Predicting the behavior of COVID-19 pandemics

PEDRO MIGUEL DA SILVA FERREIRA

BIOMEDICAL SCIENCES DOCTORAL DEGREE – ICBAS

SUPERVISOR: CRISTINA P. VIEIRA, PHD

CO-SUPERVISORS: HUGO LÓPEZ-FERNÁNDEZ, PHD; JORGE VIEIRA, PHD

PHENOTYPIC EVOLUTION GROUP – IBMC/I3S

Fundação para a Ciência e a Tecnologia

State of the Art

State of the Art

- **Coronaviruses** belong to

Orthocoronavirinae

subfamily that splits into

four genera

α-CoV, β-CoV, δ-CoV, γ-CoV

only infect Mammals | **infect Avian species**

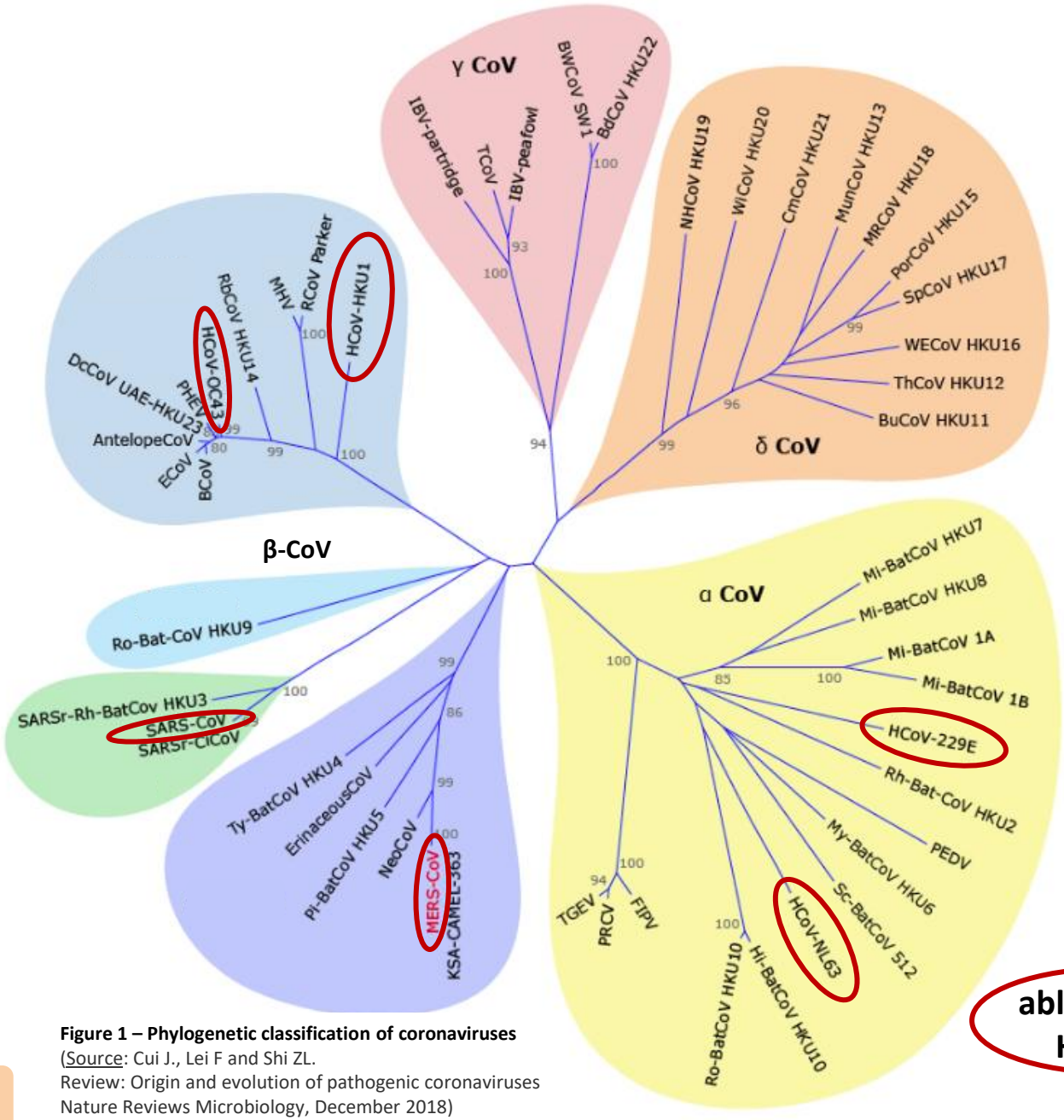


Figure 1 – Phylogenetic classification of coronaviruses
(Source: Cui J., Lei F and Shi ZL.
Review: Origin and evolution of pathogenic coronaviruses
Nature Reviews Microbiology, December 2018)

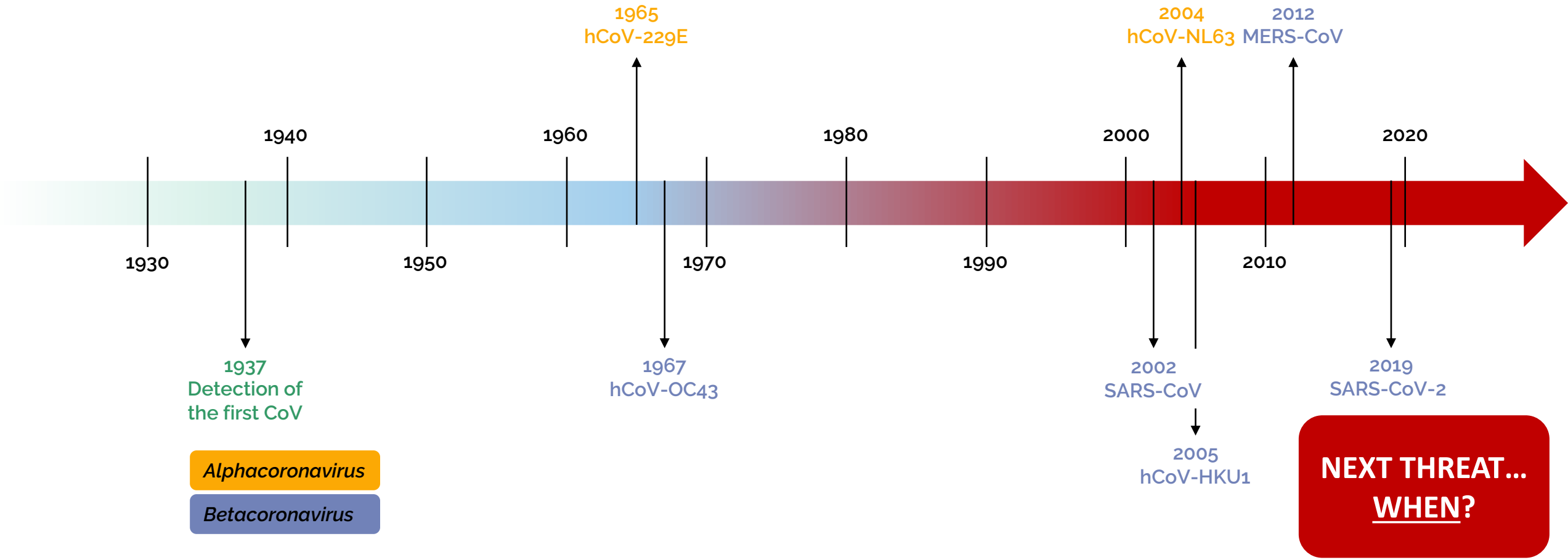
able to infect Humans

U.PORTO

ICBAS | INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
SCHOOL OF MEDICINE AND BIOMEDICAL SCIENCES

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Timeline of **Human Coronaviruses**



