A computational approach to investigate the structural and functional consequences of residue mutations in human disease and protein design

Mahnaz Abbasian

A thesis submitted for the degree of

Doctor of Philosophy

January 2025

Institute of Structural and Molecular Biology

University College London

Declaration

I, Mahnaz Abbasian, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Genes encode information translated into proteins, essential for biological functions. Genetic mutations alter DNA sequences, potentially modifying protein structures and functions with diverse effects. In nature, mutations drive genetic diversity, enabling beneficial innovations or detrimental changes, including extinction. In health and disease, mutations influence susceptibility, resistance, and disease progression, including cancer. These impacts underscore the importance of mutations in evolution, health, and disease.

The first project focused on *Ideonella sakaiensis* PETase (*Is*PETase), aiming to discover potent PETases with improved substrate binding affinity. Computational algorithms analysed over a billion metagenomic sequences from the MGnify database, narrowing down a PETase subset. Twenty-seven putative PETase sequences were shortlisted, and three demonstrated in vitro activity, marking a discovery in identifying naturally evolved PETases.

The second project examined amino acid changes in human and SARS-CoV-2 proteins affecting binding affinity in human:SARS-CoV-2 complexes. Among 450 human protein missense mutations, sixteen were predicted to enhance binding affinity ($\Delta\Delta G \geq 0.5$ kcal/mol) with SARS-CoV-2 proteins, involving cell-entry receptors, immune responses, and translation machinery. Additionally, three SARS-CoV-2 mutations were predicted to strengthen interactions with human proteins. This research highlighted genetic variations' potential role in COVID-19 susceptibility across ethnic groups.

The final project studied mutations impacting protein function in lung adenocarcinoma (LUAD), analysing data from the TRACERx database. Using paralog protein structures in CATH, FunVar identified functional impact events (FIEs) in known and novel driver genes. Pre-genome duplication FIEs dominated, while post-duplication FIEs contributed to LUAD specialisation. Genes with FIE mutations revealed enriched metabolic pathways in LUAD, providing insights into tumour evolution.

Together, these projects advanced understanding of the role of genetic mutations in enzyme evolution, disease susceptibility, and cancer progression.

Impact statement

This work advances our understanding of the complex impact of genetic mutations on evolution, health, and disease. By employing computational approaches, the studies allowed us a deeper understanding into protein sequence-structure-function relationships and their implications for enzyme optimisation, disease susceptibility, and cancer progression.

Using CATH FunFams, functional determinant (FD) positions were identified based on differential conservation across FunFams from over one billion metagenomic sequences. FD positions near catalytic triads and binding pockets facilitated the shortlisting of efficient putative PETases for experimental validation. Out of these, three sequences were confirmed to exhibit PETase activity, marking a valuable advancement in identifying functional enzymes. By integrating machine learning with structural prediction, the study achieved remarkable accuracy in modelling enzyme structures and guiding substitutions to enhance solubility, stability, and performance. These findings provide a robust framework for engineering enzymes tailored to degrade diverse plastics and function under varying industrial conditions, contributing to scalable solutions to critical environmental challenges.

The second research study explored the impact of amino acid changes in human and SARS-CoV-2 proteins on the binding affinity of human:SARS-CoV-2 protein complexes, shedding light on genetic factors influencing COVID-19 susceptibility across diverse ethnic groups. Using publicly available human and viral databases, 12 protein complexes critical to viral entry, immune response, and cellular translation were analysed.

Computational analysis identified 16 human missense variants that increased binding affinity to SARS-CoV-2 proteins ($\Delta\Delta G \geq 0.5$ kcal/mol), including three proteins associated with spike-binding, immune responses, and translation. Among over 200 SARS-CoV-2 missense residues, five mutations were found to enhance binding affinity with human cell receptors and immune-related proteins. These findings underscore the influence of genetic variation in both host and virus on infection dynamics and immune response. This work offers critical insights for understanding COVID-19 susceptibility and guiding personalized therapeutic strategies across populations.

The third research study integrated advanced computational tools to deepen our understanding of Functional Impact Events (FIEs) from genetic variations on lung adenocarcinoma (LUAD) progression. By employing the TRACERx consortium dataset, the study identified FIEs near critical protein functional sites, analysing their roles in protein stability, ligand affinity, and tumorigenesis. The characterisation of FIEs as loss-of-function, gain-of-function, or neofunctionalisation provided insights into their biological and evolutionary timing, pre- or post-genome duplication. Pathway enrichment analyses using g:Profiler and KEGG databases linked FIE-affected genes to key cancer pathways. Genes with post-duplication FIEs exhibited greater diversity and were found on the same pathways with pre-duplication FIE genes, suggesting their role in amplifying the functional impact of genetic variations on cancer genome evolution. This work underscores the transformative impact of bioinformatics and high-throughput technologies like AlphaFold2 in protein structure prediction and mutation analysis. By connecting molecular mechanisms with tumorigenesis, the study provides a framework for identifying novel

therapeutic targets and improving our understanding of LUAD pathology, advancing the fight against cancer.

