Supplementary Tables

	b-phipsi	w-rdist	t-alpha
b-phipsi	1.000000	0.118785	0.075813
w-rdist	0.118785	1.000000	0.089612
t-alpha	0.075813	0.089612	1.000000

Table 1: Correlation matrix of b-phipsi, w-rdist and t-alpha based on whole structure comparisons between SARS-CoV-2 Spike (native) and the proteins in the paper's target dataset. We see that the metrics are segregated and therefore there is no redundancy of information.

Hydrophobicity group	Aminoacids
0	Ile, Val
1	Leu
2	Phe,Cys,Met,Ala
3	Gly,Thr,Ser,Trp,Tyr,Pro
4	His,Asn,Gln,Asp,Glu,Lys
5	Arg

Table 2: Hydrophobicity groups based on Hydropathy Index, which are encoded into the mixed sequence alignments in the method's constrained mode search.

Site Index	Residue positions
0	1, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 53, 195, 197, 198, 200, 202, 203, 204, 225, 227, 228, 274, 276, 281, 291, 298, 301, 302, 303, 304, 305, 312, 313, 314, 315, 316, 317, 319, 320, 321, 322, 355, 378, 379, 380, 381, 382, 383, 384, 386, 393, 396, 412, 426, 427, 428, 429, 430, 464, 514, 515, 516, 517, 518, 520, 540, 547, 548, 549, 550, 567, 568, 569, 570, 571, 574, 589, 590, 592, 596, 613, 735, 736, 737, 738, 739, 740, 743, 744, 745, 746, 747, 748, 750, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 773, 854, 855, 856, 858, 951, 954, 957, 958, 959, 960, 961, 962, 963, 964, 965, 967, 968, 969, 970, 971, 972, 973, 978, 981, 982, 983, 984, 985, 986, 987, 988, 989, 990, 991, 992, 993, 994, 995, 998, 999, 1001, 1002, 1003, 1005, 1006, 1007, 1008, 1009, 1010, 1012, 1013, 1014, 1015, 1016, 1017, 1019
1	37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 49, 50, 53, 195, 197, 198, 200, 202, 203, 204, 225, 226, 227, 228, 230, 274, 291, 298, 301, 302, 303, 304, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 355, 357, 380, 381, 382, 390, 391, 392, 393, 394, 396, 412, 426, 427, 428, 429, 430, 464, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 541, 543, 544, 545, 546, 547, 548, 549, 550, 557, 565, 567, 568, 569, 571, 574, 589, 590, 591, 592, 595, 596, 611, 612, 613, 614, 646, 647, 661, 662, 665, 666, 667, 697, 698, 699, 700, 733, 735, 736, 737, 738, 739, 740, 743, 744, 745, 746, 747, 750, 754, 755, 757, 758, 761, 762, 763, 764, 765, 767, 768, 771, 772, 775, 776, 779, 783, 785, 786, 854, 861, 862, 864, 865, 873, 957, 960, 961, 964, 965, 967, 968, 969, 973, 974, 975, 976, 978, 979, 982, 983, 1004

2	1, 725, 727, 728, 769, 770, 772, 773, 776, 777, 780, 781, 784, 785, 786, 789, 888, 889, 891, 947, 950, 951, 954, 1012, 1013, 1014, 1015, 1016, 1017, 1018, 1019, 1020, 1021, 1022, 1023, 1024, 1026, 1027, 1028, 1030, 1031, 1034, 1039, 1040, 1041, 1042, 1043, 1044, 1045, 1064
3	86, 87, 88, 89, 107, 108, 109, 113, 114, 115, 130, 132, 167, 168, 195, 196, 197, 198, 199, 200, 229, 230, 231, 232, 233, 234, 235, 236, 353, 354, 355, 356, 357, 396, 454, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 486, 487, 488, 489, 490, 491, 492
4	41, 197, 198, 200, 202, 228, 355, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 390, 391, 392, 393, 396, 403, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 421, 424, 426, 427, 428, 429, 430, 433, 436, 437, 439, 440, 464, 505, 514, 515, 516, 517, 518, 519, 520, 522, 544, 545, 546, 547, 565, 567, 571, 973, 974, 976, 978, 979, 982, 983, 984, 985, 986, 987, 988, 991, 992

Table 3: Predicted binding sites by Schrödinger SiteMap for SARS-CoV-2 Spike protein (native). The predictions derive from PDB 6VXX.A that was prepared with Protein Preparation Wizard of Schrödinger Maestro Suite.

Supplementary Figures

a

Structural Comparison Report for 6VXX_A - whole structures (total: 58)

1

- **Protein name:** Spike glycoprotein
- **Organism:** Severe acute respiratory syndrome coronavirus
- **Uniprot Accession Number:** P59594
- **Protein sequence length:** 1255 aa
- **1D identity (%):** 76.35
- **1D identity (%) [Gaps excluded]:** 77.94
- **1D identity - Alignment Gaps:** 26
- **Common reported functions (%):** 100.0
- **Common reported locations (%):** 62.5
- **Common reported processes (%):** 90.0

- **PDB ID:** 6NB6
- **Chain:** A
- **Crystallized protein length:** 1052 aa
- **Resolution:** 4.2 Å
- **b-phipsi:** 0.001058
- **w-rdist:** 0.5121
- **t-alpha:** 0.002481
- **Chemical similarity (Tanimoto Index) (%):** 94.56
- **1D identity (%) [PDB]:** 69.31
- **1D identity (%) [Gaps excluded][PDB]:** 78.58
- **1D identity - Alignment Gaps [PDB]:** 128
- **2D identity (%) [PDB]:** 72.68
- **2D identity (%) [Gaps excluded][PDB]:** 86.39
- **2D identity - Alignment Gaps [PDB]:** 176
- **3D similarity (TM-Score) (%) [PDB]:** 96.46

- **Gene name:** S
- **RefSeq ID:** NC_004718
- **Genomic sequence length:** 29751
- **5-UTR|CDS|3-UTR identity (%):** 88.52 | 73.15 | 22.38
- **5-UTR|CDS|3-UTR identity (%) [Gaps excluded]:** 92.28 | 78.79 | 98.18
- **5-UTR|CDS|3-UTR identity [Alignment Gaps]:** 11 | 282 | 745

Uniprot Description:

Spike glycoprotein May down-regulate host tetherin (BST2) by lysosomal degradation, thereby counteracting its antiviral activity.

Homotrimer; each monomer consists of a S1 and a S2 subunit. The resulting peplomers protrude from the virus surface as spikes (By similarity). Binds to human and palm civet ACE2 and human CLEC4M/DC-SIGNR. Interacts with the accessory proteins 3a and 7a.

Gene Ontology Information:

Molecular Function

- host cell surface receptor binding
- identical protein binding

Location

- host cell endoplasmic reticulum-Golgi intermediate compartment membrane
- host cell plasma membrane
- integral component of membrane
- viral envelope
- virion membrane

Biological process

- endocytosis involved in viral entry into host cell
- fusion of virus membrane with host endosome membrane
- fusion of virus membrane with host plasma membrane
- pathogenesis
- receptor-mediated virion attachment to host cell
- suppression by virus of host tetherin activity
- suppression by virus of host type I interferon-mediated signaling pathway
- viral protein processing
- viral translation

b

Structural Comparison Report for 6VXX_A - whole structures (total: 58)

1

Protein name: Spike glycoprotein **Organism:** Severe acute respiratory syndrome coronavirus **Uniprot Accession Number:** P59594 **Protein sequence length:** 1255 aa **1D identity (%):** 76.35 **1D identity (%) [Gaps excluded]:** 77.94 **1D identity - Alignment Gaps:** 26 **Common reported functions (%):** 100.0 **Common reported locations (%):** 62.5 **Common reported processes (%):** 90.0

PDB ID: 6NB6 **Chain:** A **Crystallized protein length:** 1052 aa **Resolution:** 4.2 Å **b-phipsi:** 0.001058 **w-rdist:** 0.5121 **t-alpha:** 0.002481 **Chemical similarity (Tanimoto Index) (%):** 94.56 **1D identity (%) [PDB]:** 69.31 **1D identity (%) [Gaps excluded][PDB]:** 78.58 **1D identity - Alignment Gaps [PDB]:** 128 **2D identity (%) [PDB]:** 72.68 **2D identity (%) [Gaps excluded][PDB]:** 86.39 **2D identity - Alignment Gaps [PDB]:** 176 **3D similarity (TM-Score) (%) [PDB]:** 96.46

Gene name: S **RefSeq ID:** NC_004718 **Genomic sequence length:** 29751 **5-UTR|CDS|3-UTR identity (%):** 88.52 | 73.15 | 22.38 **5-UTR|CDS|3-UTR identity (%) [Gaps excluded]:** 92.28 | 78.79 | 98.18 **5-UTR|CDS|3-UTR identity [Alignment Gaps]:** 11 | 282 | 745

Figure 1: Sample final report of Machaon in printer-friendly, readable formats. This is a formatted output for a protein profile built by Machaon in Hypertext Markup Language (HMTL) and Cascading Style Sheets (CSS). Figure **a** is the browser display format and **b** is the compact printed PDF view format.

