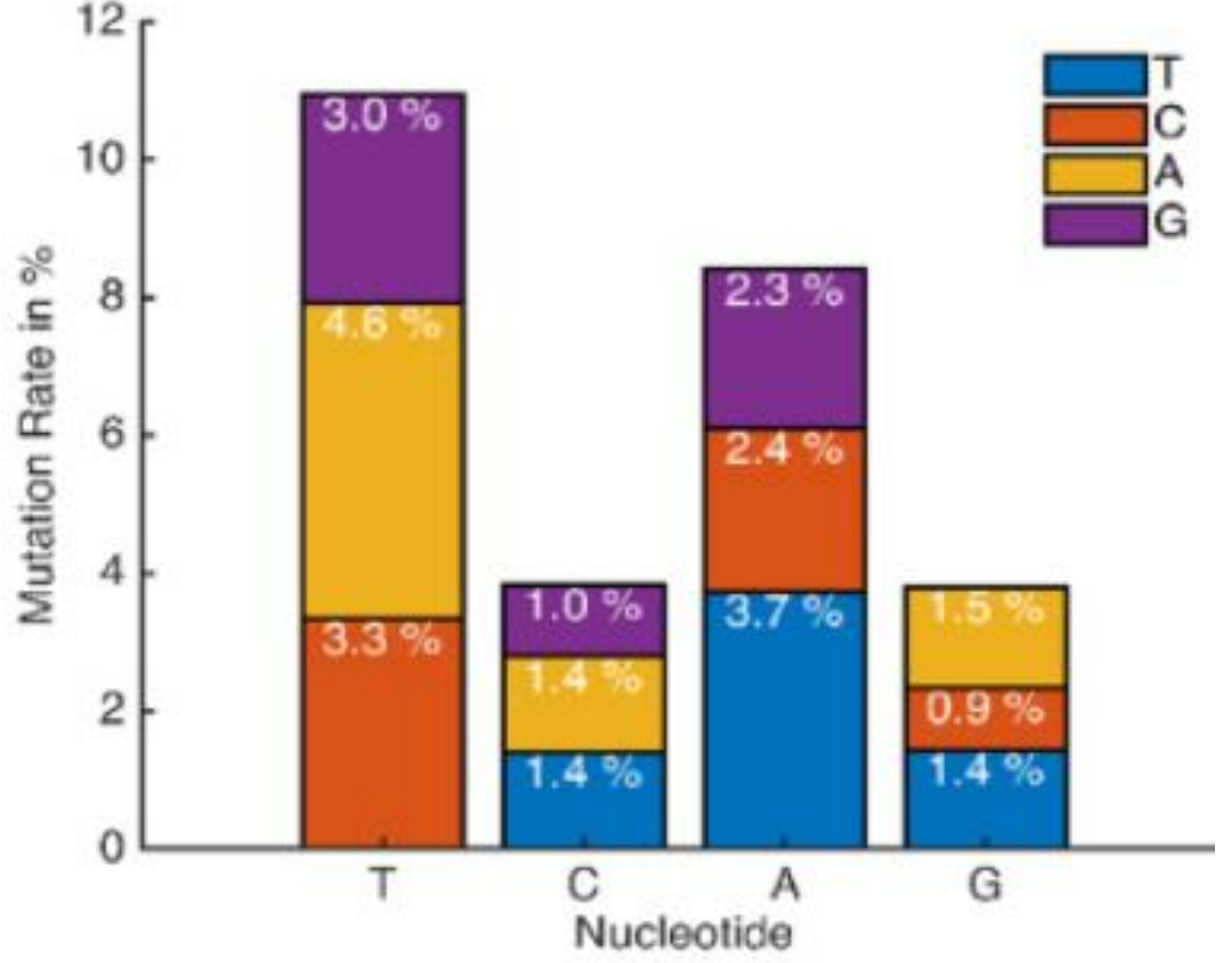

Fig 6: *Mutations rates (%)*

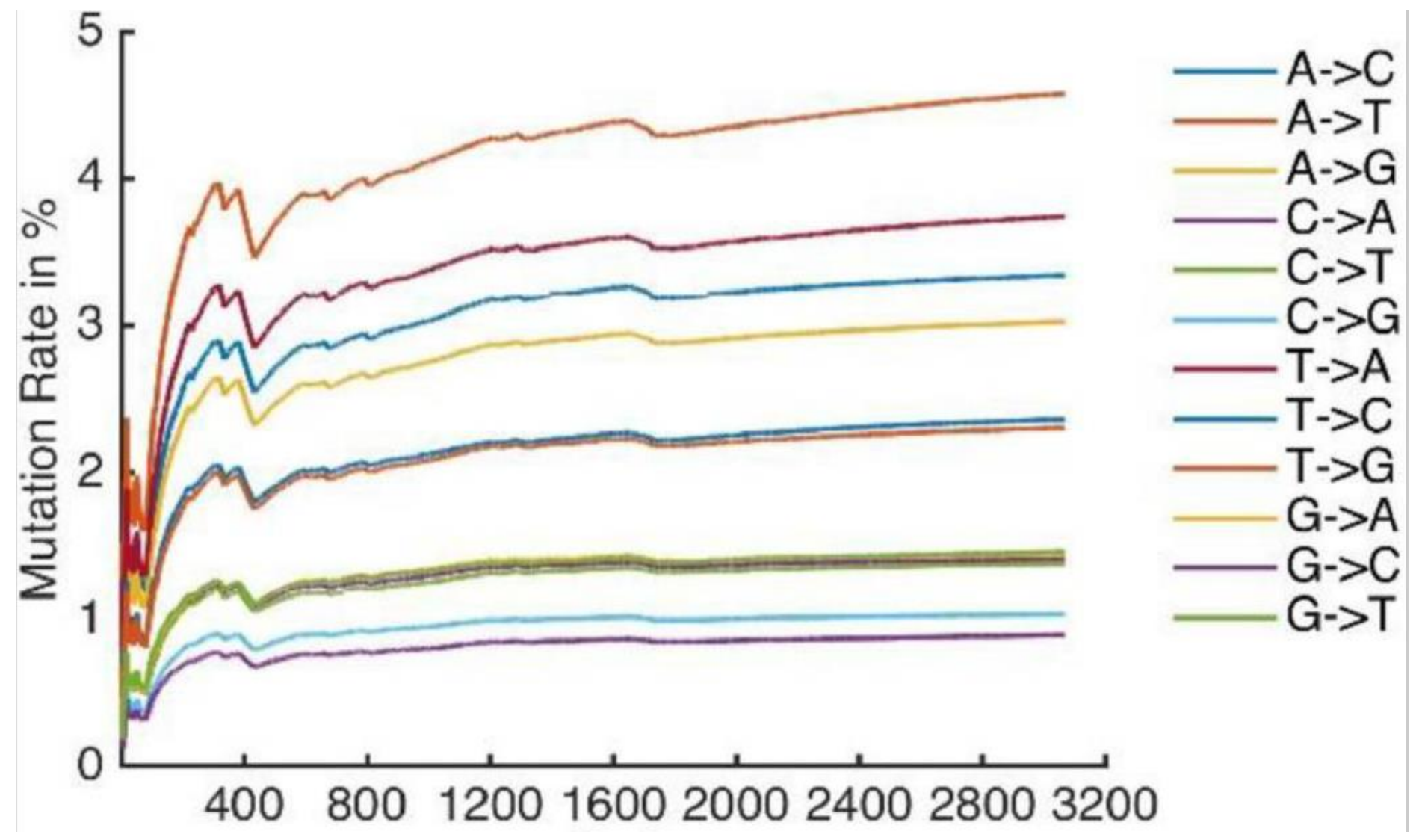

Fig 7: *Simulated time series*

Each VOC has a mean Poisson arrival rate based on their divergence from the original strain (**Fig 8**), which are multiplied with the above Markov transition probabilities to predict specific nucleotide mutations rate given as a 4x4 matrix for each of the four key VOCs (**Table 1**).

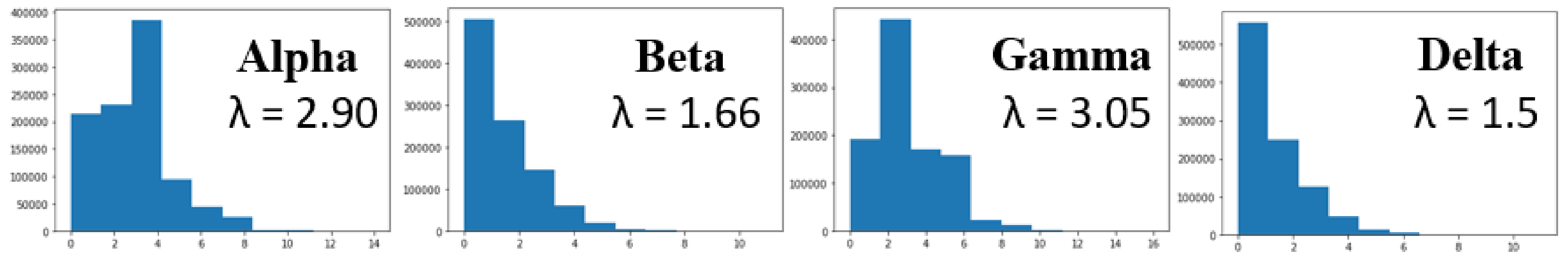

Fig 8: *Poisson process histograms of 4 key VOCs*

Alpha (B.1.1.7)				Beta (B.1.351)				Gamma (P.1)				Delta (B.1.617.2)			
2.664	0.030	0.069	0.045	1.426	0.016	0.037	0.024	2.718	0.031	0.070	0.046	1.337	0.015	0.035	0.023
0.090	2.876	0.072	0.027	0.048	1.539	0.038	0.014	0.092	2.934	0.073	0.027	0.045	1.443	0.036	0.014
0.138	0.042	2.739	0.042	0.074	0.022	1.466	0.022	0.140	0.043	2.794	0.043	0.069	0.021	1.374	0.021
0.099	0.042	0.111	2.876	0.053	0.022	0.059	1.539	0.101	0.043	0.113	2.934	0.050	0.021	0.056	1.443

Table 1: *4x4 transition rate matrices of 4 key VOCs*

Discussions & Conclusion

- Our study presents a method to model the mutation rate of SARS-CoV-2 virus as a Markov Chain embedded Poisson process.
- The calculated mean mutation rate is λ = 2.27 nucleotide mutations per month, which closely matches the literature value of 2.
- We can thus expect a new mutation to occur with inter-arrival time given by an exponential distribution with a mean of every 13 days.
- We found that the mutation rates for each key VOC differs slightly, suggesting that each strain could have unique virus characteristics.
- It is also worth noting that the nucleotides T (Thymine) and A (Adenine) have higher propensities to be substituted in the chain.
- **Limitations:** Need further expert opinion for good/bad mutations;
- Dataset could be biased toward reported 'bad' strains only;
- Data is skewed toward countries with more genome submissions.
- **Suggested future work**: Apply this model to descendants of VOCs.