Advancements in machine learning and deep learning are transforming the analysis of large-scale datasets, unveiling patterns critical to understanding mutations that drive evolution, biodiversity, and their pivotal roles in health and disease.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Christine Orengo, for her guidance, encouragement and support throughout my PhD journey. Her mentorship has been invaluable and has shaped both this thesis and my growth as a researcher.

I am also sincerely thankful to my secondary supervisor, Professor Andrew Martin, and my thesis committee chair, Professor Joanne Santini, for their insightful feedback, encouragement and support during the course of my research.

This work would not have been possible without the remarkable contributions of the CATH team: Dr Paul Ashford, Dr Nico Bordin, Dr Sayoni Das, Dr Natalie Dawson, Miss Weining Lei, Dr Clemens Rauer, Dr Joel Roca, Dr Ian Sillitoe, Dr Neeladri Sen, Dr Su Datt Lam and Dr Vaishali Waman for their collaborative efforts and support.

I am also thankful to our collaborators, Professors Rob Finn and Florian Hollfelder and their research groups, for their invaluable contributions, which have greatly enriched this thesis. I extend my appreciation to all members of the Institute of Structural and Molecular Biology, whose assistance have been instrumental during my time at UCL.

I am deeply grateful to the people in my life outside of academia who provided me with emotional support. To my family for believing in me and igniting hope when challenges seemed insurmountable.

Contents

Chapter 1: Introduction 19

1.1 Characteristics of amino acids and different levels of protein structures 20
1.2 Domains, motifs, binding sites, catalytic sites 21
1.3 Evolvability 22
1.4 Methodological focus and tool selection 23
1.5 Bioinformatics methods to analyse protein sequences similarity 23
1.5.1 Multiple Alignment using Fast Fourier Transform (MAFFT) 24
1.5.2 Hidden Markov Models (HMMs) 25
1.5.3 HH-suite 28
1.5.4 Cluster Database at High Identity with Tolerance (CD-HIT) 29
1.6 Quantifying residue conservation 29
1.6.1 Scorecons 30
1.6.2 Diversity of positions (DOPs) score 30
1.6.3 Identification of specificity determining positions in a protein family 31
1.7 Resources for protein classification 33
1.7.1 Pfam 33
1.7.2 CATH 33
1.7.2.1 Sub-classification of CATH evolutionary superfamilies into functional families 37
1.7.2.2 FunFamer 40
1.7.2.3 FunFam generation by Multidomain Architecture-based Clustering 41
1.8 MGnify: resource for storing and assembling metagenomic samples 43
1.9 Protein structure prediction 44
1.9.1 Homology modelling 45
1.9.2 SWISS-MODEL 45
1.9.3 Deep learning-based modelling 48
1.9.4 AlphaFold2 49
1.10 Protein functional annotation 51
1.10.1 Enzyme commission number 51
1.10.2 Gene Ontology (GO) 52
1.10.3 Analysing metabolic pathways 53
1.10.4 Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) 54
1.11 Genetic variants and disease susceptibility 55
1.11.1 Genomic instability and cancer transformation 56

1.11.2 Prediction of mutation effects on proteins 56
1.12 Predicting the impact of genetic variations 57
1.12.1 mCSM-PPI2 57
1.12.2 MutPred2 58
1.12.3 Polymorphism Phenotyping 58
1.12.4 Sorting Intolerant From Tolerant (SIFT) 59
1.12.5 Combined annotation dependent depletion (CADD) 59
1.12.6 VarMap 60
1.13 Resources providing information of known and predicted functional sites 61
1.13.1 BioLip 61
1.13.2 Mechanism and Catalytic Site Atlas (M-CSA) 62
1.13.3 Inferred Biomolecular Interaction Server (IBIS) 62
1.14 Overview of the thesis 63

1 Chapter 2: Exploration of naturally evolved polyethylene terephthalate hydrolases from metagenomic data 66

Chapter 3: Computational analysis of the effect of amino acid changes in SARS-CoV-2 proteins and human interactor proteins on the binding affinity between the host and viral proteins 67