COVID-19 Worldwide / Portugal

Total deaths

7 million

Total confirmed cases

705 million

Total deaths

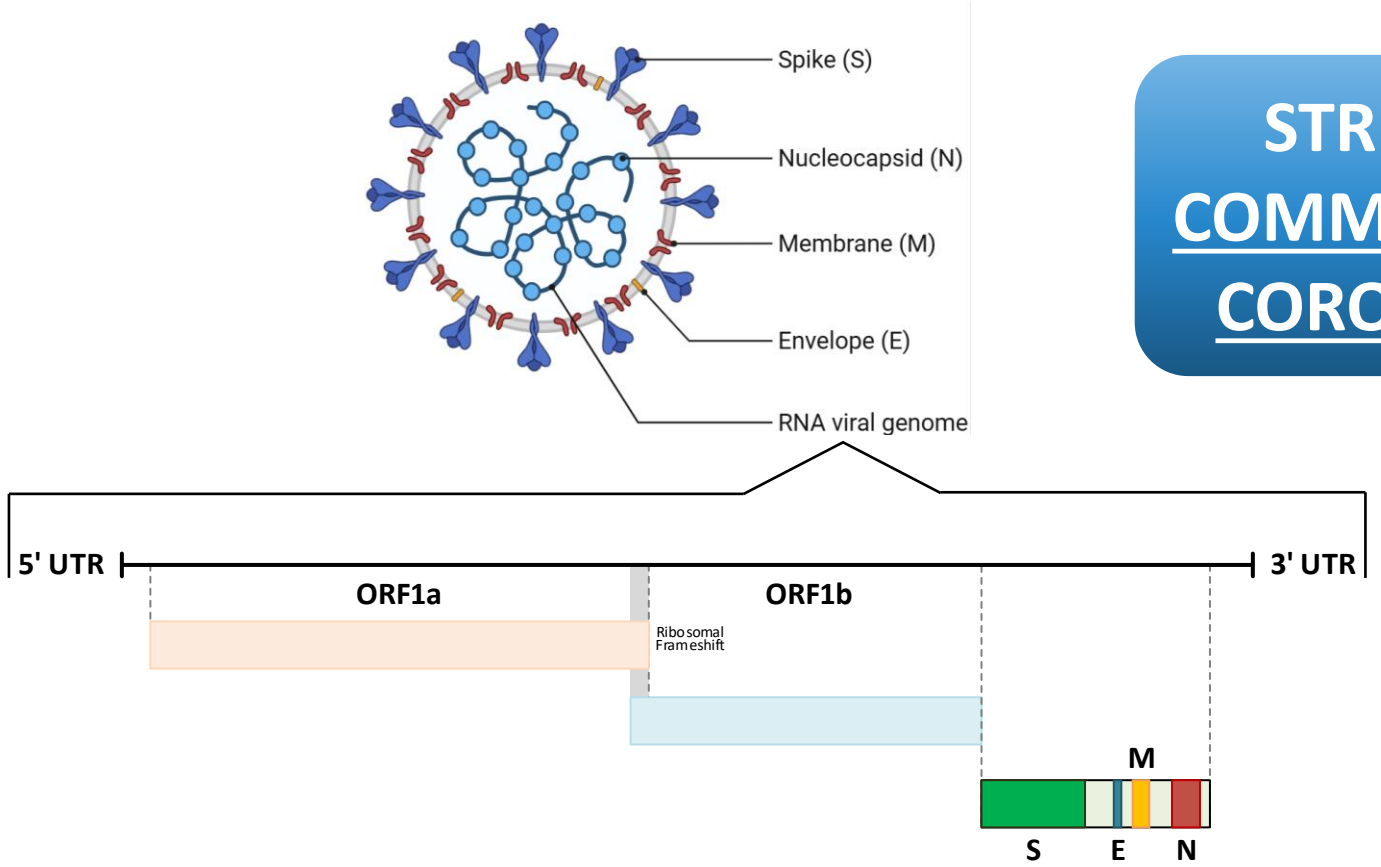
28 thousand

Total confirmed cases

5.6 million

Source: Worldometer, April 13th 2024
https://www.worldometers.info/coronavirus/

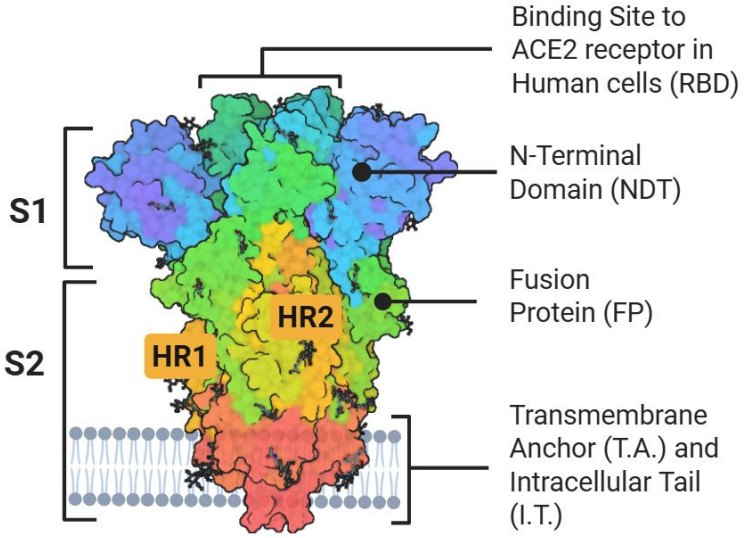
State of the Art - **Coronavirus**



STRUCTURE COMMON TO ALL CORONAVIRUS

Figure 2 – Coronavirus genome structure. Adapted from Rastogi *et. al.*, 2020.

State of the Art – **Spike** Protein (S)



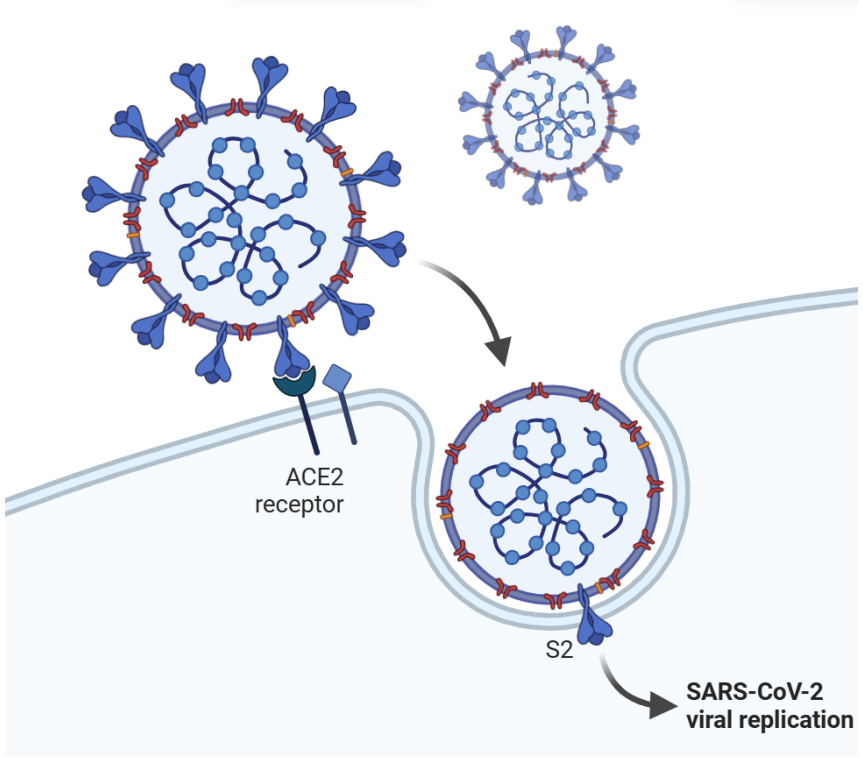
Source: Biorender

- **S** protein is responsible for **receptor binding**.
- It **forms homotrimers** through the viral surface.
- **S** has two functional **subunits** called **S1** and **S2**:
  - **S1** subunit constitutes the **trimers' apex** and includes the **receptor-binding domain** (RBD), responsible for **binding to host cell receptors**.
  - **S2** subunit is anchored in the viral membrane and **mediates membrane fusion**, enabling **CoVs entry into the host**.

State of the Art – **ACE2** Receptor

- **Using the S protein** on its surface, the **SARS-CoV-2** virus **binds to ACE2** prior to entry and infection of host's cells.

- **<u>ACE2 Receptor</u>:** protein on the surface of many cell types.
  - This enzyme is envolved in the regulation of cardiovascular and renal functions.



<u>Source</u>: Biorender

State of the Art – PSS

- **Positively selected amino acid sites** (PSAAS or **PSS**) provide important information about a **protein's function**

  - **PSS**: Amino acid positions that **show more changes than expected** under a neutral evolutionary model (non-synonymous / synonymous ratio > 1)

State of the Art – PSS

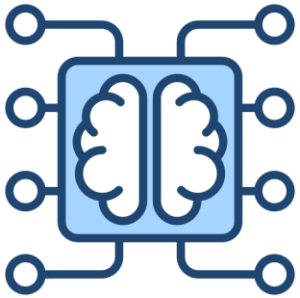
- **PSS** can be identified by **analyzing protein-coding DNA sequences**, using Markov models of codon evolution combined with maximum likelihood methods (**FUBAR and codeML**) as well as a population genetics approximation to the coalescent with recombination (**omegaMap**).

- Application of **more than one method** will strengthen the **PSS** inferences.

Objectives

- Develop efficient and **simple-to-use tools** to conduct **large-scale analyses**

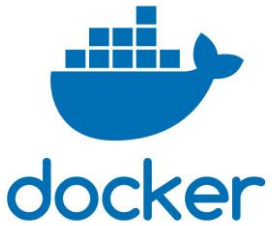
- Infer **Positively Selected Amino Acid Sites** (**PSS**) in **Coronavirus** species using **Auto-PSS-Genome**



- Develop **machine learning models** based on **PSS** found with **Auto-PSS-Genome**
  - Predict **SARS-CoV-2** spread **patterns** that will allow making **predictions** about the future behavior of COVID-19