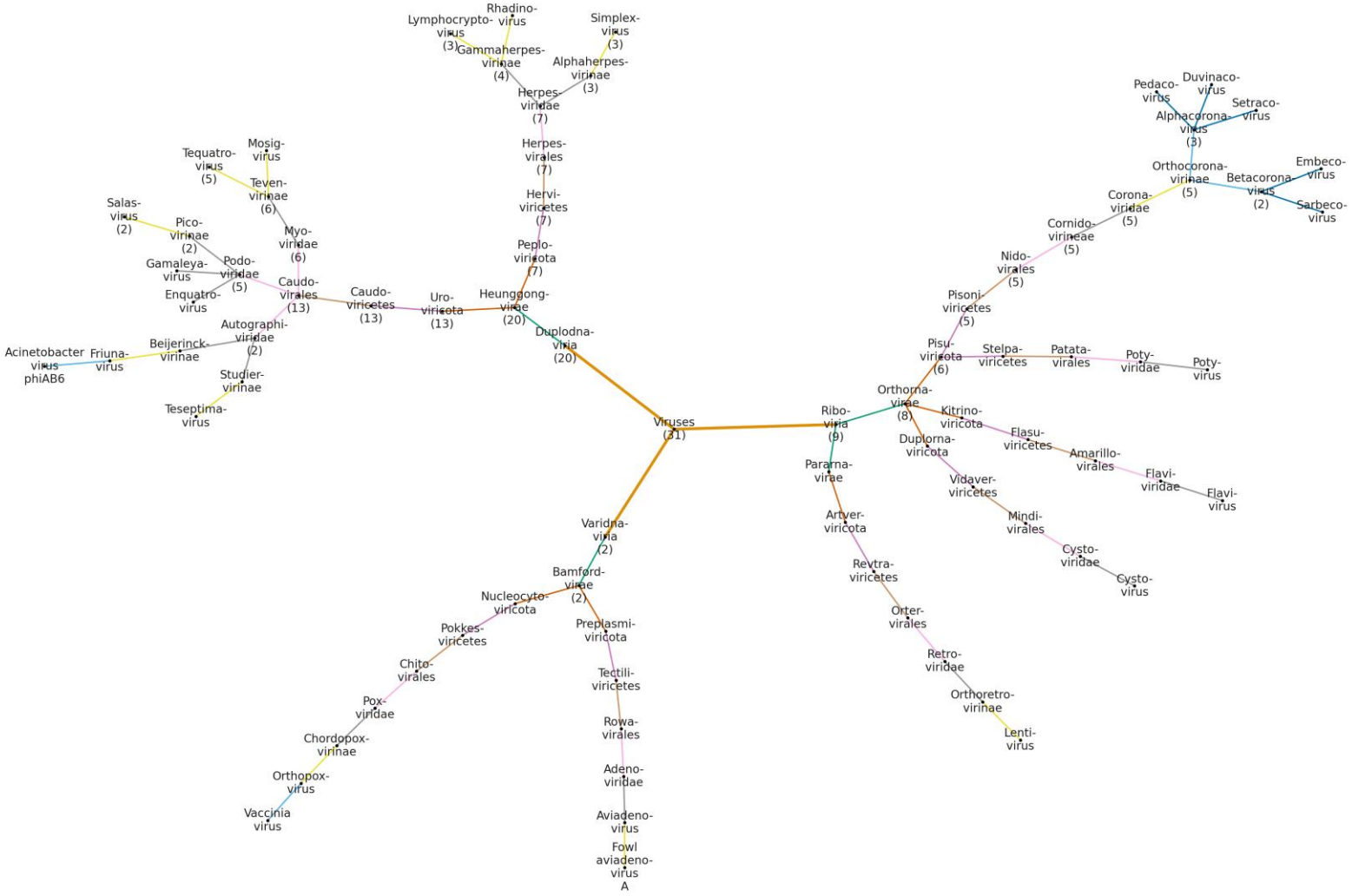
Viruses (31)
Duplodnaviria (20)
Heunggongvirae (20)
Uroviricota (13)
Caudoviricetes (13)
Caudovirales (13)
Myoviridae (6)
Podoviridae (5)
Autographiviridae (2)
Teseptimavirus
Studiervirinae
Beijerinckvirinae
Acinetobacter virus phiAB6
Friunavirus
Enquatrovirus
Gamaleyavirus
Pico-virinae (2)
Salasvirus (2)
Tevenvirinae (6)
Tequatrovirus (5)
Mosigvirus
Peploviricota (7)
Herviviricetes (7)
Herpesvirales (7)
Herpesviridae (7)
Gammaherpesvirinae (4)
Alphaherpesvirinae (3)
Lymphocryptovirus (3)
Rhadinovirus
Simplexvirus (3)
Varidnaviria (2)
Bamfordvirae (2)
Nucleocytoviricota
Pokkesviricetes
Chitovirales
Poxviridae
Chordopoxvirinae
Orthopoxvirus
Vaccinia virus
Preplasmiviricota
Tectiliviricetes
Rowavirales
Adenoviridae
Aviadenovirus
Fowl aviadenovirus A
Riboviria (9)
Pararnavirae
Artverviricota
Revtraviricetes
Ortervirales
Retroviridae
Orthoretrovirinae
Lentivirus
Orthornavirae (8)
Pisuviricota (6)
Pisoniviricetes (5)
Nidovirales (5)
Cornidovirineae (5)
Coronaviridae (5)
Orthocoronavirinae (5)
Alphacoronavirus (3)
Pedacovirus
Duvinacovirus
Setracovirus
Betacoronavirus (2)
Embecovirus
Sarbecovirus
Stelpaviricetes
Patatavirales
Potyviridae
Potyvirus
Kitrinoviricota
Duplornaviricota
Flasuviricetes
Amarillovirales
Flaviviridae
Flavivirus
Vidaverviricetes
Mindivirales
Cystoviridae
Cystovirus

Figure 2: The lineage tree of the viral proteins in the final set of the whole structural comparisons with Delta variant's Spike. These trees are generated by Machaon's presentation module. Each family name carries also a population number if there are more than one protein categorized under it. The root of the tree starts with thicker branches and the colors designated the branch levels. The tree refers to the viral proteins that were found to be the most structurally relevant. These trees could also be considered as distant evolutionary trees according to the structural traits of the proteins in the results. The lineage information is retrieved from UniProt. (Zoom to review)