3.1. Introduction 67
3.1.1 Comparison of SARS-CoV and SARS-CoV-2 68
3.1.2. SARS-CoV-2 structure 70
3.1.3 SARS-CoV-2 spike protein structure 72
3.1.4 SARS-CoV-2 RBD:hACE2 interface binding and the mechanism of host infection 75
3.1.5 SARS-CoV-2 variants of concern 79
3.1.5.1. Alpha variant 80
3.1.5.2. Beta variant 81
3.1.5.3. Gamma variant 82
3.1.5.4. Delta variant 82
3.1.5.5. Omicron variant 83
3.1.6 The role of other SARS-CoV-2 proteins in COVID-19 susceptibility and infection 86
3.1.7 Aim of the project 89
3.2 Methods 90

3.2.1. Viral zone 90
3.2.2. CoV-GLUE-Viz 91
3.2.3. The Genome Aggregation Database (gnomAD) 91
3.2.4. All of Us 92
3.2.5. SweGen 92
3.2.6. GenomeAsia 100K 93
3.2.7. IndiGenomes 93
3.2.8. jMorp 94
3.2.9. Dataset of human:SARS-CoV-2 interactor protein complexes 94
3.2.10. HADDOCK 97
3.2.11. Validating the AlphaFold2 models conformation using HADDOCK 97
3.2.12. OHM, Allostery site prediction 101
3.2.13. Predicting the impact of human and viral protein variants on the binding affinity of the corresponding complexes using mCSM-PPI2 101
3.2.14. Predicting the impact of human protein variants on the binding affinity of the human:SCov2 complexes using PROtein binDIng enerGY prediction 103
3.2.15. Other algorithms and methods 104
3.3. Results 105
3.3.1. Impact of hTOM70 mutations on hTOM70:SARS-CoV-2 ORF9b interaction 108
3.3.2. SARS-CoV-2 papain-like protease targets hISG15, hIFIH1 and hIFIT2 to supress the immune response 111
3.3.3. Impact of hIFIH1 mutations on hIFIH1:SARS-CoV-2 PLpro interaction 112
3.3.4. Impact of hIFIT2 mutations on hIFIT2:SARS-CoV-2 PLpro interaction 114
3.3.5. Impact of SARS-CoV-2 mutations on binding affinity to human interactor proteins 118
3.4. Conclusion and future work 120

Chapter 4: Identifying driver mutations in lung adenocarcinoma 126

Chapter 5: Conclusions and future directions 127

References 130

List of figures

Figure 1.1 Classification of amino acids according to their sidechain or R group's physicochemical properties.
Figure 1.2. The four levels of hierarchy in protein conformation.
Figure 1.3. Illustration of motif and domain.
Figure 1.4. A schematic model of a profile HMM.
Figure 1.5. Sequence alignment-based filtering of columns by GroupSim.
Figure 1.6. An example CATH classification.
Figure 1.7. The basic GeMMA clustering pipeline algorithm.
Figure 1.8. The FunFamer algorithm determines an optimal cut of the tree by incorporating Scorecons and GroupSim algorithms.
Figure 1.9. Multi-Domain Architecture.
Figure 1.10. A schematic pipeline of FunFam-MARC, multiple parallel and separate GeMMA runs.
Figure 1.11. Procedure for transforming raw reads from environmental samples into meaningful biological information.
Figure 1.12. Building homology mode using SWISS-MODEL.
Figure 1.13. GO annotations.
Figure 1.14. An example of STRING protein-protein interactions.
Figure 3.1. Similarities and differences in SARS-CoV-2 and SARS-CoV RBDs.
Figure 3.2. Schematic structure of SARS-CoV-2.
Figure 3.3. Sequence and structure of the Spike protein.
Figure. 3.4. Sequence and structure of SARS-CoV-2 RBD.
Figure 3.5. The structures of the prefusion and post-fusion trimeric spike for SARS-CoV-2.
Figure 3.6. Stepwise demonstration of conformational changes of RBD.
Figure 3.7. A cartoon representation showing the pre- to the post-fusion transition of the SARS-CoV-2 Spike.
Figure 3.8. A summary of the ACE2:RBD interactions.
Figure 3.9. The interaction network of RBD with ACE2.
Figure 3.10. A schematic presentation of the SARS-CoV-2 genome and encoded proteins.

Figure 3.11. The number of people in the gnomAD database, broken down by demographic and subpopulation.
Figure 3.12. Ribbon (left) and space-filled (right) models of hACE2:SCoV2 RBD complex.
Figure 3.13. Positions of hTOM70 affinity-enhancing residues: Val556, Lys576 and Ala591
Figure 3.14. The position of affinity enhancing residues in hIFIT2 and their proximity to conserved residues in space-filled and ribbon models.