Methodology

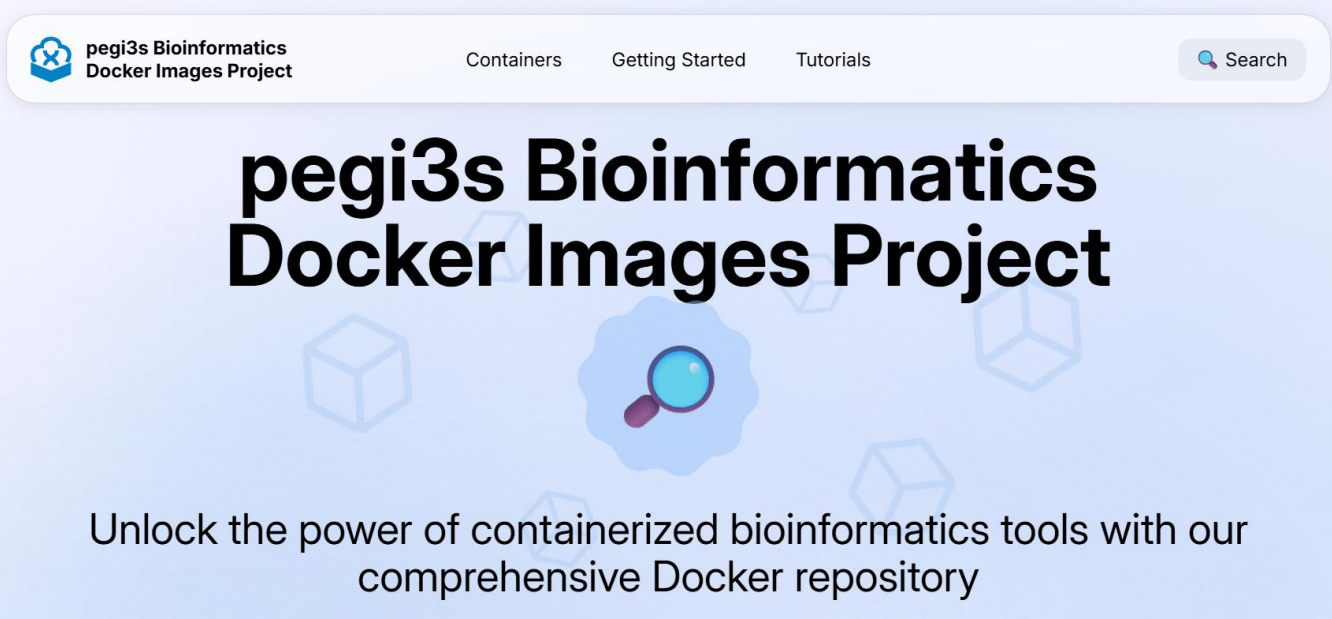
Auto-PSS-Genome



Docker Images Project

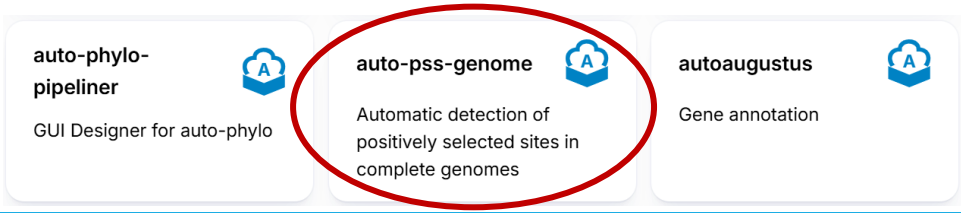
+ 278K Pulls

+ 150 Images



bdip.i3s.up.pt

- **Easy to use** since it only requires the installation of Docker;
- **Portable** between computers and immutable;
- Docker images can be **downloaded when needed** and **erased** when **no longer needed**;
- Tools classified in broad categories;
- **Clear instructions** on how to use the images, with test cases;
- **Docker images** can be **incorpored** in **pipelines**.



Auto-PSS-Genome

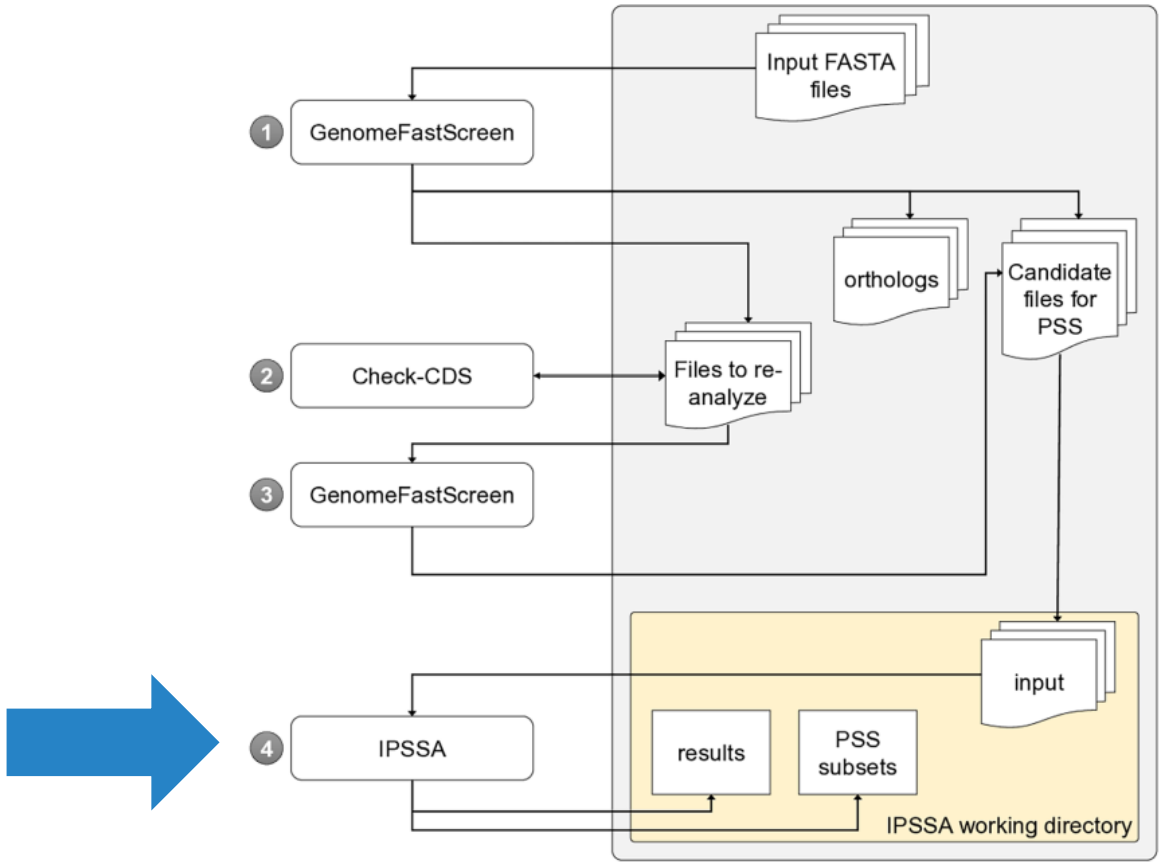


Figure 3 – Steps and files involved in the Auto-PSS-Genome pipeline.

Relies on the usage of 3 COMPI pipelines:

- ✓ **GenomeFastScreen**
- ✓ **CheckCDS**
- ✓ **IPSSA**

docker

COMPUTATIONAL PIPELINES COMPI

IPSSA (Running Times)

Interdisciplinary Sciences: Computational Life Sciences

Running times of the main steps involved in the IPSSA pipeline using a different number of sequences

# Sequences	Alignment method's execution times (s)					MrBayes (min)	FUBAR (s)	codeml (h)	omegaMap (min)
	Clustalw	Muscle	Kalign	t_coffee	amap				
10	3	2	2	20	5	8.35	17	0.15	0.12
20	5	3	3	86	18	16.17	18	1.19	9.92
30	7	3	3	166	42	36.13	25	**4.24**	21.12
⋮									
80	37	7	4	1281	413	182.7	67	n.a	102.67
90	48	**8**	5	1559	546	**187.65** ≈3h	**68** ≈1min	n.a	**132.22** ≈2h

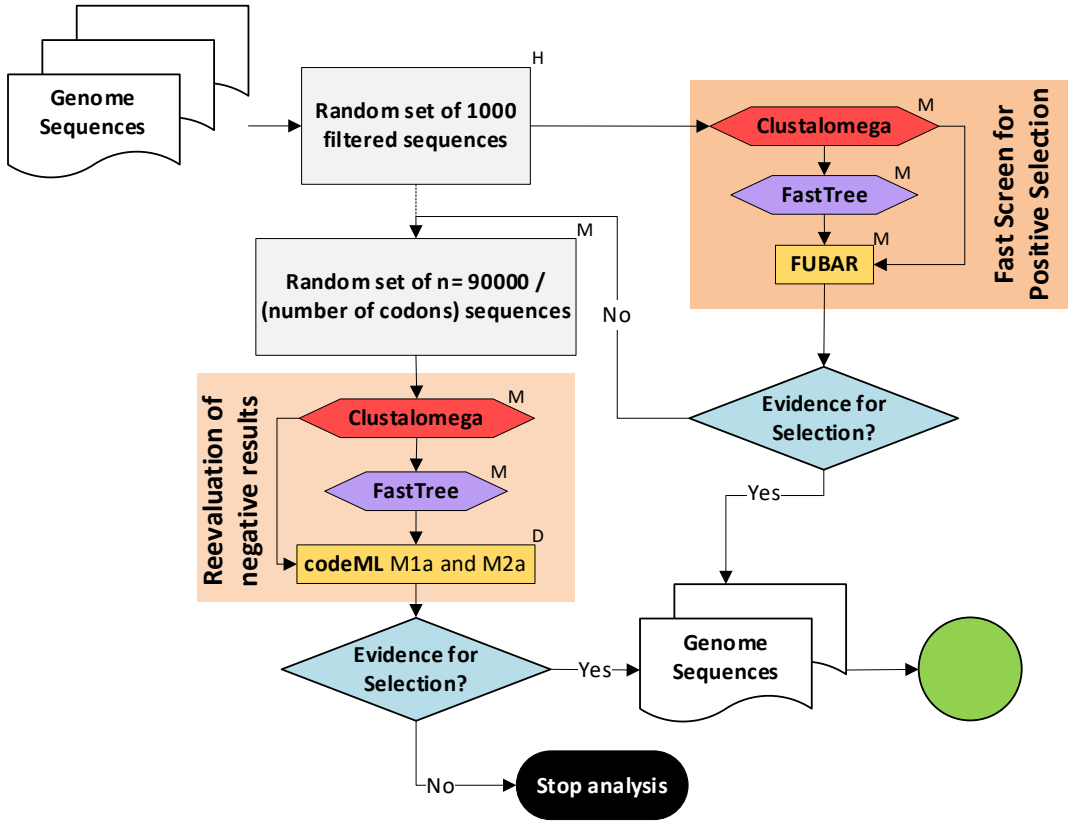
In bold-underline are the default values for IPSSA