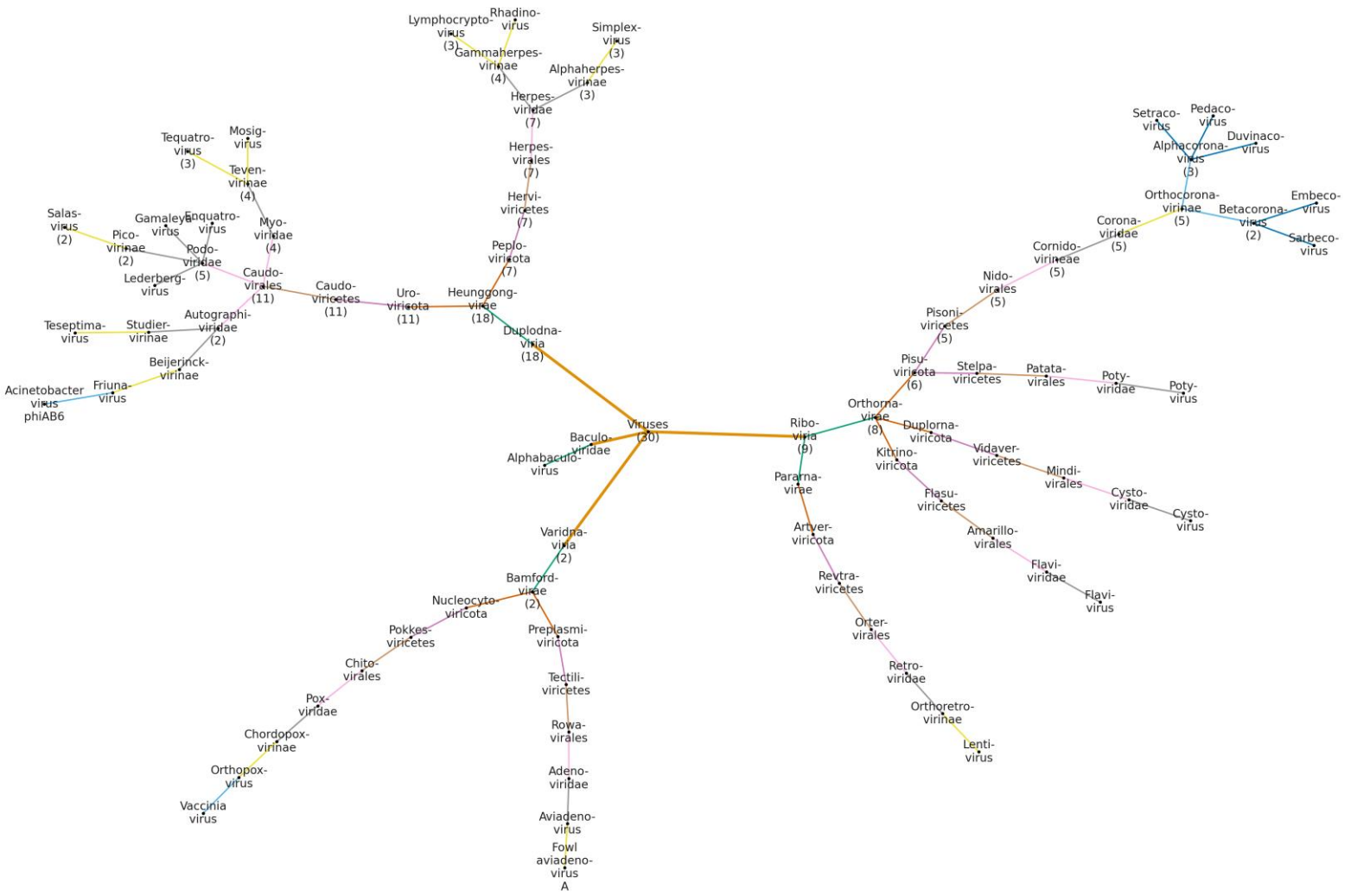
Viruses (30)
Duplodnaviria (18)
Heunggongvirae (18)
Uroviricota (11)
Caudoviricetes (11)
Caudovirales (11)
Myoviridae (4)
Podoviridae (5)
Autographiviridae (2)
Beijerinckvirinae
Tevenvirinae (4)
Picovirinae (2)
Studiervirinae
Peploviricota (7)
Herviviricetes (7)
Herpesvirales (7)
Herpesviridae (7)
Gammaherpesvirinae (4)
Alphaherpesvirinae (3)
Lymphocryptovirus (3)
Rhadinovirus
Simplexvirus (3)
Mosigvirus
Tequatrovirus (3)
Salasvirus (2)
Gamaleyavirus
Enquatrovirus
Lederbergvirus
Teseptimavirus
Acinetobacter virus phiAB6
Friunavirus
Varidnaviria (2)
Bamfordvirae (2)
Nucleocytoviricota
Pokkesviricetes
Chitovirales
Poxviridae
Chordopoxvirinae
Orthopoxvirus
Vaccinia virus
Preplasmiviricota
Tectiliviricetes
Rowavirales
Adenoviridae
Aviadenovirus
Fowl aviadenovirus A
Baculoviridae
Alphabaculovirus
Riboviria (9)
Orthornavirae (8)
Pisuviricota (6)
Pisoniviricetes (5)
Nidovirales (5)
Cornidovirineae (5)
Coronaviridae (5)
Orthocoronavirinae (5)
Alphacoronavirus (3)
Setracovirus
Pedacovirus
Duvinacovirus
Betacoronavirus (2)
Embecovirus
Sarbecovirus
Stelpaviricetes
Patatavirales
Potyviridae
Potyvirus
Kitrinoviricota
Vidaveriviricetes
Mindivirales
Cystoviridae
Cystovirus
Flasuviricetes
Amarillovirales
Flaviviridae
Flavivirus
Duplornaviricota
Pararnavirae
Artverviricota
Revtraviricetes
Ortervirales
Retroviridae
Orthoretrovirinae
Lentivirus

Figure 3: The lineage tree of the viral proteins in the final set of the whole structural comparisons with Omicron variant's Spike. These trees are generated by Machaon's presentation module. Each family name carries also a population number if there are more than one protein categorized under it. The root of the tree starts with thicker branches and the colors designated the branch levels. The tree refers to the viral proteins that were found most structurally relevant. These trees could also be considered as distant evolutionary trees according to the structural traits of the proteins in the results. The lineage information is retrieved from UniProt. (Zoom to review)