List of tables

Table 3.1. Summary of SARS-CoV-2 and SARS-CoV interactions with ACE2.
Table 3.2. List of Spike's NTD and RBD mutations in Alpha variant.
Table 3.3. List of Spike's NTD and RBD mutations in Beta variant.
Table 3.4. List of Spike's NTD and RBD mutations in Gamma variant.
Table 3.5. List of all mutations in Spike's Delta variant.
Table 3.6. List of common Spike's NTD and RBD mutations in Omicron sub-lineages.
Table 3.7: Details of complexes for downstream human: SARS-CoV-2 interactor protein-protein analysis.
Table 3.8. Verification of conformation and interaction in AF modelled complexes.
Table 3.9. A summary of reported human mutation in each human:SARS-CoV-2 complex in gnomAD and the number of affinity-enhancing mutations in the corresponding complex.
Table 3.10. Summary of three hTOM70 substitutions with predicted enhancing binding affinity.
Table 3.12. Summary of two hIFIH1 substitutions with predicted enhancing binding affinity.
Table 3.13. Summary of six hIFIT2 substitution with predicted enhancing binding affinity.
Table 3.14. SARS-CoV-2 mutations and their effects on human protein interactions.
Table 3.15. Summary of affinity-enhancing human protein to the corresponding SCoV-2 protein interactors and their associated functional features.

List of abbreviations

Abbreviation	Phrase
+ssRNA	positive-strand single-stranded RNA
2-HG	D-2-hydroxyglutarate
3D	3-dimensional
ACE2	angiotensin-converting enzyme 2
AMF	autocrine motility factor
ARF6	ADP-ribosylation factor 6
BHET	Bis-(2-Hydroxyethyl) Terephthalate
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrix
CADD	Combined annotation dependent depletion
CASP	Critical Assessment of Structural Prediction
CATH	Class, Architecture, Topology, Homologous Superfamily
CD-HIT	Cluster Database at High Identity with Tolerance
COVID-19	Coronavirus Disease-19
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats / CRISPR-associated protein 9
CTD1	C-terminal domain 1
DC	direct contact
DCEX	direct contact extended residues
df	Degrees of freedom
DOPs	Diversity of positions
DOPS	Diversity of Position Scores
E protein	Envelope protein
E-value	expectation value
EC	Enzyme Classification
EC	Enzyme Commission
EG	Ethylene Glycol
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
ENA	European Nucleotide Archive
FD	Functional Determinant
FFT	Fast Fourier Transform
FIEs	Functional Impact Events
FRAN	FunFam generation by RANdom splitting
FunFam	Functional Family
FunFam-MARC	FunFam generation by Multidomain Architecture-based Clustering
GAP	GTPase-activating protein
GDT	Global Distance Test
GMQE	Global Model Quality Estimation
gnomAD	Genome Aggregation Database

Abbreviation	Phrase
GO	Gene Ontology
GOBP	Gene Ontology Biological Process
GPI	Glucose-6-phosphate isomerase
HADDOCK	High Ambiguity Driven biomolecular DOCKing
HMM	Hidden Markov Models
HPLC	High-Performance Liquid Chromatography
hRPS3	Human Ribosomal Protein S3
HSP90	heat shock protein 90
hTRIM25	Human tripartite motif protein 25
IBIS	Inferred Biomolecular Interaction Server
ICTV	International Committee on Taxonomy of Viruses
IDH1	isocitrate dehydrogenase 1
IFIT2	Interferon-induced protein with tetratricopeptide repeats 2
IFN	interferon
ITH	intra-tumour heterogeneity
iTOL	interactive tree of life
JMML	juvenile myelomonocytic leukaemia
KO	KEGG Orthology
KRAS	Kirsten rat sarcoma viral oncogene homologue
LC	Lung cancer
LC–MS	liquid chromatography-mass spectrometry
LCC	Leaf-Branch Compost Cutinase
LGG	low-grade gliomas
LUAD	lung adenocarcinoma
M protein	Membrane proteins
M-CSA	Mechanism and Catalytic Site Atlas
MAFFT	Multiple Alignment using Fast Fourier Transform
MDAs	MULTI-DOMAIN ARCHITECTURES
MERS-CoV	Middle East Respiratory Syndrome coronavirus
MHET	mono-2-hydroxyethyl terephthalate
MME	membrane metalloendopeptidase
MMseqs	Many-against-Many sequence searching
MSA	Multiple sequence alignment
N protein	Nucleocapsid protein
NIH	National Institutes of Health
NMR	Nuclear Magnetic Resonance
NSCLC	non-small cell lung cancer
NSPs	non-structural proteins
nsSNP	nonsynonymous Single Nucleotide Polymorphisms
NTD	N-terminal domain
NTP	nucleoside triphosphate
ORFs	open-reading frames
PALS1	Protein Associated with LIN7 1