PC specs

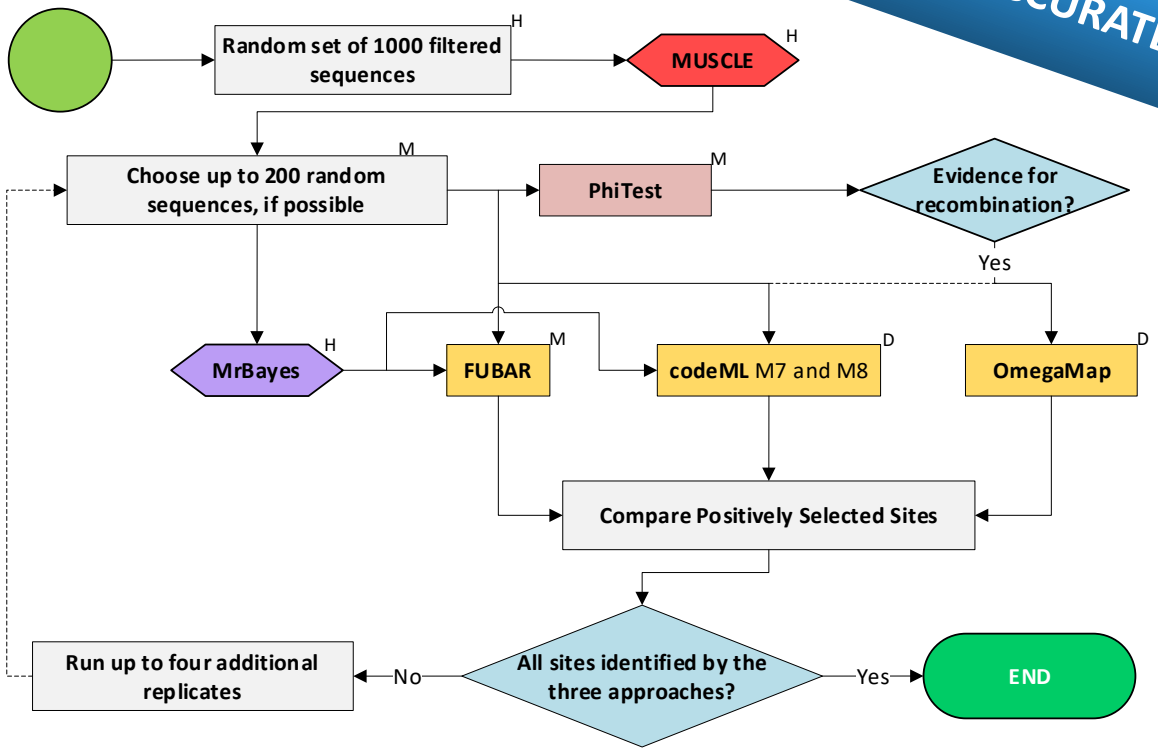
Memory 7,7 GiB
Processor Intel® Core™ i7-3540M CPU @ 3.00GHz × 4

Auto-PSS-Genome

MORE TIME-CONSUMING BUT MORE ACCURATE



1. Fast Screen for potencial PSS



2. Detailed PSS analyses

Data

Datasets – BV-BRC

non SARS-CoV-2

- **15** Datasets / CoVs Species
- **1630** Sequences
- **945** Seqs α-CoV
- **625** Seqs *β-CoV*
- **60** Seqs δ-CoV

SARS-CoV-2

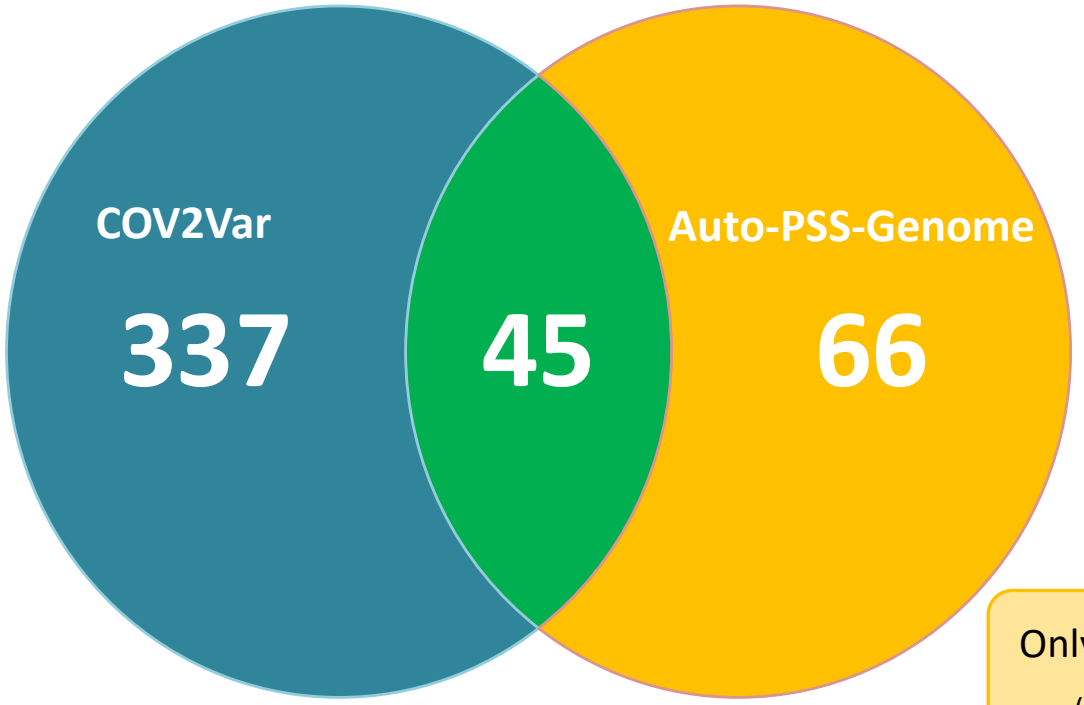
- **2019 – 2023** Years
- **100 Runs** 30 Seqs per Run (6 per year)

Multiple Lines of Evidence support 199 SARS-CoV-2 PSS

Lines of Evidence – Support PSS

Multiple Lines of Evidence support the inference of SARS-CoV-2 PSS

COV2Var vs Auto-PSS-Genome



Only **SARS-CoV-2** Data

(**FEL, FUBAR** and **MEME**)

Only **SARS-CoV-2** Data

(**FUBAR** and **codeML**)

≈ 40.5%

Our **PSS** **predicted** only **in SARS-CoV-2** are **also COV2Var PSS**

(**45 out of 111**)

Figure 4 – Venn diagram showing the overlap of the PSS in the COV2Var database (COV2Var-int-5%) identified by the three methods used (FEL, FUBAR, and MEME), and those here identified (Auto-PSS-Genome (yellow)).

GISAID vs Auto-PSS-Genome

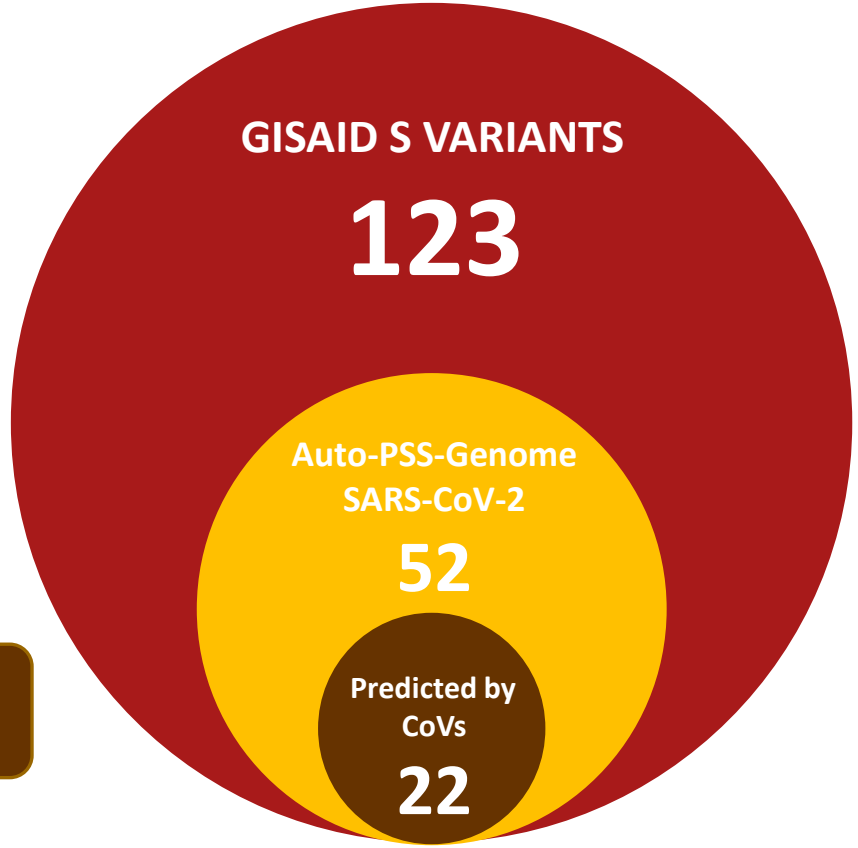
Only **SARS-CoV-2** Data

Only **SARS-CoV-2** Data

(**FUBAR** and **codeML**)

Only **non-SARS-CoV-2** Data

(**FUBAR** and **codeML**)



≈ 42.3%

a) PSS in Spike SARS-CoV-2 are **also PSS predicted by CoVs**

(**22 out of 52**)

b) GISAID sites are **PSS in Spike SARS-CoV-2**

(**52 out of 123**)

Figure 5 – Venn diagram showing the overlap of the sites in the GISAID database and the PSS here identified (Auto-PSS-Genome (yellow and brown)).

PSS Prediction Main **Results**

Table 1 – PSS identified in more than one non-SARS-CoV-2 coronavirus.