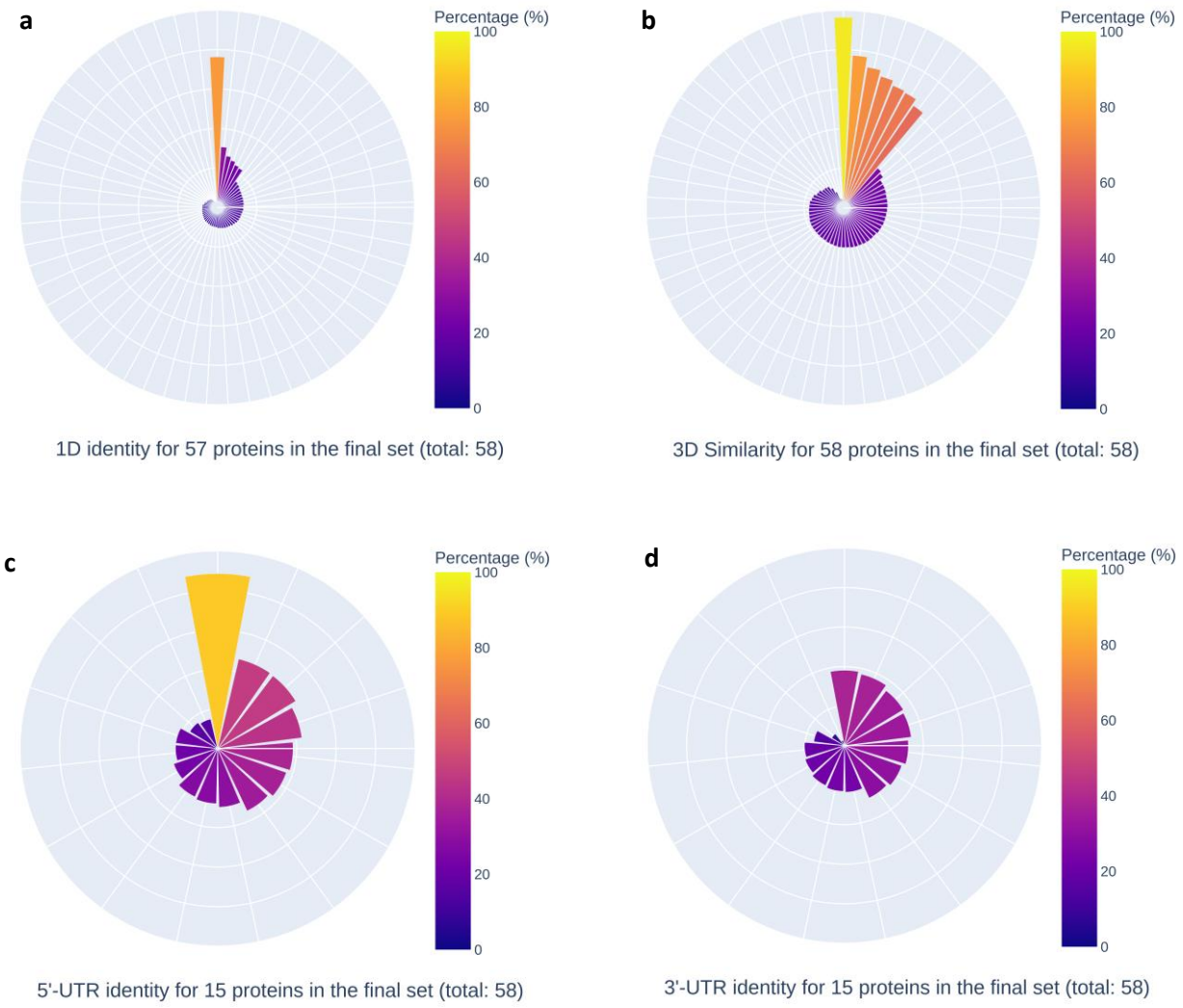

Figure 4: Evaluation of whole structure comparison results with SARS-CoV-2 Spike protein (native). Each protruding bar from the center of each radial plot represents a protein entry from the final set yielded by Machaon. The plots are generated by the presentation module. The measurements were carried out by the evaluation module wherever the underlying data allowed it (missing or malformed relevant data). **a)** Protein 1D sequence identities **b)** 3D similarities (TM-Score) **c)** 5'-end Untranslated Region identities **d)** 3'-end Untranslated Region sequence identities.

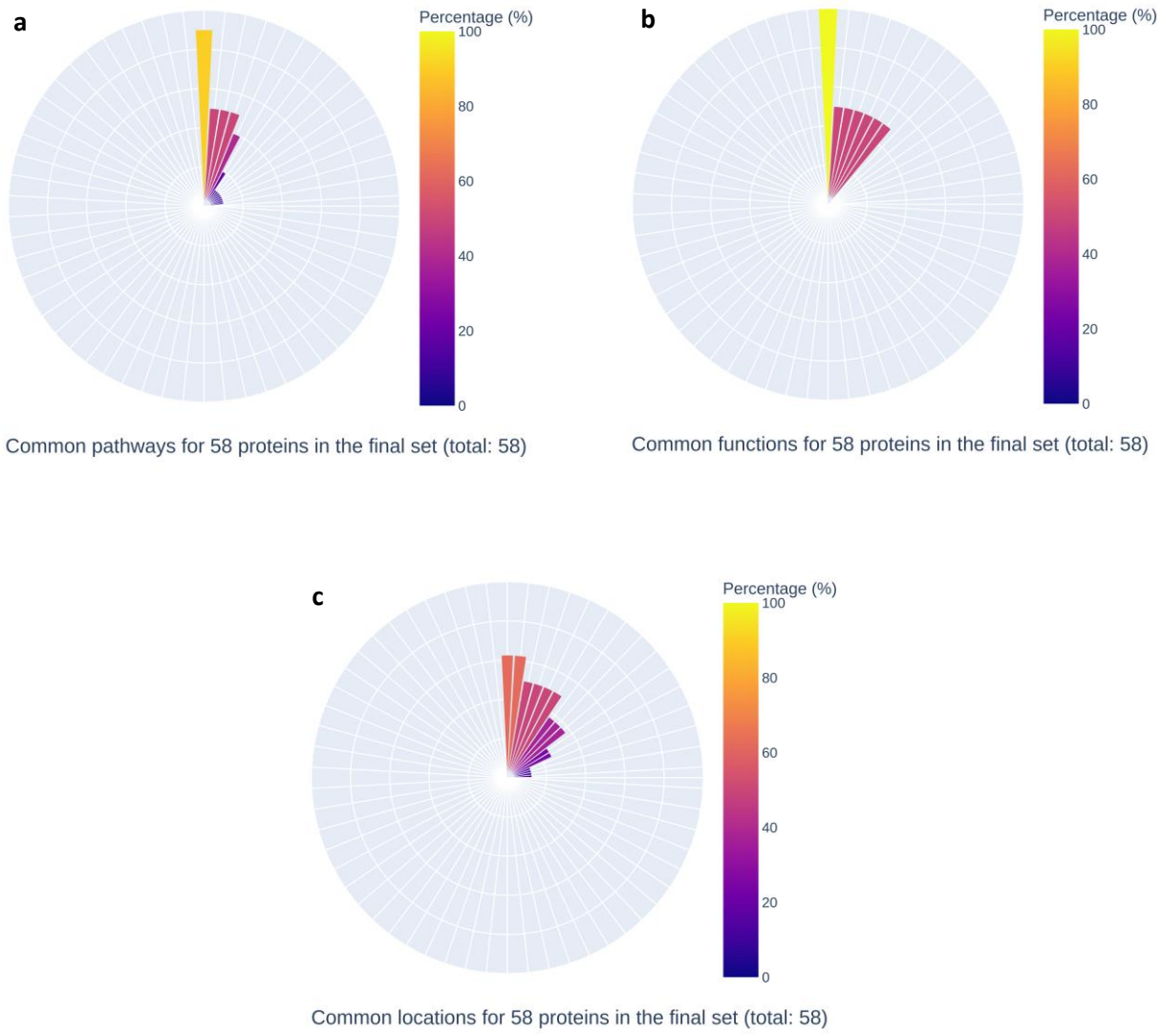

Figure 5: Common Gene Ontology terms on whole structure comparison results with SARS-CoV-2 Spike protein (native). Each protruding bar from the center of each radial plot represents a protein entry from the final set yielded by Machaon. The plots are generated by the presentation module. **a-c)** Common Gene Ontology terms retrieved from UniProt/EMBL QuickGo: **a** for biological processes, **b** for molecular functions and **c** for cellular locations.