Abbreviation	Phrase
PAM	Point Accepted Mutation
PAZy	Plastics-Active Enzymes
PBSA	poly-butylene succinate-co-butylene adipate
PDB	Protein Database
PET	Polyethylene Terephthalate
pNPB	p-nitrophenyl butyrate
PolyPhen	Polymorphism Phenotyping
PPIs	Protein-Protein Interactions
PRODIGY	PROtein binDIng enerGY prediction
PSC	Percentage of Scorecons
PSI-BLAST	Position-Specific Iterated Basic Local Alignment Search Tool
RBD	receptor binding domain
RBM	Receptor Binding Motif
RMSD	Root-Mean-Square Deviation
RNP	ribonucleocapsid
S protein	Spike protein
SARS-CoV	Severe Acute Respiratory Syndrome coronavirus
SARS-CoV-2/ScoV2	Severe viral acute respiratory syndrome coronavirus 2
SCLC	small-cell lung cancer
SD1	subdomain1
sdFunFams	sequence-diverse FunFams
SDPs	Specificity Determining Positions
SIFT	Sorting Intolerant From Tolerant
SMILE	Simplified Molecular Input Line Entry System
SNPs	Single Nucleotide Polymorphisms
SSAP	Secondary Structure Alignment Program
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SWI/SNF	SWItch/Sucrose Non-Fermentable
TA	Terephthalic Acid
TRACERx	Tracking Cancer Evolution through Therapy
TRIMM	Tripartite motif
UniProtKB	UniProt Knowledgebase
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
VOCs	Variant Of Concerns
WGD	whole genome duplication
WHO	World Health Organisation
WT	Wild-type

List of publications

Rauer C., Bordin N., Abbasian M., Roca-Martinez J., Range M., Holstein J. M, Schäfer E., Hollfelder F., Finn R., Orengo C. “Discovery of PETases using a computational classification system”. University College London and University of Cambridge,2025 (Unpublished manuscript)

Ashford P, Frankell AM, Piszka Z, Pang CSM, Abbasian M, Al Bakir M, Jamal-Hanjani M, McGranahan N, Swanton C, Orengo CA. “*Gene duplication is associated with gene diversification and potential neofunctionalisation in lung cancer evolution*”. University College London, 2024 (Manuscript submitted for publication)

Waman VP, Ashford P, Lam SD, Sen N, Abbasian M, Woodridge L, Goldtzvik Y, Bordin N, Wu J, Sillitoe I, Orengo CA. “Predicting human and viral protein variants affecting COVID-19 susceptibility and repurposing therapeutics”. *Scientific Reports*, 2024, DOI: 10.1038/s41598-024-61541-1

Rauer C, Sen N, Waman VP, Abbasian M, Orengo CA. “Computational approaches to predict protein functional families and functional sites”. *Current Opinion in Structural Biology,* 2021, DOI: 10.1016/j.sbi.2021.05.012

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla S M Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, Mahnaz Abbasian, Sean Le Cornu, Su Datt Lam, Karel Berka, Ivana Hutařová Varekova, Radka Svobodova, Jon Lees, Christine A Orengo, “CATH: increased structural coverage of functional space”, *Nucleic Acids Research*, 2021, DOI: 10.1093/nar/gkaa1079

S. D. Lam, N. Bordin, V. P. Waman, H. M. Scholes, P. Ashford, N. Sen, L. van Dorp, C. Rauer, N. L. Dawson, C. S. M. Pang, M. Abbasian, I. Sillitoe, S. J. L. Edwards, F. Fraternali, J. G. Lees, J. M. Santini & C. A. Orengo. SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. Sci Rep 10, 16471, 2020, DOI: 10.1038/s41598-020-71936-5

Chapter 1: Introduction

Proteins are essential macromolecules that have evolved to perform a broad spectrum of functions in various biological systems, for example: enzymes (*e.g.*, DNA ligase), antibodies (*e.g.*, immunoglobulin G), cell receptors (*e.g.*, angiotensin-converting enzyme-2) cell signalling and cellular pathways (*e.g.*, kinases), structural components (*e.g.,* actin), messenger molecules (*e.g.*, growth hormone) and, transport and storage molecules (*e.g.*, and haemoglobin and ferritin). Understanding the relationship between protein sequence, structure, and function facilitates understanding of the mechanisms underlying protein evolution.

Although the number of deposited protein sequences in metagenomic databases is exponentially increasing (exceeding 2.4 billion non-redundant sequences according to the EMBL-EBI updates (https://www.ebi.ac.uk), the structure of only a fraction of these sequences has been resolved due to the cost and time constraints. This gap of knowledge, known as the protein sequence structure gap, has driven scientists to employ computational methods *e.g.*, AlphaFold2 (1) to predict the proteins' structures and functions.