Protein		Datasets	SARS-CoV-2 Positions
S		PEDV–Betacoronavirus1	75
		PEDV–Alphacoronavirus1	Gap 97-98
M		Unknown Bat-CoV–Bat-CoV-HKU9–Bat-CoV-HKU10	4
		Bat-CoV-HKU2–Bat-CoV-HKU9–Bat-CoV-HKU10–Alphacoronavirus1	3
		Bat-CoV-HKU2–Bat-CoV-HKU10	6
N		Porcine-CoV-HKU15–Bat-CoV-HKU9	62
		Alphacoronavirus1–hCoV-HKU1	91
		hCoV-HKU1–Murine-CoV	289
ORF1ab	NSP3	hCoV-HKU1–hCoV-NL63	112
		Murine-CoV–PEDV	162
		Murine-CoV–Betacoronavirus1	1234
	NSP6	Murine-CoV–PEDV	138
	NSP12	MERS-CoV–PEDV	9

13

- **PSS** identified in **more than one CoV**

(in structural and non-structural proteins)

PSS Prediction Main Results

Table 2 – PSS identified in non-SARS-CoV-2 coronavirus. PSS Common represents sites identified by both methods. Homologs in SARS-CoV-2 (the PSS that can be aligned with the SARS-CoV-2 sequences) are shown in brackets.

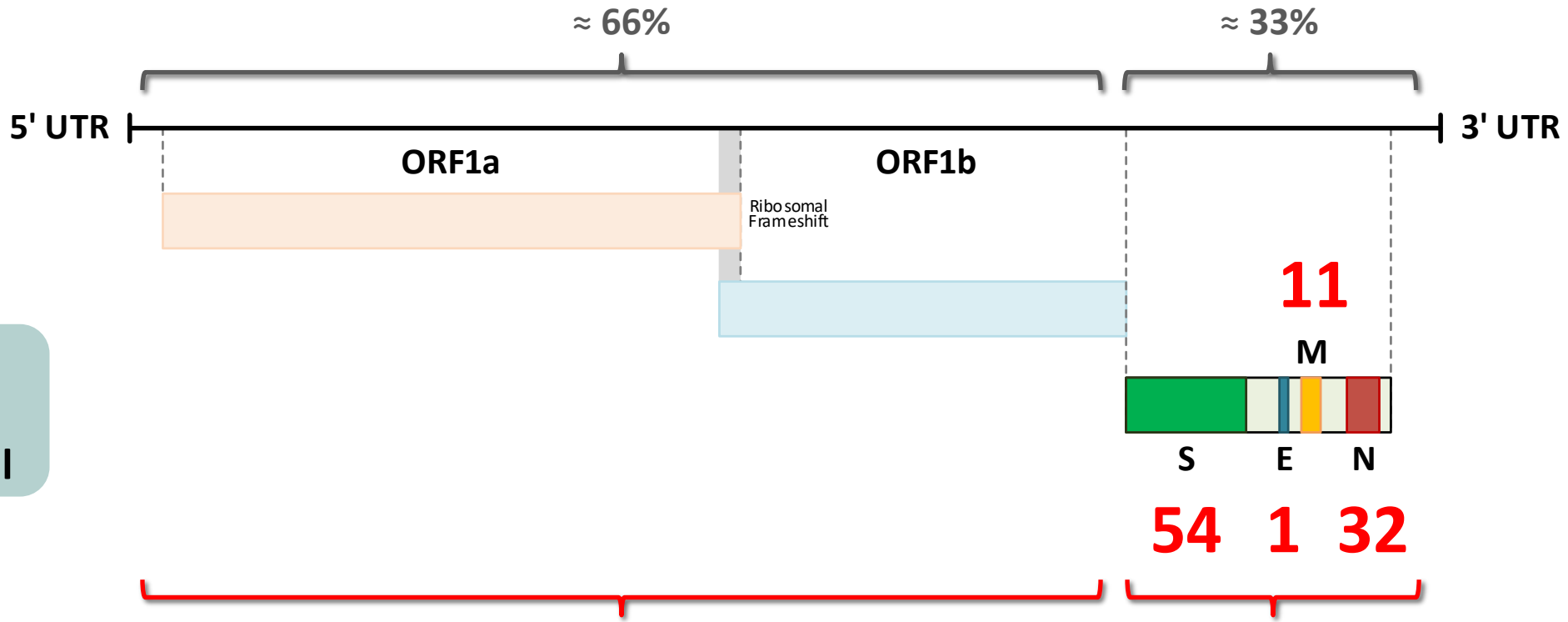
Protein		PSS-FUBAR	PSS-codeML	PSS Common
Structural	S	50 (35)	51 (37)	15 (13)
	M	12 (8)	13 (7)	4 (4)
	N	25 (18)	11 (9)	4 (4)
	E	1 (1)	NA	NA
Non-Structural	nsp1	8 (4)	4 (1)	2 (1)
	nsp2	15 (14)	2 (2)	NA
	nsp3	58 (52)	8 (6)	4 (3)
	nsp4	5 (5)	2 (2)	1 (1)
	nsp5	3 (3)	NA	NA
	nsp6	7 (7)	NA	NA
	nsp7	1 (1)	NA	NA
	nsp8	3 (3)	NA	NA
	nsp9	NA	NA	NA
	nsp10	1 (1)	NA	NA
	nsp12	9 (6)	2 (2)	1 (1)
	nsp13	1 (1)	1 (1)	NA
	nsp14	2 (2)	NA	NA
	nsp15	9 (8)	2 (2)	2 (2)
	nsp16	6 (5)	NA	NA
Accessory		44	5	2
TOTAL		260 (174)	101 (69)	35 (29)

29

- **PSS** identified with **both FUBAR and codeML** with **homologs** in SARS-CoV-2

(in structural and non-structural proteins)

199 PSS Distribution



199

PSS in Total

81 PSS
(≈ 40.7%)

98 PSS
(≈ 49.2%)

***20 PSS - Accessory**
(≈ 10.1%)

PSS on SARS-CoV-2 Structural Proteins

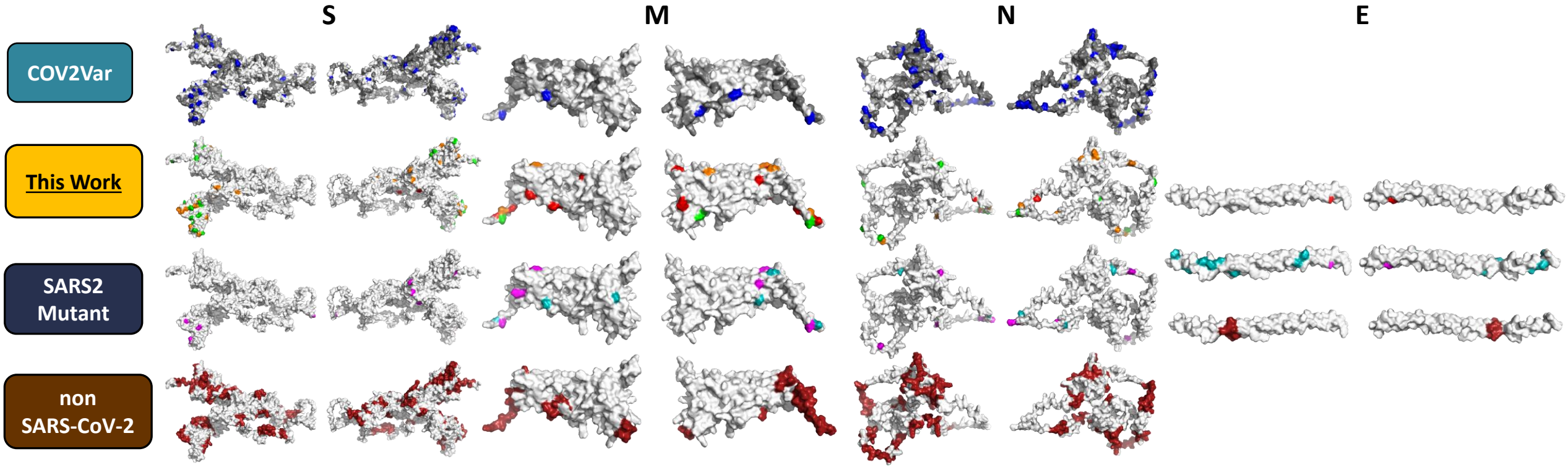


Figure 6 – PSS location on SARS-CoV-2 protein monomers. For each protein, from top to bottom: PSS present in the COV2Var-int-5% list (in blue are PSS, in gray are variable amino acid sites), PSS identified in this work (green – identified in this work and in the COV2Var-int-5% list; orange – identified in this work and in at least one other method in the COV2Var-int-5%; and red – only identified in this work), the top 10 variants (pink if it has a frequency over 5%, cyan otherwise) and regions identified as hot PSS regions in non-SARS-CoV-2 species (dark red).

Location of PSS on SARS-CoV-2 Spike

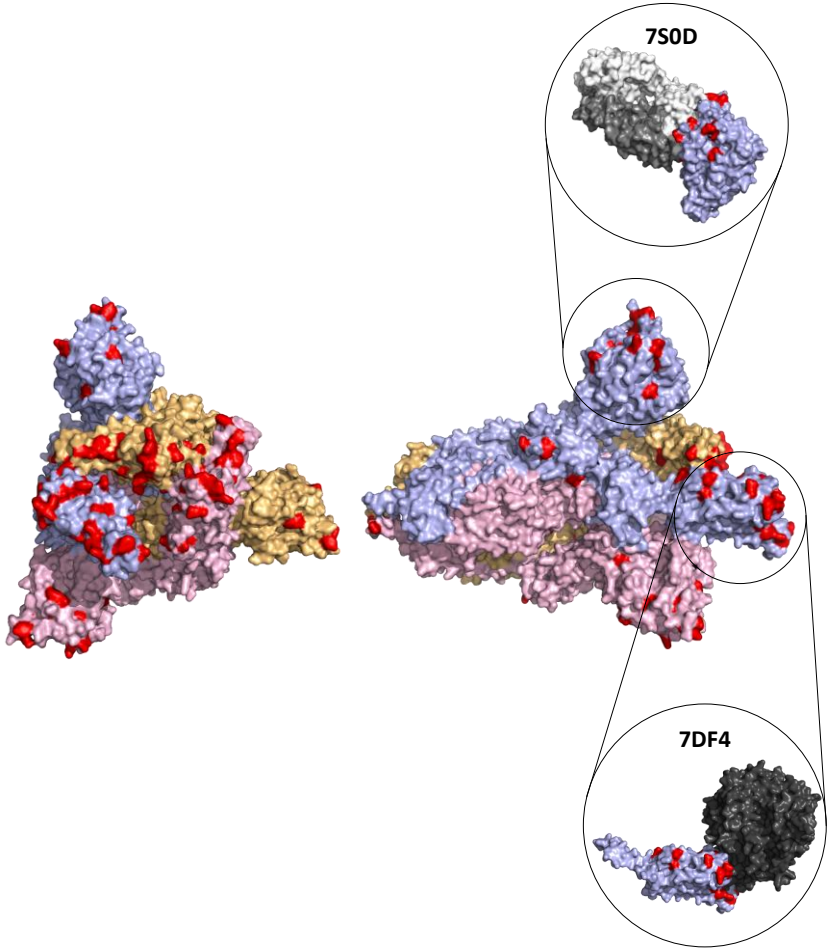


Figure 7 – Location of PSS (labeled in **red**) supported by more than one type of evidence in the SARS-CoV-2 S homotrimer protein structure (PDB accession number 7DF4). Each monomer is shown in a different color. The PDB accession numbers of the docking partners of the S protein are shown above the respective structure (PDB accession numbers 7S0D and 7DF4).

