CoVtRec

TDALife

Topological surveillance of potentially adaptive mutations in the evolution of the coronavirus SARS-CoV-2

CoVtRec report as of 10 January 2024

Michael Bleher[2*], Lukas Hahn[2*], Maximilian Neumann[*], Samuel Braun[3], Holger Obermaier[3], Mehmet Soysal[3], René Caspart[3], Andreas Ott[1,2*]

[1]Institute for Biological Interfaces, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Institute for Mathematics, Heidelberg University, Heidelberg, Germany
[3]Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany

[*]Correspondence: mbleher@mathi.uni–heidelberg.de (M.B.)
lhahn@mathi.uni–heidelberg.de (L.H.)
neumann.mn@icloud.com (M.N.)
andreas.ott@kit.edu (A.O.)

Abstract

The appearance of new variants in the evolution of the coronavirus SARS-CoV-2 underlines the importance of being able to quickly identify mutations that could confer some adaptive advantage to the virus, such as immune evasion or higher infectivity. Here we apply CoVtRec, a fast and scalable surveillance system based on Topological Data Analysis, for the identification of emerging potentially adaptive mutations in the ongoing evolution of SARS-CoV-2. CoVtRec is based on a new topological approach to the study of recurrent mutations in large genomic datasets developed in [1].

Results

We analyzed topological signals for the ongoing convergent evolution of the coronavirus SARS-CoV-2 on the Spike gene from 15 January 2019 until 10 January 2024. To that end, we performed a topological recurrence analysis for a curated alignment of 14,786,600 high-quality SARS-CoV-2 Spike gene sequences shared via GISAID, the global data science initiative [2, 3]. For each Spike mutation we computed its topological recurrence index (tRI) and the corresponding time series analysis chart. The topological recurrence index is a topological measure for the potential adaptiveness of a given mutation (see [1, 4] for details).

We present a list featuring the top ten amino acid changes on the Spike gene with currently highest tRI growth rate in December 2023 (see Table 1). Here signals with tRI $\geq$ 74 are statistically significant ($p < 0.05$). It was demonstrated in [1] that these mutations are potentially adaptive and might therefore appear in future variants. We also present time series analysis charts (see Figure 1) showing (i) the development of the topological signal as well as its significance over time, (ii) major lineages containing the mutation, and (iii) the date from which on the topological signal became significant.

SAAV	tRI	notable variants
L455F	439	Omicron
G184S	82	Omicron
A475V	101	Omicron
E554K	122	Omicron
P521T	109	Omicron
N185D	133	Omicron
L452R	451	Delta, Omicron, Kappa
F157L	406	Omicron
Y248H	144	Omicron
S256L	350	Omicron