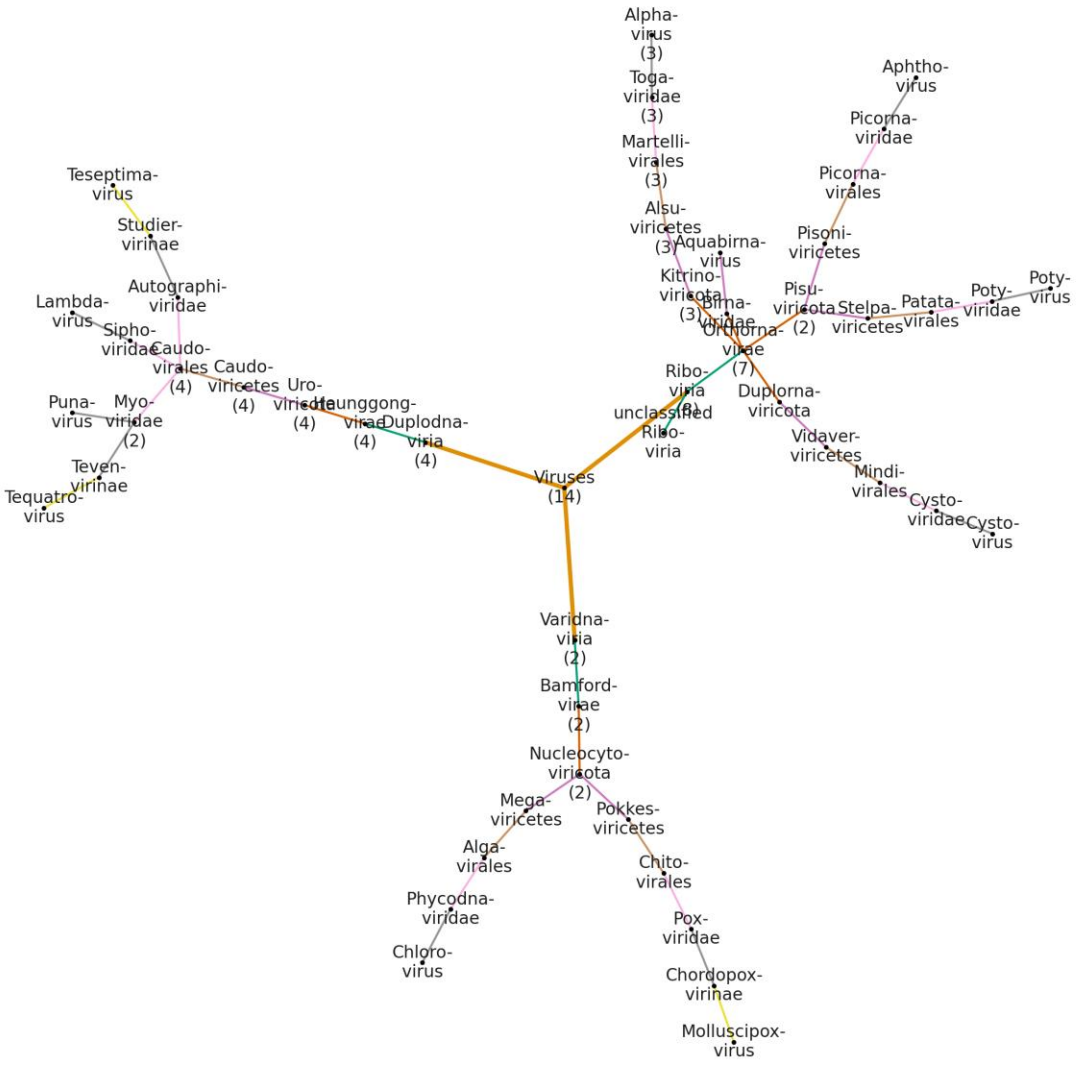

Figure 6: The lineage tree of the viral proteins in the final set of the constrained structural comparisons with S1 N-terminal domain (NTD) of Spike protein (native). These trees are generated by Machaon's presentation module. Each family name carries also a population number if there are more than one protein categorized under it. The root of the tree starts with thicker branches and the colors designated the branch levels. The tree refers to the viral proteins that were found most structurally relevant. These trees could also be considered as distant evolutionary trees according to the structural traits of the proteins in the results. The lineage information is retrieved from UniProt. (Zoom to review)

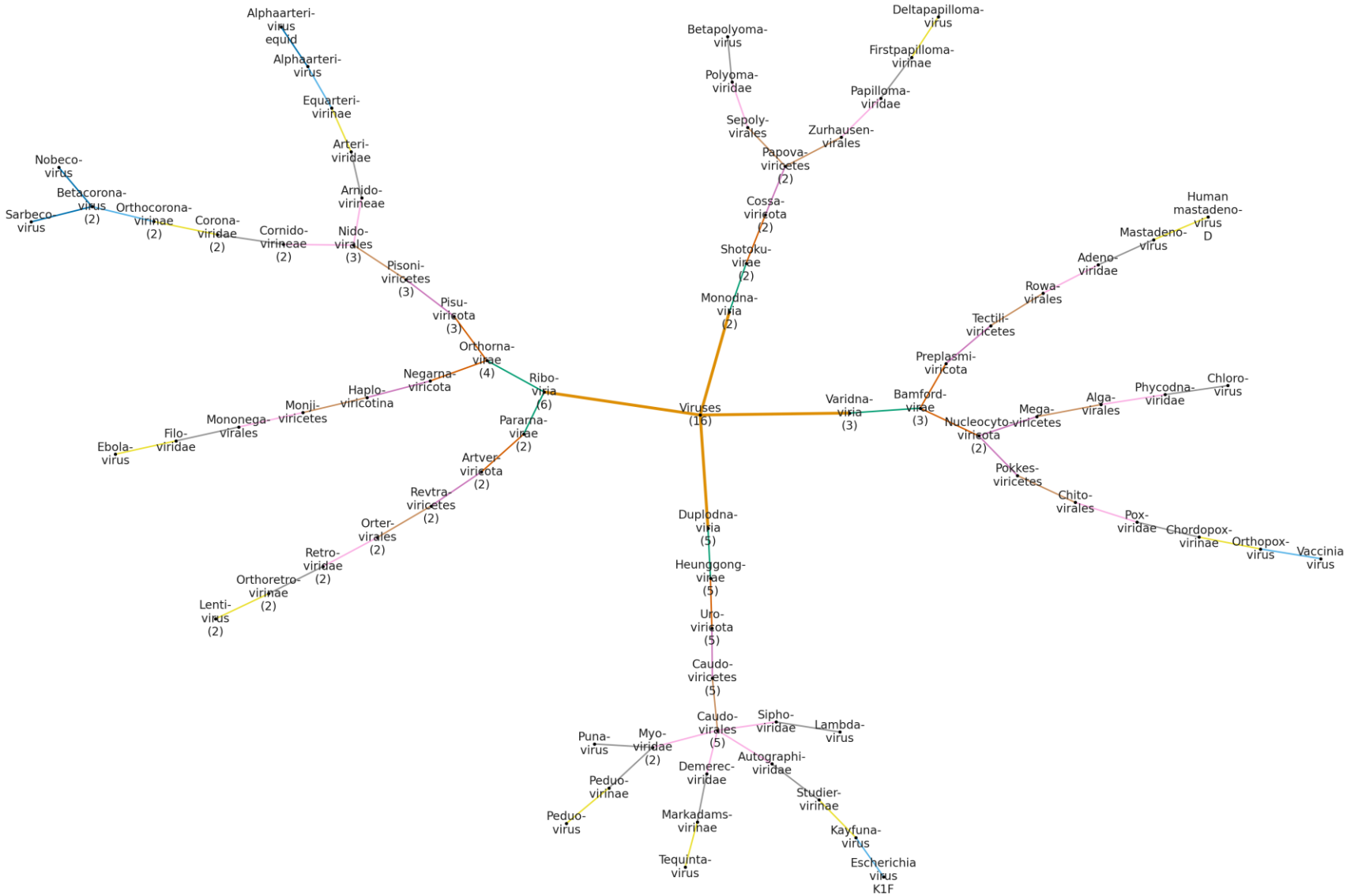
Alphaarteri-virus equid
Alphaarteri-virus
Equarteri-virinae
Arteri-viridae
Arnido-virineae
Nobeco-virus
Betacorona-virus (2)
Sarbeco-virus
Orthocorona-virinae (2)
Corona-viridae (2)
Cornido-virineae (2)
Nido-virales (3)
Pisoni-viricetes (3)
Pisu-viricota (3)
Orthorna-virae (4)
Negarna-viricota
Haplo-viricotina
Monji-viricetes
Mononega-virales
Filo-viridae
Ebola-virus
Ribo-viria (6)
Parama-virae (2)
Artver-viricota (2)
Revtra-viricetes (2)
Orter-virales (2)
Retro-viridae (2)
Orthoretro-virinae (2)
Lenti-virus (2)
Betapolyoma-virus
Polyoma-viridae
Sepoly-virales
Papova-viricetes (2)
Deltapapilloma-virus
Firstpapilloma-virinae
Papilloma-viridae
Zurhausen-virales
Cossa-viricota (2)
Shotoku-virae (2)
Monodna-viria (2)
Viruses (16)
Duplodna-viria (5)
Heunggong-virae (5)
Uro-viricota (5)
Caudo-viricetes (5)
Caudo-virales (5)
Myo-viridae (2)
Puna-virus
Peduo-virinae
Peduo-virus
Demerec-viridae
Markadams-virinae
Tequinta-virus
Sipho-viridae
Lambda-virus
Autographi-viridae
Studier-virinae
Kayfuna-virus
Escherichia virus K1F
Varidna-viria (3)
Bamford-virae (3)
Preplasmi-viricota
Tectili-viricetes
Rowa-virales
Adeno-viridae
Mastadeno-virus
Human mastadeno-virus D
Nucleocyto-viricota (2)
Mega-viricetes
Alga-virales
Phycodna-viridae
Chloro-virus
Pokkes-viricetes
Chito-virales
Pox-viridae
Chordopox-virinae
Orthopox-virus
Vaccinia virus

Figure 7: The lineage tree of the viral proteins in the final set of the constrained structural comparisons with S1 C-terminal domain (CTD) of Spike protein (native). These trees are generated by Machaon's presentation module. Each family name carries also a population number if there are more than one protein categorized under it. The root of the tree starts with thicker branches and the colors designated the branch levels. The tree refers to the viral proteins that were found most structurally relevant. These trees could also be considered as distant evolutionary trees according to the structural traits of the proteins in the results. The lineage information is retrieved from UniProt. (Zoom to review)