Prediction of PSS using Machine Learning Models

199 PSS

Can we predict more?

Prediction of **PSS** using **Machine Learning**

Objectives

- Identify possible **patterns shared** between the **PSS** to find out if **selective pressures are acting** on **SARS-CoV-2** in predictable ways.

- Identification of **new PSS in unseen data**, easing the understanding of **SARS-CoV-2 evolution**.

Machine Learning **Algorithms**

Table 3 – Machine learning methods and their respective algorithms applied in the prediction of PSS.

Machine Learning Method	Bayesian Networks	Support Vector Machines	Decision Trees	Classification Rules	Ensemble
Algorithms	BayesNet *naïve Bayes*	SMO	DecisionStump J48 NBTree RandomForest SimpleCart	DTNB OneR PART ZeroR	AdaBoostM1 Bagging MultiBoostAB

Validation of ML Methods

Table 4 – Types of results validation when applying machine learning methods.

Validation Types	Description	Pros (+) / Cons (-)
Repeated Random Sub-Sampling	Divides randomly the data set for training and validation	(-) A given sample of elements may never be selected, while another may be chosen several times
N fold Cross-Validation	Data is split into *N* subsets (folds) for a learning of *N* iterations. *N*-1 blocks are used, and only one for testing, which is different for each iteration	(+) Uses all the available data
Leave-One-Out	Similar method to *N* fold cross-validation, but sample size is only one element in the test set	
Hold-Out Percentage Split	Test set randomly chosen, usually around 20 to 30% of the elements. Remaining data subject to train and then validated in test set	

Prediction of **PSS** – **Amino Acid** Properties

Table 5 – List of amino acid properties used by Selbig *et. al.*, 1992.

PROPERTIES	
HYDROPHOBIC	SIDE CHAIN WITH <4 HEAVY ATOMS
POLAR	SIDE CHAIN WITH >4 HEAVY ATOMS
NEGATIV CHARGED	LINEAR SIDE CHAIN
POSITIV CHARGED	BULKY SIDE CHAIN
SMALL	SIDE CHAIN WITH O (OXYGEN)
TINY	SIDE CHAIN WITHOUT NH, OH, SH
ALIPHATIC	CHARGED SIDE CHAIN
AROMATIC	SIDE CHAIN WITH <3 CH

Extra Features

PROTEIN DOMAIN
STRUCTURE TYPE
SURFACE

Binary / Nominal Datasets

Table 6 – Part of Binary and Nominal Databases with features before the Feature Selection process.

FEATURE	DESCRIPTION
PSS_POSITION	Numerical Position
PROTEIN_DOMAIN	[0,1] ⇔ [No, Yes]
AMINOACID_ORIGIN	Amino acid (Nominal)
NUMBER_OF_TRANSFORMS	Numeral
HYDROPHOBIC	[0,1] ⇔ [No, Yes]
POLAR	[0,1] ⇔ [No, Yes]
NEGATIV_CHARGED	[0,1] ⇔ [No, Yes]
POSITIV_CHARGED	[0,1] ⇔ [No, Yes]
SMALL	[0,1] ⇔ [No, Yes]
TINY	[0,1] ⇔ [No, Yes]
ALIPHATIC	[0,1] ⇔ [No, Yes]
AROMATIC	[0,1] ⇔ [No, Yes]

Binary

FEATURE	DESCRIPTION
PSS_POSITION	Numerical Position
PROTEIN_DOMAIN* *(from Sup. Figs. S1 & S2 – Chapter IV)	[NTD, RBD, RBM, HR1, TM, RNA_Binding, Linker, Dimerization, C_Tail, N_Tail, EC, TM1, TM2, TM3, CTD, NA]
AMINOACID_ORIGIN	Amino acid (Nominal)
HYDROPHOBICITY	[Hydrophobic, Hydrophilic]
POLARITY	[Polar, NonPolar]
CHARGE	[Positive, Negative, NonCharged]
SIZE	[Big, Small, Tiny]
BENZENE_RING	[Aromatic, Aliphatic, None]

Nominal

Datasets Distribution

108

54+/54-

- **Spike**

22

11+/11-

- **Membrane**

64

32+/32-

- **Nucleocapsid**

2

1+/1-

- **Envelope**

199+ PSS

162

81+/81-

- **All_NSP**

40

20+/20-

- **All_ORF**

Distribution of instances per structural protein or groups of proteins. In **red** are all the **199 PSS** found in Ferreira, Soares *et al*. (2024) in SARS-CoV-2 and non-SARS-CoV-2.
All_NSP – All Non-Structural Proteins from ORF1ab;
All_ORF – All Accessory Proteins.

Feature Selection – Binary / Nominal DB

Table 7 – Part of **Feature Selection** (with **10x 10 fold cross-validation**) results for **Binary** and **Nominal Datasets** with ***CfsSubsetEval*** as atribute evaluator and ***Best First*** as search method.

Binary

Binary Dataset (w/o PSS_POSITION & NUMBER_OF_TRANSFORMS)		
	Attribute Selection	**Num Folds (%)**
All_SP	PROTEIN_DOMAIN	1 (10)
	AMINOACID_ORIGIN	10 (100)
	TINY	2 (20)
	SIDE_CHAIN_LESS_4_HEAVY_ATOMS	1 (10)
	SURFACE	10 (100)
All_NSP (w/ extra info)	AMINOACID_ORIGIN	10 (100)
	HYDROPHOBIC	10 (100)
	ALIPHATIC	6 (60)
	SURFACE	4 (40)
All_ORF (w/ extra info)	AMINOACID_ORIGIN	10 (100)
	NEGATIV_CHARGED	1 (10)
	TINY	2 (20)
	AROMATIC	9 (90)
	SIDE_CHAIN_MORE_4_HEAVY_ATOMS	2 (20)
	STRUCTURED	1 (10)

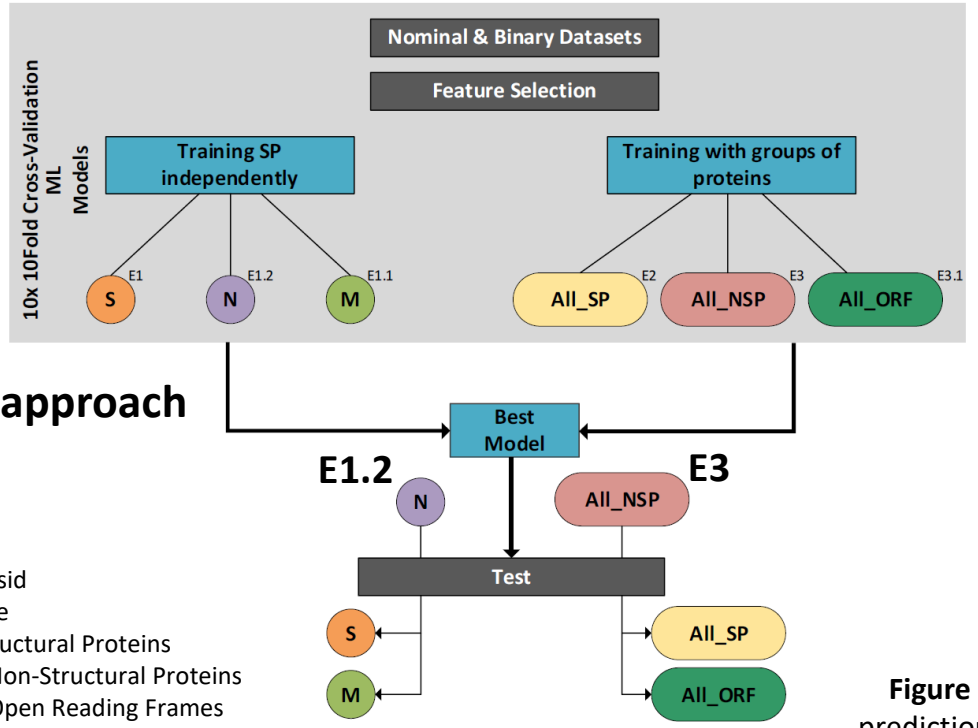
Nominal

Nominal Dataset (w/o PSS_POSITION)		
	Attribute Selection	**Num Folds (%)**
All_SP	PROTEIN_DOMAIN	10 (100)
	AMINOACID_ORIGIN	9 (90)
	SURFACE	8 (80)
All_NSP (w/ extra info)	AMINOACID_ORIGIN	10 (100)
	HYDROPHOBICITY	10 (100)
	BENZENE_RING	8 (80)
	SURFACE	6 (60)
All_ORF (w/ extra info)	AMINOACID_ORIGIN	10 (100)
	CHARGE	1 (10)
	SIDE_CHAIN_WITH_HEAVY_ATOMS	1 (10)