This chapter describes general and fundamental bioinformatics concepts, methods, and resources relevant to the work presented in this thesis, followed by an outline of the following chapters.

1.1 Characteristics of amino acids and different levels of protein structures

A combination of 20 amino acids is linked together by covalent bonds between the carboxyl and amino groups to form an oligopeptide. Despite their identical core structures, the side chain or R group assigns each amino acid with a specific physicochemical property (non-polar aliphatic, polar uncharged, positively charged, negatively charged and non-polar aromatic), giving proteins specific structures and functions (Figure 1.1).

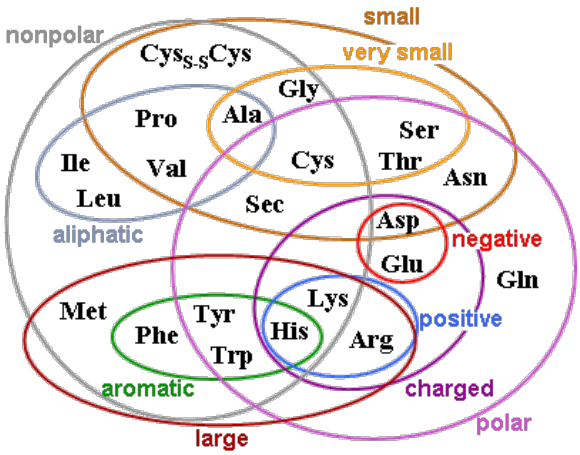

Figure 1.1. Classification of amino acids according to their side chain or R group's physicochemical properties such as polarity, electric charge, hydrophobicity and size (2).

The four levels of protein structures are: the primary structure relates to the sequence of amino acids; secondary structure describes local regular structures such as the alpha helix, beta-sheets, turns and loops; the tertiary structure is the three-dimensional (3D) structure of the entire protein; and, quaternary structure refers to structural assembly of proteins that consist of two or more polypeptide chains (Figure 1.2).

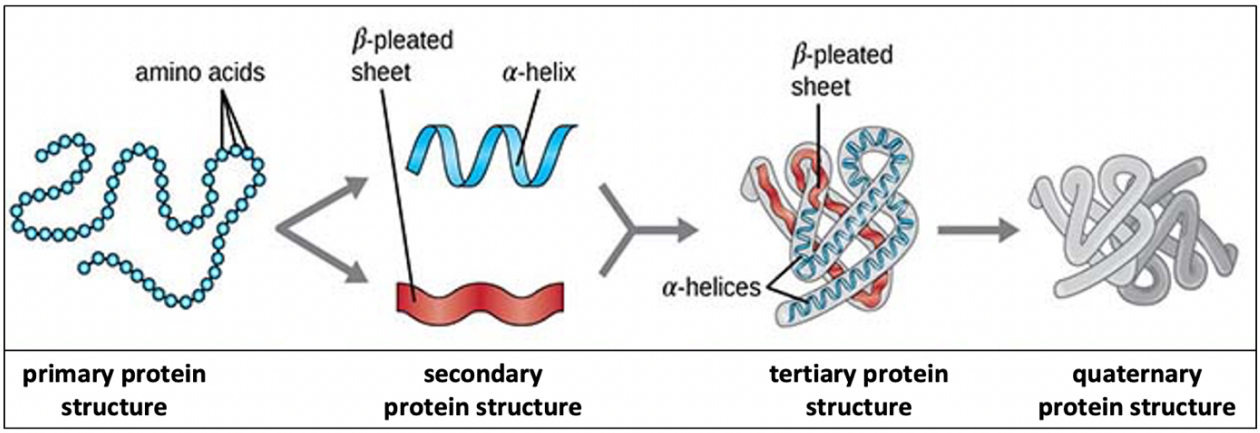

Figure 1.2. The four levels of hierarchy in protein conformation. From left to right: the primary protein structure (the sequence of a chain of amino acids), the secondary protein structure (local folding of the polypeptide chain into helices or sheets), the tertiary protein structure (3-dimentional folding pattern of a protein due to side chain interactions) and, the quaternary protein structure (assembly consisting of more than one amino acid chain). Image source: https://microbenotes.com

1.2 Domains, motifs, binding sites, catalytic sites

Structural motifs are short, conserved segments of protein 3D structure, which are spatially close but not necessarily adjacent in the sequence. Structural domains are globular structural units which are independently stable tertiary structures and which tend to have their own function (3). In enzymes, substrates often bind to a small pocket (active site) on the tertiary structure. The regions that form non-covalent bonds with the substrate and catalyse a reaction of a substrate are called the binding site and the catalytic site, respectively (4) (Figure 1.3).