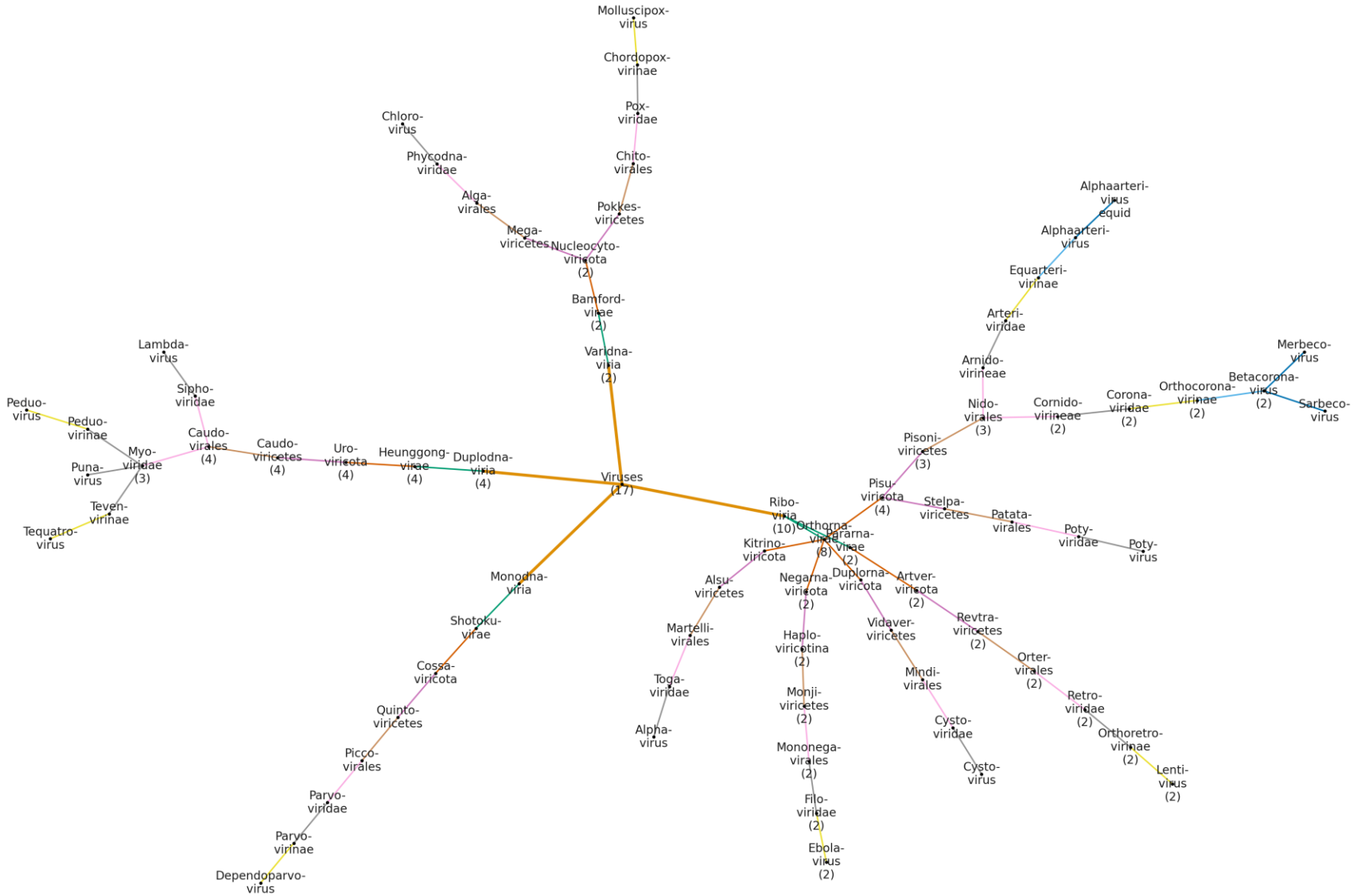
Molluscipox-virus
Chordopox-virinae
Pox-viridae
Chloro-virus
Phycodna-viridae
Chito-virales
Alga-virales
Pokkes-viricetes
Mega-viricetes
Nucleocyto-viricota (2)
Bamford-virae (2)
Varidna-viria (2)
Alphaarteri-virus equid
Alphaarteri-virus
Equarteri-virinae
Arteri-viridae
Arnido-virineae
Merbeco-virus
Betacorona-virus (2)
Orthocorona-virinae (2)
Corona-viridae (2)
Cornido-virineae (2)
Nido-virales (3)
Sarbeco-virus
Pisoni-viricetes (3)
Lambda-virus
Sipho-viridae
Peduo-virus
Peduo-virinae
Caudo-virales (4)
Myo-viridae (3)
Puna-virus
Teven-virinae
Tequatro-virus
Caudo-viricetes (4)
Uro-viricota (4)
Heunggong-virae (4)
Duplodna-viria (4)
Viruses (17)
Pisu-viricota (4)
Stelpa-viricetes
Patata-virales
Poty-viridae
Poty-virus
Ribo-viria (10)
Orthorna-virae (8)
Pararna-virae (2)
Kitrino-viricota
Monodna-viria
Shotoku-virae
Cossa-viricota
Quinto-viricetes
Picco-virales
Parvo-viridae
Parvo-virinae
Dependoparvo-virus
Alsu-viricetes
Martelli-virales
Toga-viridae
Alpha-virus
Negarna-viricota (2)
Duplorna-viricota
Artver-viricota (2)
Haplo-viricotina (2)
Vidaver-viricetes
Revtra-viricetes (2)
Monji-viricetes (2)
Mindi-virales
Orter-virales (2)
Mononega-virales (2)
Cysto-viridae
Retro-viridae (2)
Filo-viridae (2)
Cysto-virus
Orthoretro-virinae (2)
Ebola-virus (2)
Lenti-virus (2)

Figure 8: The lineage tree of the viral proteins in the final set of the constrained structural comparisons with S1 receptor-binding domain (RBD) of Spike protein (native). These trees are generated by Machaon's presentation module. Each family name carries also a population number if there are more than one protein categorized under it. The root of the tree starts with thicker branches and the colors designated the branch levels. The tree refers to the viral proteins that were found most structurally relevant. These trees could also be considered as distant evolutionary trees according to the structural traits of the proteins in the results. The lineage information is retrieved from UniProt. (Zoom to review)