10x10 fold CV

Prediction of **PSS** using **Machine Learning**

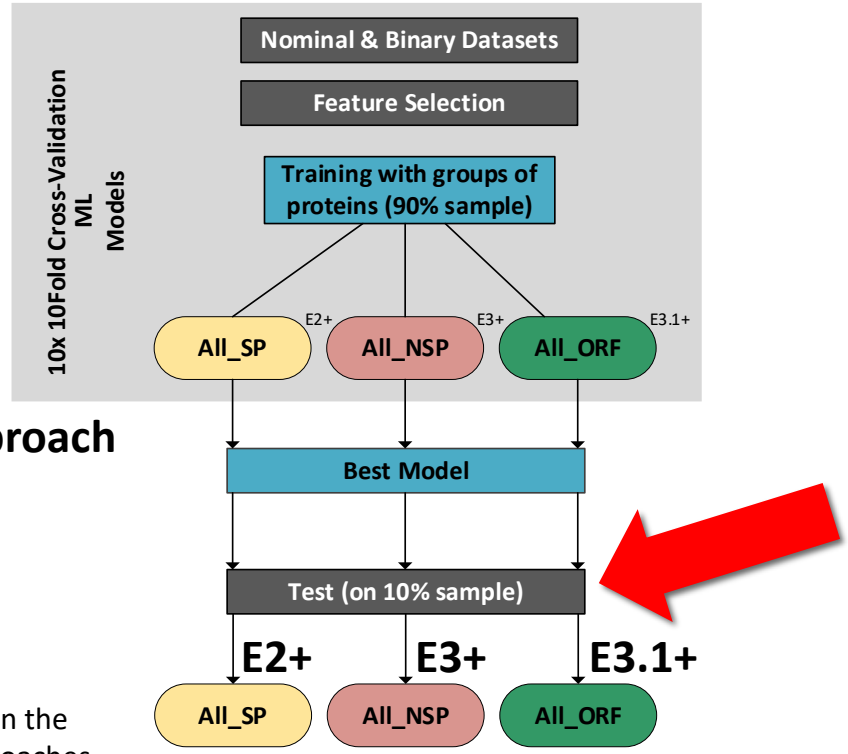
- Using **all available data** to produce a model that could predict PSS, genome wide



❖ **1st approach**

S – Spike
N – Nucleocapsid
M – Membrane
All_SP – All Structural Proteins
All_NSP – All Non-Structural Proteins
All_ORF – All Open Reading Frames
ML – Machine Learning

- More conserved in the percentage of data used for training, using **90% for training and 10% for testing**

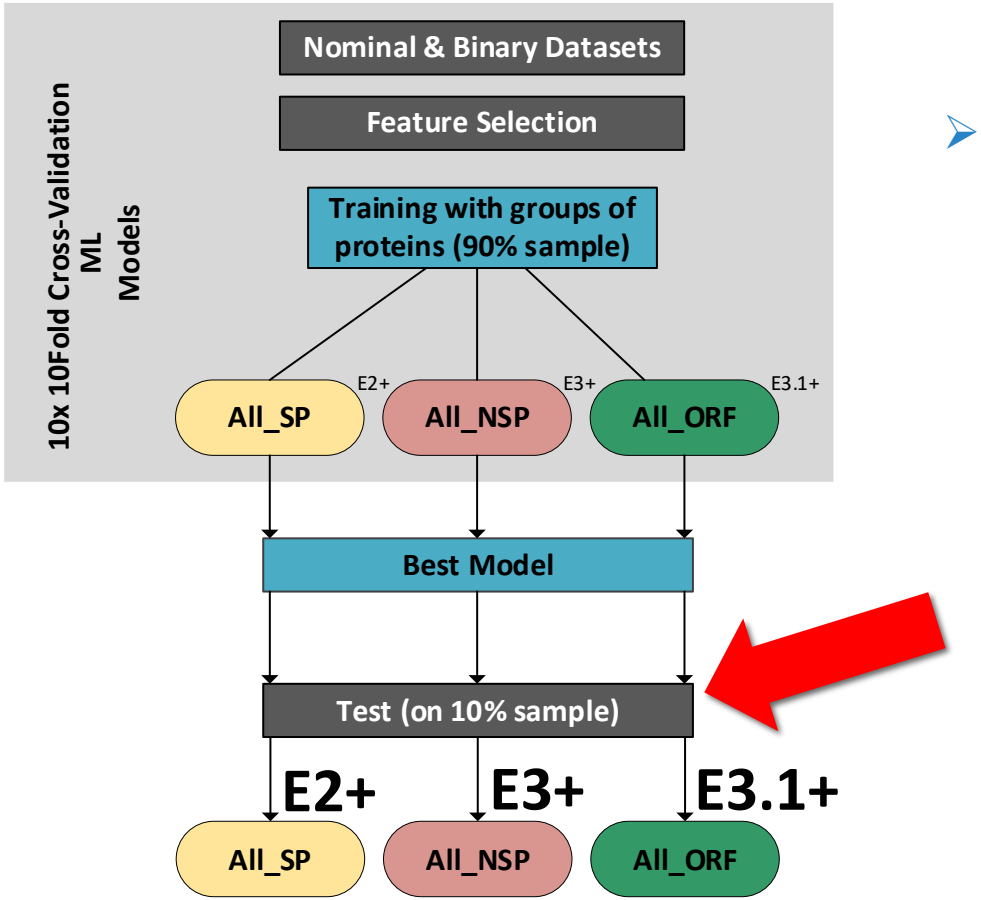


❖ **2nd approach**

Figure 8 – Steps and datasets involved in the prediction of PSS using two different approaches.

Prediction of **PSS** – 90% **Train** / 10% **Test**

- **2nd approach**



- More conserved in the percentage of data used for training, using **90% for training and 10% for testing**

Figure 9 – Steps and datasets involved in the prediction of PSS using 90% of the datasets as training sets and 10% as test sets.

All_SP – All Structural Proteins
All_NSP – All Non-Structural Proteins
All_ORF – All Open Reading Frames
ML – Machine Learning

U.PORTO

ICBAS | INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
SCHOOL OF MEDICINE AND BIOMEDICAL SCIENCES

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Nominal

Results Nominal 10% DB

Table 8 – Test Results on 10% Samples in Groups of Proteins (GP) for Nominal Datasets.

Nominal Datasets – 10% Samples							
TEST		Metrics					
		CCI	F-Measure	Kappa Statistic	Precision	TPR	AUROC
E2+ (BayesNet K2)	All_SP_10%	60.00	0.60	0.20	0.60	0.60	0.61
E3+ (DTNB)	All_NSP_10%	70.00	0.70	0.40	0.71	0.70	0.74
E3+ (AdaBoost M1)	All_NSP_10%	80.00	0.79	0.60	0.86	0.80	0.84
E3.1+ (Decision Stump)	All_ORF_10%	50.00	0.33	0.00	0.25	0.50	0.50

Test

Best in unseen data

Results Binary **10%** DB

Binary

Test

Table 9 – Test Results on 10% Samples in Groups of Proteins (GP) for Binary Datasets.

Binary Datasets – 10% Samples							
		Metrics					
TEST		**CCI**	**F-Measure**	**Kappa Statistic**	**Precision**	**TPR**	**AUROC**
E2+ (BayesNet TAN)	All_SP_10%	70.00	0.70	0.40	0.71	0.70	0.70
E3+ (*naïve Bayes*)	All_NSP_10%	90.00	0.90	0.80	0.92	0.90	0.96
E3.1+ (*naïve Bayes*)	All_ORF_10%	50.00	0.33	0.00	0.25	0.50	0.50

Best in unseen data

General Discussion

General **Discussion**

Auto-PSS-Genome

docker

COMPUTATIONAL PIPELINES ComPS

Infer **Positively Selected Amino Acid Sites** (**PSS**) in **Coronavirus** Species

- ✓ Auto-PSS-Genome allowed for this analysis with overlap of **FUBAR** and **codeML** results.
- ❖ None of the attempts to identify **PSS** in SARS-CoV-2 used cross information from SARS-CoV-2 and a high array of coronaviruses species sequences, unlike us.

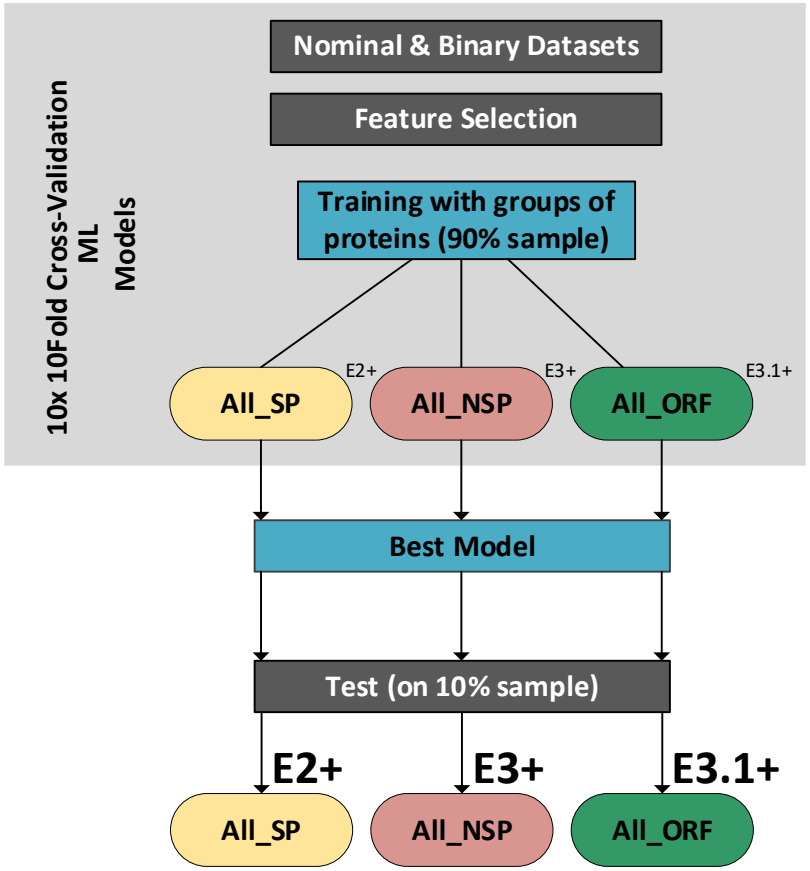
General **Discussion**

- **Spike** → Responsible for cell tropism and host range → Mediates the fusion of the host's and viral's membranes
  - **54 PSS**
- **Nucleocapsid** → Presents a domain involved in the inhibition of IFN-β → **Main target of Positive Selection**
  - **32 PSS**