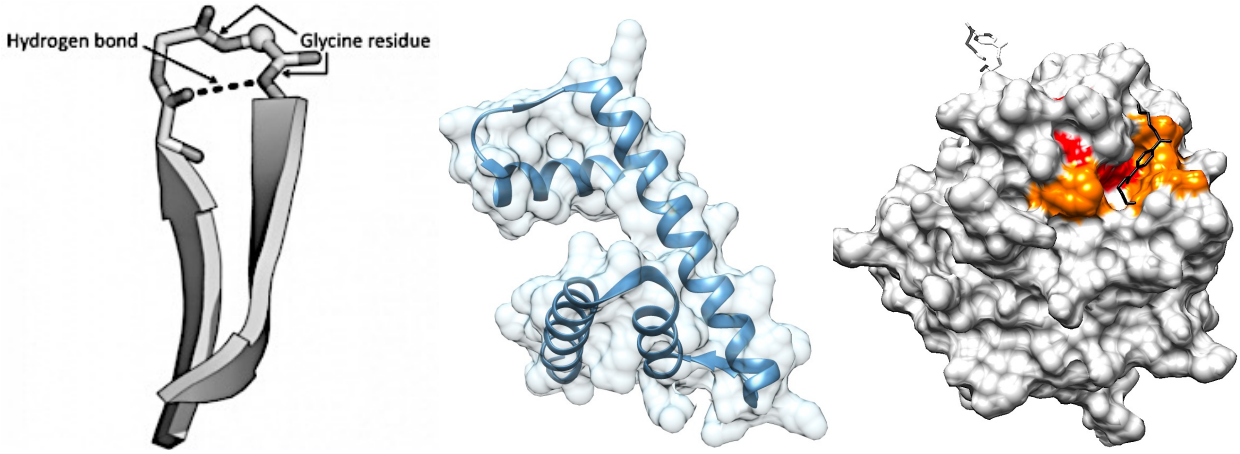

Figure 1.3. Illustration of motif and domain. Left) The beta-turn structural motif consists of four consecutive residues, where the polypeptide chain folds back on itself by nearly 180 degrees. Middle) Ribbon structure of an all α-helical domain in HIST1H2B (PDBe: 5KGF). Right) The binding pocket (orange) and catalytic site (red) shown with BHET as a substrate in *Is*PETase (PDBe: 5XJH).

1.3 Evolvability

Evolvability is the emergence of change in the sequence and structure of a protein, which leads to changes in the function. Mutational robustness is a mechanism that all organisms possess to support beneficial mutations and acquire new variations favoured by natural selection. This phenomenon is described as divergent evolution. Moreover, as a result of mutations, enzymes from distinct organisms can sometimes evolve to catalyse identical reactions. This can give rise to proteins having a variety of structures and sequences across different host species which are functionally the same. This phenomenon is described as convergent evolution (5).

Homologous proteins originate from a common ancestor (6) and can be categorised as orthologues or paralogs. Orthologous proteins, typically but not always exhibit highly similar functions, and arise from speciation events. Conversely, paralogues emerge through gene duplication within the same genome, often acquiring modified functions, particularly within catalytic or substrate-binding sites (7).

The evolution of the globin family serves as an example of divergence through paralogy. Haemoglobin, a tetrameric protein composed of four subunits: two α-globin and two β-globin (which originated via tandem duplication of a single-copy globin gene in the common ancestor of vertebrates), is responsible for oxygen transport in red blood cells (8,9). Myoglobin, which stores oxygen in muscle tissues, also belongs to the globin family. It is hypothesised that both haemoglobin and myoglobin evolved from a single ancestral globin gene encoding a protein with an oxygen transport function. Over time, gene duplication events followed by accumulation of mutations modified the structure and function of the proteins, resulting in different oligomeric states and properties such as oxygen affinity, kinetics, and stability compared to the original ancestral protein. These changes likely conferred selective advantages, aligning with environmental pressures such as variations in oxygen availability and metabolic demands (10,11).

1.4 Methodological focus and tool selection

Given the breadth of bioinformatics as a discipline—encompassing a wide array of algorithms, analytical frameworks, and software tools—this chapter focuses exclusively on the methods that were directly employed in the present study. Rather than offering a comprehensive overview of all available solutions, the discussion is limited to those tools selected based on their suitability for the specific biological questions and data types addressed.