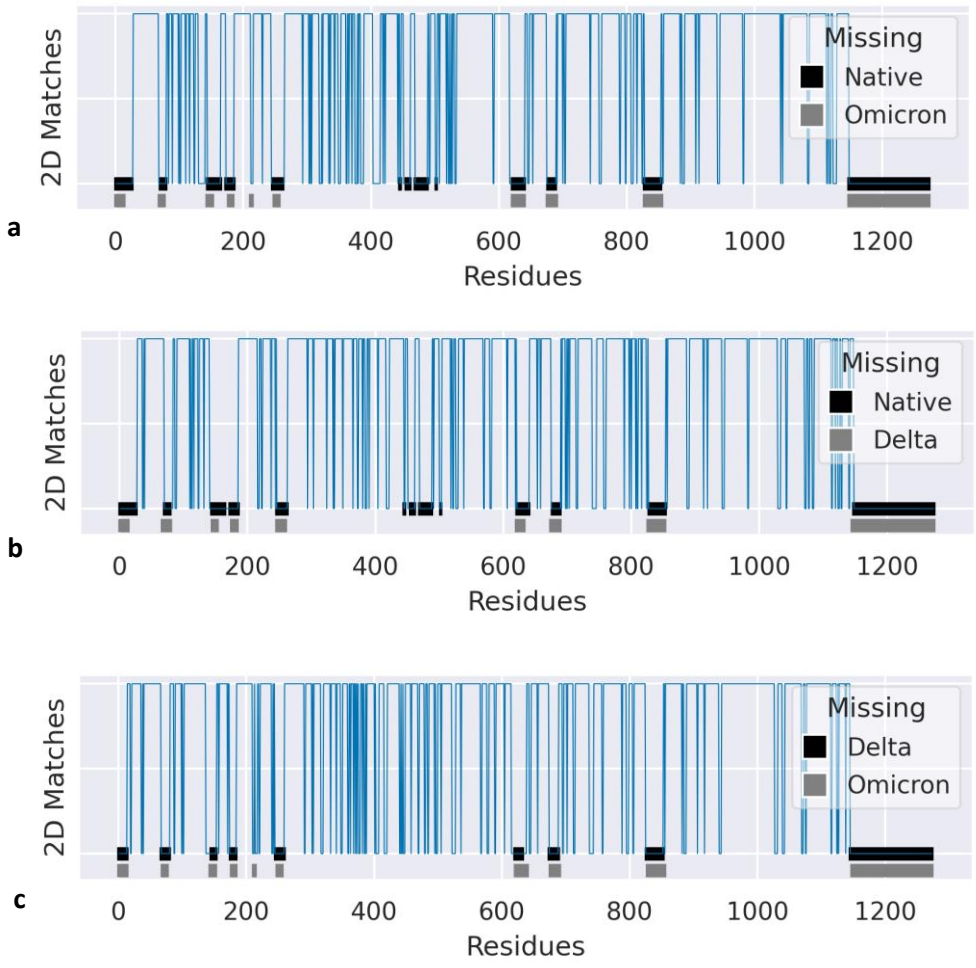

Figure 9: The difference in secondary proteins structures between the variants of SARS-CoV-2 Spike protein (native, Delta, Omicron). The depicted comparisons are between: **a)** Native and Delta (6VXX.A & 7V7Q.A), **b)** Native and Delta (6VXX.A & 7T9K.A), **c)** Delta and Omicron (7V7Q.A & 7T9K.A). These data were produced by using Machaon's modules programmatically. Areas with missing data in the reference PDB file are annotated with black and gray boxes on the x-axis.

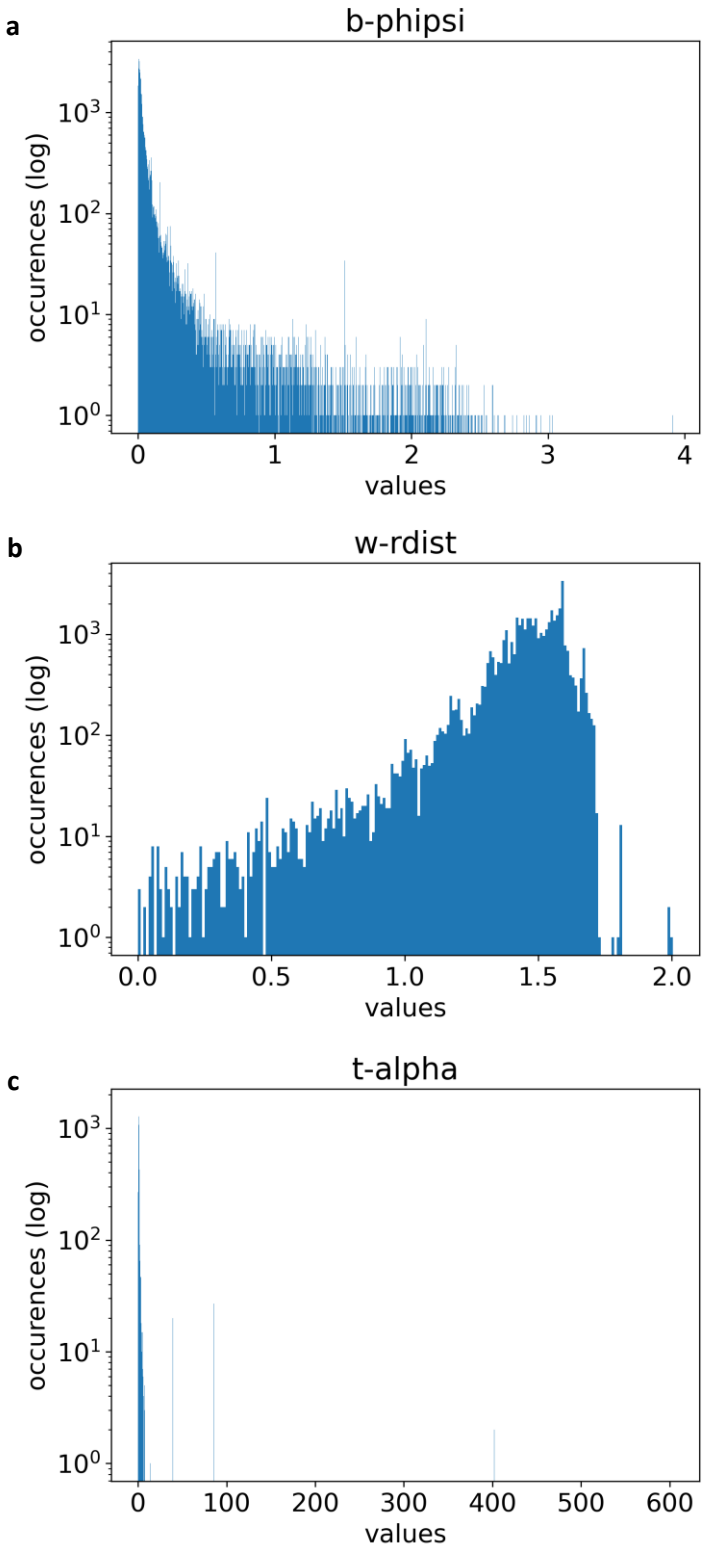
a
b-phipsi
occurences (log)
values
b
w-rdist
occurences (log)
values
c
t-alpha
occurences (log)
values

Figure 10: The numerical space of the metrics as calculated between SARS-CoV-2 Spike protein (native strain) and our study's viral dataset. The histograms include computed metrics on more than 41000 PDB chains and their bins are determined by Freedman-Diaconis rule; **a** plot refers to b-phipsi, **b** to w-rdist and **c** to t-alpha metric. We can notice that the metrics' distributions have peaks localizing near to zero and outliers within a finite small margin.

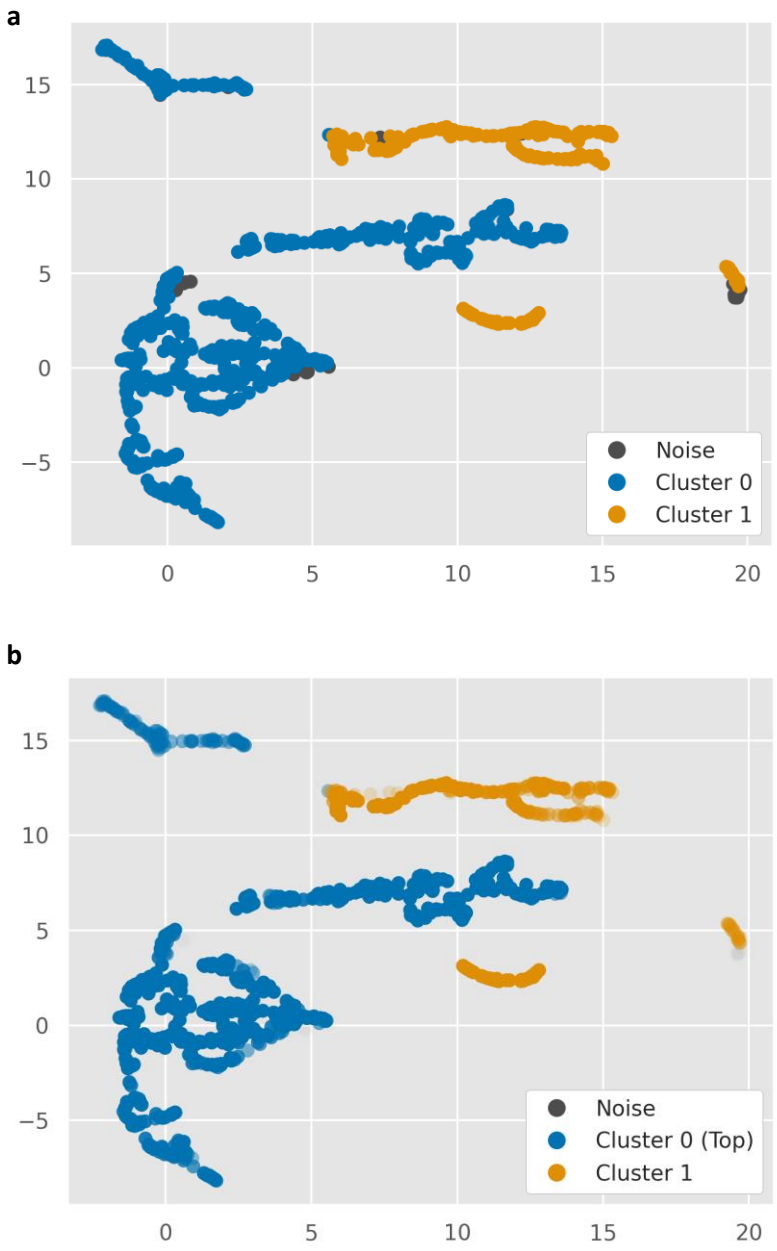

Figure 11: The clustering result on whole structure comparisons with SARS-CoV-2 Spike protein (native). Each point corresponds to a PDB structure in the dataset, which is represented by a three-dimensional vector (the values of the proposed metrics). Some proteins participate with multiple PDBs in the dataset. The visualization is generated by UMAP method by reducing the dimensions of the data to 2D. **a** depicts the clustering result by HDSCAN and **b** the final selection and form of the top cluster determined by the ranking via b-phipsi, w-rdist and t-alpha. The blue cluster ('Top') is chosen as the preferred one, a selection based on the order of the vectors derived from rank aggregation. The transparency of a point matches its clustering probability. The light greyed data points are interpreted as noise by Hierarchical DBSCAN (not visible in **b** plot due to their low clustering probability). Also, every point that has clustering probability below 0.1 is discarded from the clusters, labelled as noise. The figure displays the separation of the clusters.