Highest normalized proportions

Signalling protein that is released in virus-infected cells causing nearby cells to strengthen their antiviral defenses

General Discussion – PSS with ML



- **2nd approach**:
  Training only with Group of Proteins.
  Using samples of 10% as Test Sets

- **Methodology**:
  Using 90% for Training, while 10% for Testing

- **Results**:
  More Robust in the Test Sets
  (CCI ≈ 70%-90% ; FM ≈ 0.70-0.90)

Best in unseen data

- **Conclusions**:
  Accuracy increased due to presence of 10% of data of each protein for Training

General Discussion – PSS with ML

LIMITATIONS

- **Lack of uniform distribution of PSS** makes it impossible to train models and test them using data from every single protein individually and, **if proteins are grouped**, the **models** can become **skewed**;
- The **true negative PSS** sites were **picked** at random from a **pool of sites** that were **deemed as "conserved"**, even if by chance, the **sample** can also be **skewed**;

POSSIBLE SOLUTIONS

- Data from all **PSS identified in COV2Var should be used**, even though they are not well supported **PSS**;
- Regarding the **true negatives**, **multiple samples should be made** to guarantee **minimal biases** in the **training** of models.

Conclusions and Future Perspectives

Conclusions

- Development of a multitude of **Bioinformatic tools**, **available to the community** in a more accessible way

pegi3s Bioinformatics Docker Images Project

- Development of **Auto-PSS-Genome** pipeline
  - Can be used to **analyze positive selection** in a variety of datasets, not limited to bacterial or viral.

docker

COMPUTATIONAL PIPELINES

- We present an **approach to identify PSS in Coronavirus**
  - Identification of **199 PSS** in **SARS-CoV-2 supported** by **multiple lines of evidence**;
  - Those, in principle, that **contribute to the increased transmissibility** and/or host immune system escape, **being of interest** in the study and **prediction of SARS-CoV-2** behavior and **evolution**.

- **Machine Learning models** for the **prediction of PSS**
  - Best ones having **70-90% accuracy**.

Future Perspectives

- Further **development** of **machine learning methods** by addressing the limitations pointed out

Inductive Logic Programming

Interpretable Rules

PATTERNS IN PSS SARS-CoV-2?

YES OR NO?

- **Development** of a **tool** able to **make predictions** of **PSS** in new datasets

USEFUL WHEN A NEW MUTATION / CORONAVIRUS THREAT ARISES

THUS TRACING CORONAVIRUS EVOLUTION, ASSESSING RISK OF FUTURE TRANSMISSION EVENTS

List of **Publications**

1) **Ferreira, P.**, Soares, R., López-Fernández, H., Vazquez, N., Reboiro-Jato, M., Vieira, C.P., Vieira, J.: **Multiple Lines of Evidence Support 199 SARS-CoV-2 Positively Selected Amino Acid Sites**. In: International Journal of Molecular Sciences, 2024.
DOI:10.3390/ijms25042428

2) López-Fernández, H., Vieira, C.P., **Ferreira, P.**, Gouveia, P., Fdez-Riverola, F., Reboiro-Jato, M., Vieira, J.: **On the Identification of Clinically Relevant Bacterial Amino Acid Changes at the Whole Genome Level Using Auto-PSS-Genome**. In: Interdisciplinary Sciences: Computational Life Sciences, 2021.
DOI:10.1007/s12539-021-00439-2

3) López-Fernández, H., **Ferreira, P.**, Reboiro-Jato, M., Vieira, C.P., Vieira, J.: **The pegi3s Bioinformatics Docker Images Project**. In: Proceedings of 15th International Conference on Practical Applications of Computational Biology and Bioinformatics, 2021.
DOI:10.1007/978-3-030-86258-9_4

Acknowledgments / Funding

- Cristina Vieira, PhD
- Hugo López-Fernández, PhD
- Jorge Vieira, PhD
- Ricardo Soares, MSc
- André Sousa, MSc
- *Phenotypic Evolution Group* – i3S
- *SING Research Group* – UVigo
- *Yeast Signalling Networks Group* – i3S

U.PORTO

ICBAS | INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
SCHOOL OF MEDICINE AND BIOMEDICAL SCIENCES

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Thank you all!

Appendix

U.PORTO

ICBAS | INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
SCHOOL OF MEDICINE AND BIOMEDICAL SCIENCES

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Nominal

Results Nominal DB

Table ? – Training Results for Nominal Datasets.

Nominal Datasets w/ Feature Selection								
		Metrics						
TRAIN	**Algorithm**	**CCI**	**F-Measure**	**Kappa Statistic**	**Precision**	**TPR**	**TNR**	**AUROC**
Spike	J48	69.34 (11.79) v	0.61 (0.19)	0.38 (0.24) v	0.80 (0.21) v	0.52 (0.22)	0.86 (0.13) v	0.72 (0.13) v
	DTNB	68.10 (11.67) v	0.63 (0.15) v	0.36 (0.23) v	0.77 (0.19) v	0.56 (0.18)	0.80 (0.18) v	0.72 (0.13) v
Membrane	DecisionStump	70.67 (26.92) v	0.46 (0.48)	0.40 (0.54) v	0.48 (0.50)	0.46 (0.49) *	0.95 (0.22) v	0.71 (0.27) v
	BayesNet (K2)	68.33 (29.73) v	0.56 (0.44)	0.36 (0.60)	0.54 (0.45)	0.62 (0.48)	0.75 (0.42) v	0.77 (0.38) v
Nucleocapsid	NBTree	82.36 (13.73) v	0.83 (0.14) v	0.65 (0.28) v	0.84 (0.17) v	0.85 (0.19)	0.80 (0.23) v	0.85 (0.17) v
	BayesNet (TAN)	81.74 (14.23) v	0.82 (0.16) v	0.63 (0.29) v	0.81 (0.19) v	0.87 (0.20)	0.76 (0.24) v	0.84 (0.18) v
All_SP	BayesNet (K2)	68.71 (9.85) v	0.67 (0.12)	0.37 (0.20) v	0.70 (0.11) v	0.66 (0.15)	0.71 (0.14) v	0.75 (0.10) v
All_NSP	DTNB	71.00 (13.30) v	0.69 (0.16) v	0.42 (0.27) v	0.73 (0.18) v	0.70 (0.21)	0.72 (0.20)	0.74 (0.14) v
	AdaBoostM1	70.33 (12.84) v	0.69 (0.16) v	0.40 (0.26) v	0.72 (0.18) v	0.71 (0.21)	0.69 (0.22)	0.71 (0.14) v
All_ORF	DecisionStump	61.25 (24.08) v	0.70 (0.20) v	0.25 (0.43)	0.59 (0.25) v	0.94 (0.20)	0.34 (0.40)	0.64 (0.22)

Train

Best Models

Binary

Results Binary DB

Table ? – Training Results for Binary Datasets.

Binary Datasets w/ Feature Selection								
		Metrics						
TRAIN	**Algorithm**	**CCI**	**F-Measure**	**Kappa Statistic**	**Precision**	**TPR**	**TNR**	**AUROC**
Spike	BayesNet (K2)	56.30 (14.81)	0.49 (0.21)	0.12 (0.30)	0.58 (0.23) v	0.46 (0.24)	0.67 (0.20)	0.59 (0.17)
Membrane	J48	71.00 (27.07) v	0.46 (0.49)	0.41 (0.54) v	0.41 (0.54) v	0.46 (0.49) *	0.96 (0.20) v	0.71 (0.27) v
	naïve Bayes	69.33 (30.40) v	0.58 (0.44)	0.38 (0.61)	0.38 (0.61)	0.64 (0.47)	0.75 (0.42) v	0.84 (0.31) v
	DecisionStump	70.67 (26.92) v	0.46 (0.48)	0.40 (0.54) v	0.40 (0.54) v	0.46 (0.49) *	0.95 (0.22) v	0.71 (0.27) v
Nucleocapsid	SimpleCart	78.10 (14.25) v	0.80 (0.14)	0.56 (0.29)	0.75 (0.16)	0.90 (0.18)	0.66 (0.24)	0.76 (0.16)
	RandomForest	76.57 (15.31) v	0.77 (0.16) v	0.53 (0.31) v	0.76 (0.18) v	0.84 (0.21)	0.69 (0.24) v	0.83 (0.16) v
All_SP	BayesNet (TAN)	65.71 (10.08) v	0.66 (0.11)	0.31 (0.20) v	0.66 (0.11) v	0.67 (0.14)	0.65 (0.16) v	0.68 (0.11) v
All_NSP	*naïve Bayes*	74.11 (12.34) v	0.73 (0.15) v	0.48 (0.25) v	0.75 (0.16) v	0.75 (0.21)	0.73 (0.18)	0.79 (0.15) v
All_ORF	*naïve Bayes*	53.50 (23.81)	0.53 (0.30)	0.08 (0.45)	0.50 (0.33)	0.66 (0.39)	0.44 (0.43)	0.57 (0.38)

Train

Best Models