Table 1. The top ten amino acid changes on the Spike gene with currently highest tRI growth rate in December 2023. For a given mutation, the table displays its topological recurrence index (tRI) and notable variants containing the mutation.
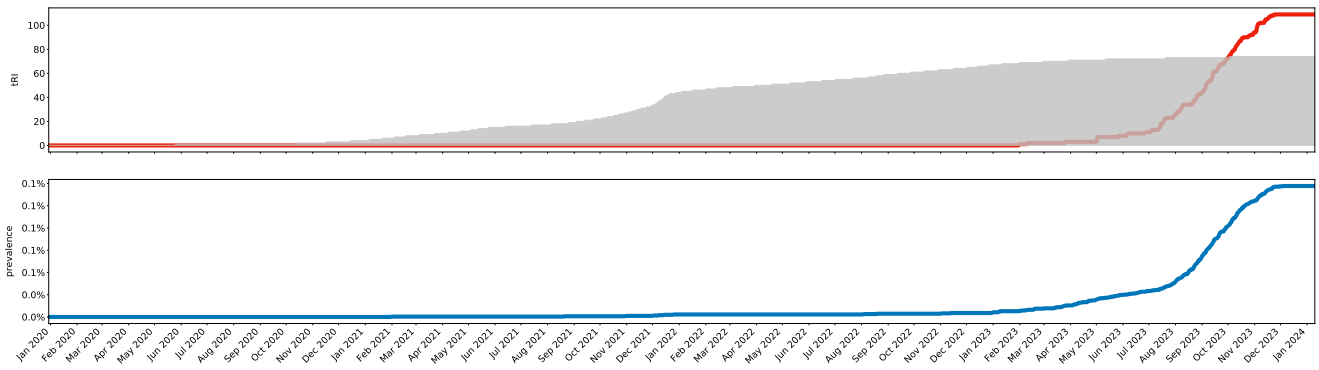
Figure 1. Time series analysis charts for the mutations listed in Table 1. Each chart shows the topological recurrence index (tRI, in red) and the prevalence (in blue) of the