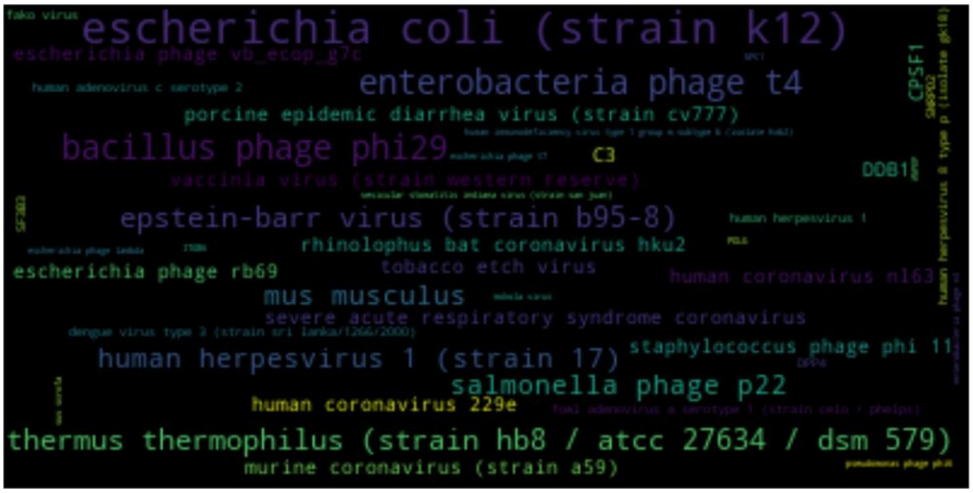

Figure 12: The world cloud visualization of the whole structure comparisons with SARS-CoV-2 Spike protein (native). The figure is based on the 'organism' and 'gene' metadata fields of the PDB files containing the data of structurally concordant proteins to the reference (Spike) as determined by Machaon. Non-human proteins are represented by the 'organism' metadata and human proteins by the 'gene' metadata. The large letter font is correlated with the increased frequency of appearance in the final output set.

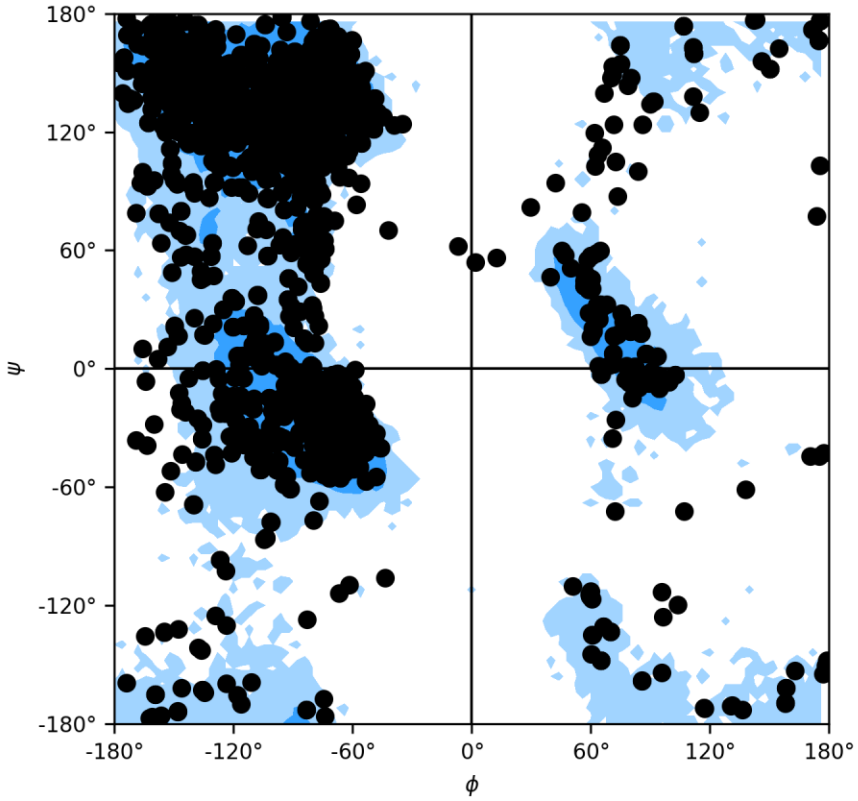

Figure 13: Ramachandran plot for preprocessed PDB data on SARS-CoV-2 Spike protein (native). This plot depicts the backbone phi-psi angles of the residues listed in the preprocessed 6VXX.A PDB file. The colored areas refer to the most favorable spatial arrangements regarding minimum potential energy. The figure was generated with MDAnalysis Python package.

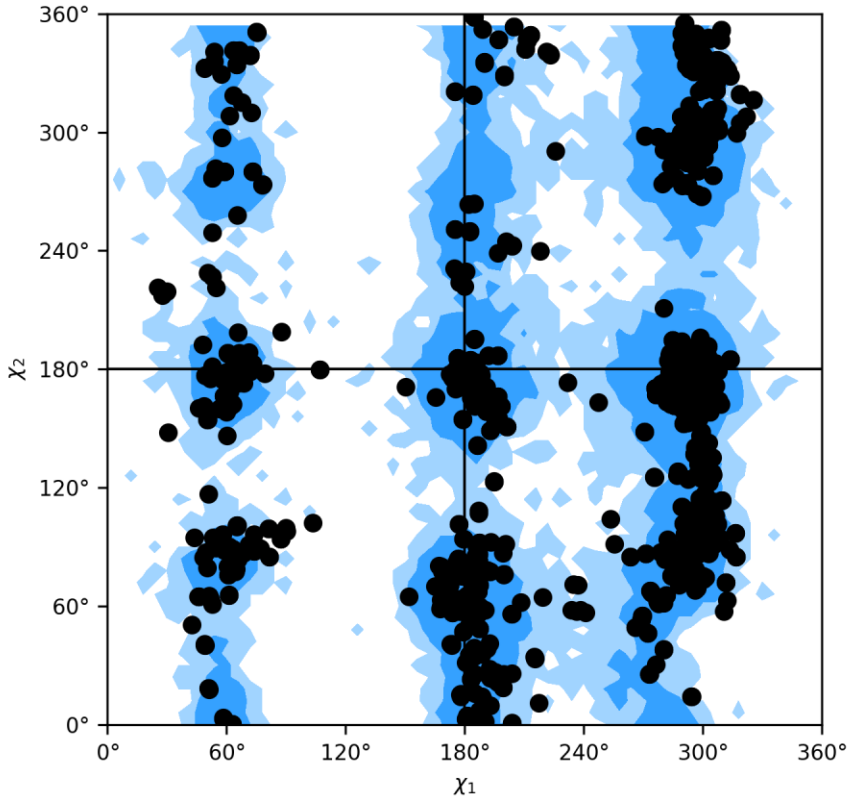

Figure 14: Janin plot for preprocessed PDB data on SARS-CoV-2 Spike protein (native). This plot depicts the side chain chi1-chi2 angles of the residues listed in the preprocessed 6VXX.A PDB file. The colored areas refer to the most favorable spatial arrangements regarding minimum potential energy. The figure was generated with MDAnalysis Python package.