respective mutation from 15 January 2019 until 10 January 2024. In the upper diagram the shaded region marks the level of significance.
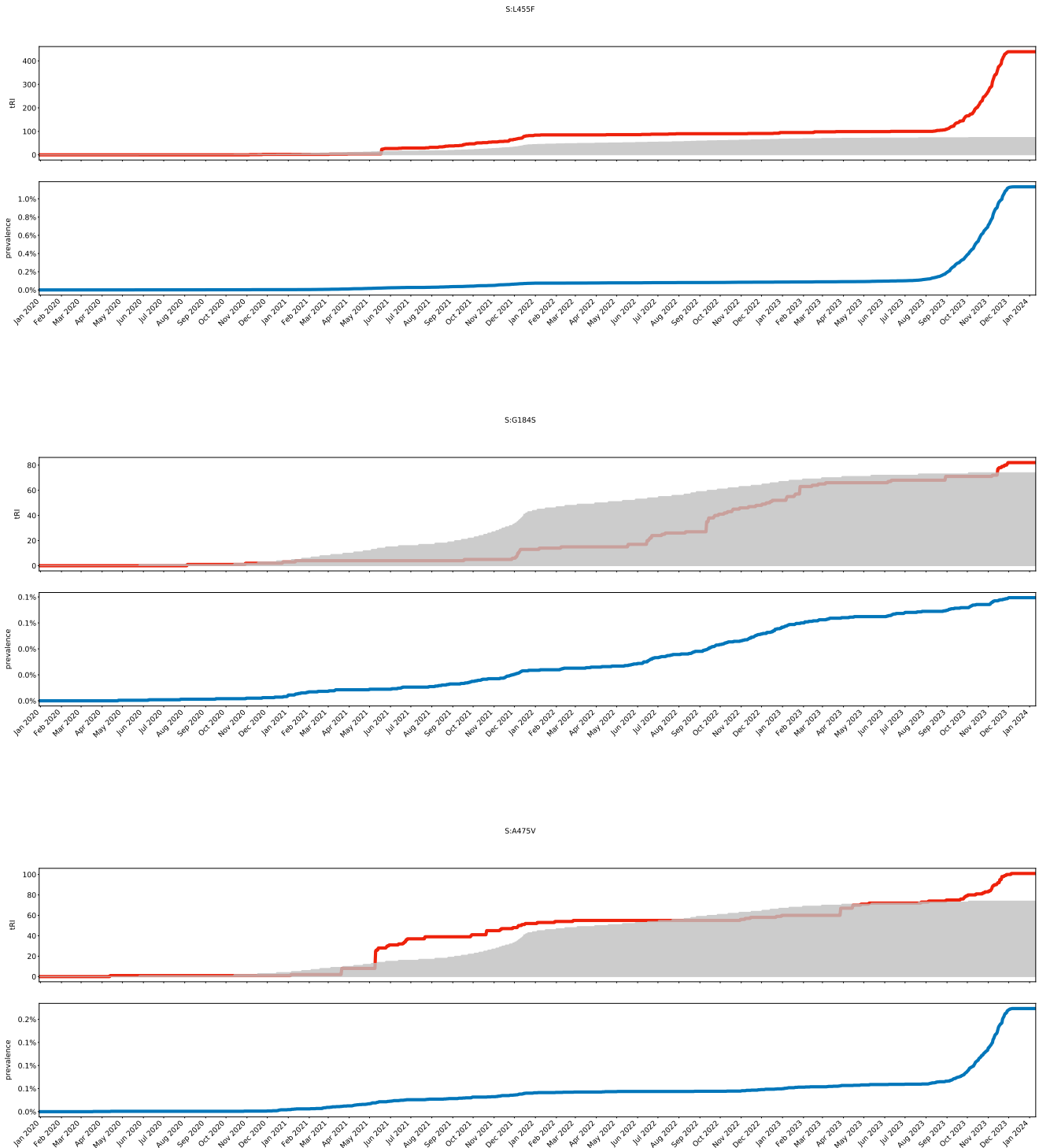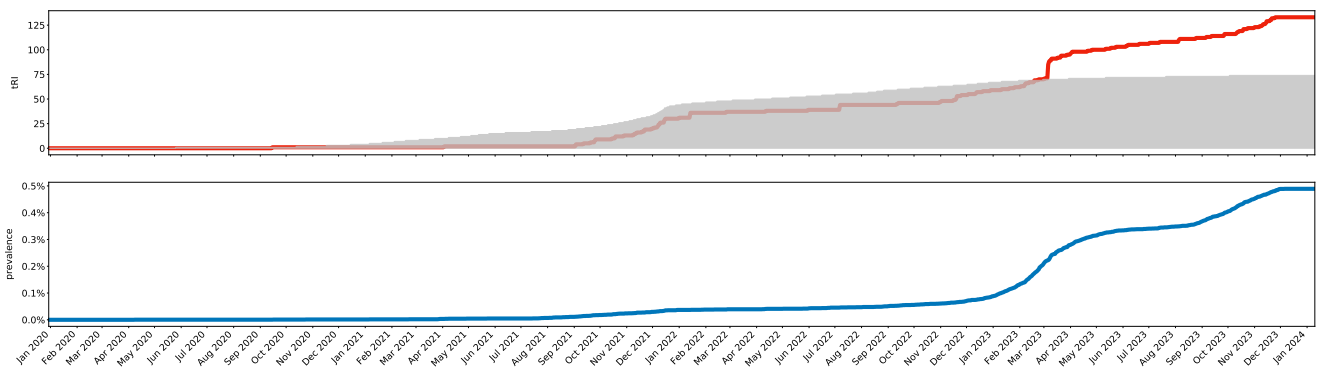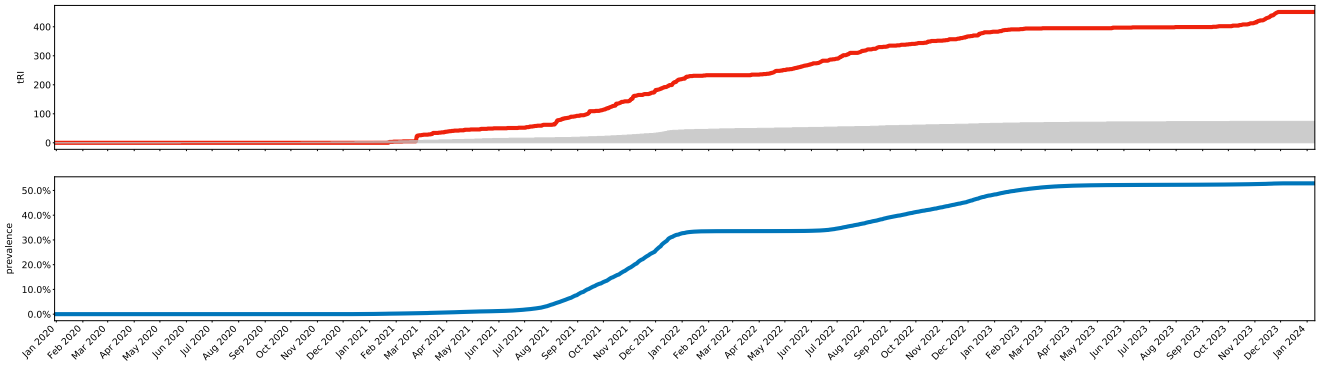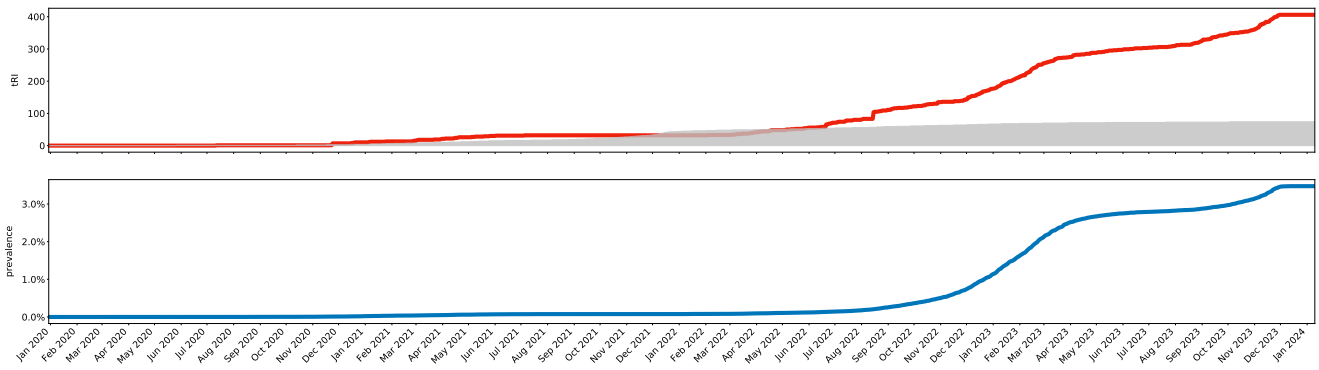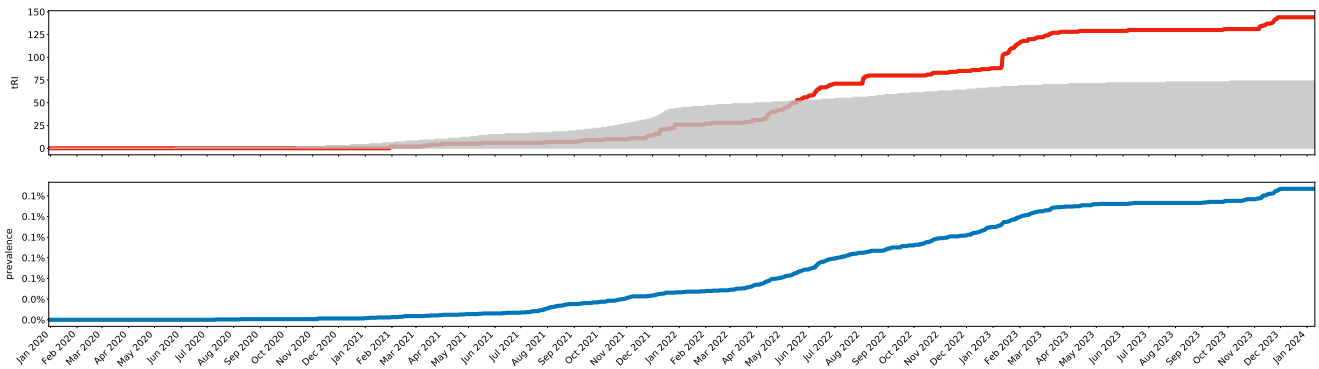
S:E554K | contained in Omicron
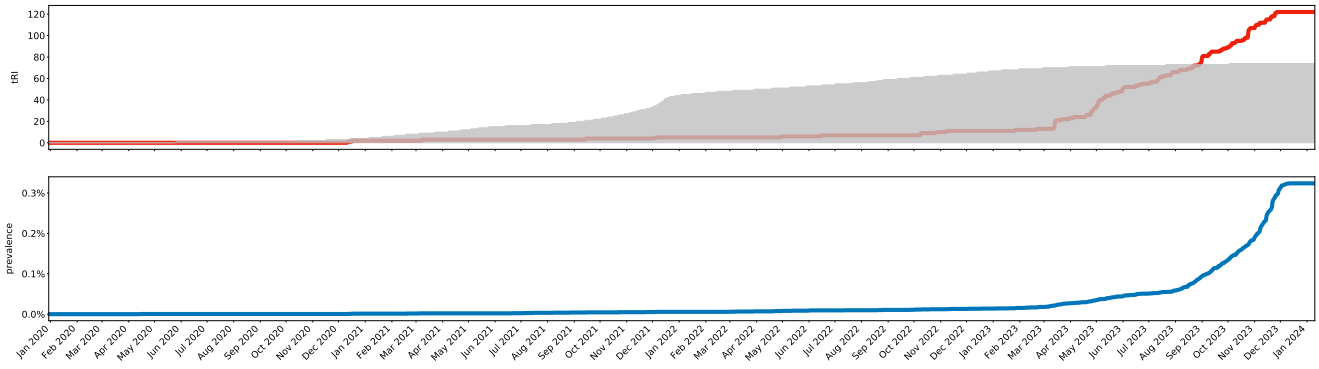
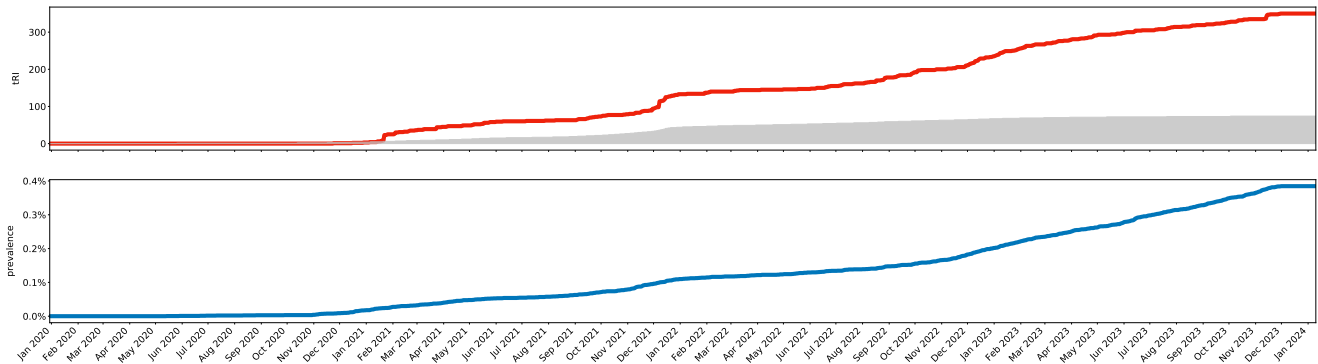S:P521T

S:N185D

S:L452R | contained in Delta, Omicron, Kappa

S:F157L | contained in Omicron

S:Y248H

S:S256L

Methods

Data acquisition and data preparation

Our analysis is based on the alignment `msa_0110.fasta` downloaded from the GISAID EpiCoV Database [2, 3] on 12 January 2024. This alignment comprises 14,786,600 SARS-CoV-2 whole genome sequences that have been aligned to the reference sequence Wuhan/WIV04 with GISAID accession number `EPI_ISL_402124` using `MAFFT` [5]. Sequences in this alignment were truncated to the Spike gene (reference site positions 21,563 to 25,384). Subsequently the following sequences were removed: (i) sequences containing any characters other than A, C, T, G or -, and (ii) sequences that occur only once in the alignment. This resulted in an alignment comprising 9,548,724 complete SARS-CoV-2 Spike genes of length 8,165nt.

Topological recurrence analysis

The Spike gene alignment contains a subalignment comprising 246,738 genetically distinct sequences. A list of accession numbers of all sequences in this subalignment, along with an acknowledgement of the contributions of both the submitting and the originating laboratories, is accessible at https://doi.org/10.55876/gis8.240118yh. We used `Hammingdist` (`Version 0.19.0`) [6] to compute the genetic distance matrix of this subalignment. Subsequently we used `Ripser` [7] to compute the representative cycles for the persistent homology of the Vietoris–Rips filtration associated to the genetic distance matrix. The computation of persistence barcodes was restricted to small genetic distance scales (`Ripser` scale parameter threshold set to 2). Next a complete list of SNV cycles (topological cycles all of whose edges correspond to single nucleotide variations) in the given alignment was generated from the corresponding `Ripser` output. Then we used custom code implemented in Python to compute the *topological recurrence index (tRI)* for each such SNV. Summing over all SNVs determining an SAAV (single amino acid variation), we computed the tRI for each SAAV. Lastly, from the distribution of the tRI measurements over the whole Spike gene we inferred the level of significance for the tRI per SAAV. Using Vietoris-Rips transformations in multipersistent homology, we computed tRI time series analysis charts at daily resolution from the natural stratification by time of genomic data. We also computed the prevalence of the given mutation in the subalignment comprising 246,738 genetically distinct sequences. For a more detailed description of the topological recurrence analysis see [1, 4].

Data availability

All SARS-CoV-2 genome data used in this work are available from the GISAID EpiCov Database [2, 3] at https://www.gisaid.org and are accessible at https://doi.org/10.55876/gis8.240118yh.

Code availability

Code used for the analyses is available at https://github.com/ssciwr/hammingdist and https://github.com/Ripser/ripser/tree/tight-representative-cycles. All other code is available from the corresponding authors upon request.

Acknowledgements

Author contributions

M.B., L.H., M.N., A.O. designed the study; A.O. curated data; A.O. performed computational analyses; M.B., L.H., M.N., A.O., S.B., H.O., M.S., R.C. developed and implemented software; M.B., L.H., A.O. acquired computing resources; M.B., L.H., A.O., M.N. drafted the manuscript; all authors contributed to the final version of the report.

Competing interests

The authors declare no competing interests.

References

[1] M. Bleher, L. Hahn, J. Á. Patiño-Galindo, et al. "Topological data analysis identifies emerging adaptive mutations in SARS-CoV-2". *medRxiv* (2021). DOI: 10.1101/2021.06.10.21258550.

[2] Yuelong Shu and John McCauley. "GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality". *Eurosurveillance* 22.13 (2017). DOI: 10.2807/1560-7917.ES.2017.22.13.30494.

[3] Shruti Khare, Céline Gurry, Lucas Freitas, et al. "GISAID's Role in Pandemic Response". *China CDC Weekly* 3.49 (2021), pp. 1049–1051. DOI: 10.46234/ccdcw2021.255.

[4] M. Bleher, L. Hahn, M. Neumann, et al. "MuRiT: efficient computation of pathwise persistence barcodes in multi-filtered flag complexes via Vietoris-Rips transformations". *arXiv* (2022). DOI: 10.48550/arXiv.2207.03394.

[5] K. Katoh. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". *Nucleic Acids Research* 30.14 (2002), pp. 3059–3066. DOI: 10.1093/nar/gkf436.

[6] Liam Keegan and Dominic Kempf. *Hammingdist: A Fast Tool to Calculate Hamming Distances*. Version 0.15.0. 2021. URL: https://github.com/ssciwr/hammingdist.

[7] Ulrich Bauer. "Ripser: efficient computation of Vietoris-Rips persistence barcodes". *Journal of Applied and Computational Topology* (2021). DOI: 10.1007/s41468-021-00071-5.

[8] Maximilian Hanussek. *VALET*. 2021. URL: https://github.com/MaximilianHanussek/VALET.