1.5 Bioinformatics methods to analyse protein sequences similarity

Protein homology and functional prediction often rely on comparing sequences with known proteins using bioinformatics tools. These approaches include pairwise alignment,

dynamic programming, and multiple sequence alignment. Such methods help identify conserved regions and infer protein function based on sequence similarity. For example, Multiple Sequence Alignment (MSA) aligns three or more sequences to reveal conserved regions, helping to infer structural, functional, or evolutionary relationships. It is essential for identifying conserved motifs and domains within protein families. MSA results often serve as input for phylogenetic analysis or structural modelling. This method was used in Chapter 2 to identify key functional residues in the metagenomic sequences by comparing them with the ABH CATH superfamily and known PETases. In addition to Chapter 2, this method was also used in Chapters 3 and 4 within relevant CATH functional families to run Scorecons and identify conserved positions in the proteins of interest. Techniques such as Multiple Sequence Alignment with Fast Fourier Transform (MAFFT), Hidden Markov Models, and HH-suite were employed in one or more of the research projects presented in this thesis.

1.5.1 Multiple Alignment using Fast Fourier Transform (MAFFT)

MAFFT (12) is a multiple sequence alignment program that incorporates two key techniques: the use of the Fast Fourier Transform (FFT) and a simplified scoring system. In the FFT technique, amino acids are converted into numerical values based on the BLOSUM (13), which allows for the swift identification of homologous regions between the sequences. The simplified scoring system helps reduce CPU time while maintaining the accuracy of sequence alignments, even in the presence of large insertions or extensions.

MAFFT employs five main steps: pairwise alignment, distance matrix calculation, guide tree construction, progressive alignment, and iterative refinement. The procedure begins

with pairwise alignment using dynamic programming to identify similar regions among all input sequences. The next step involves calculating the distance matrix based on the pairwise alignment scores. Using the distance matrix, a guide tree is constructed, representing hierarchical clusters of sequences. The guide tree facilitates the progressive alignment process, where sequences are aligned incrementally from the leaves to the root, resulting in the final multiple sequence alignment. This alignment is further refined through iterative alignment, where gaps and insertions are adjusted to enhance accuracy (14,15).

1.5.2 Hidden Markov Models (HMMs)

HMMs are algorithms that align sequences by detecting the pattern of observable events that are dependent on hidden internal factors (16). HMMs detects and aligns homologues in the same manner as a multiple sequence alignment but the higher sensitivity of this algorithm over sequence profile based algorithms is due to the calculation of position-specific insertion and deletions probabilities along the alignment (17).

An HMM first calculates the probability for the occurrence of each amino acid along the multiple alignment, then identifies the residues that are more conserved and important for defining members of the family and finally, finds the regions of the sequence where insertions and deletions are likely to occur as different residues in a protein sequence are subject to different selective pressures (18). Figure 1.4 illustrates the HMM algorithm in detail. HMMer software identifies columns of conserved residues between multiple sequences (16,19).

Entry	Consensus					
	1	2	-	3	4	
Sequence 1	A	G	-	L	D	
Sequence 2	N	G	G	F	D	(insertion)
Sequence 3	S	G	-	-	E	(deletion)
Sequence 4	T	G	-	W	Q	

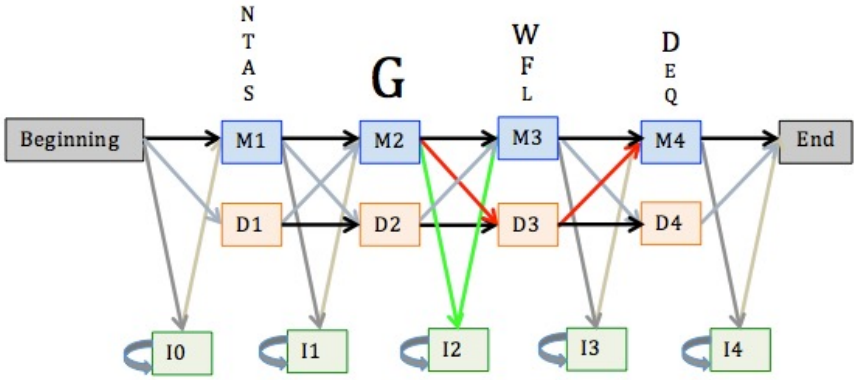

Figure 1.4**.** A schematic model of a profile HMM. The multiple alignment of four sequences is displayed (top figure) and is used to build a profile HMM (bottom figure) by assigning consensus columns in the MSA. An insertion in sequence 2 is shown with green arrows going into the insert state I2 then it goes back into the match position M3. A deletion in sequence 3 is shown with red arrows where M3 consensus is skipped (The figure is adapted from https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/).

In this example, the first consensus column, match 1, consists of small amino acids with different occurrence probabilities. The height of the letters above the consensus column represents how frequently these amino acids are seen. The second consensus column, match 2, is invariant across four sequences. So, glycine is expected with a high probability. In sequence 2, one amino acid is inserted into a sequence between the 2nd and 3rd consensus column. This is considered as an insert because the occurrence probability is less than 50% among the four sequences. In sequence 3, the consensus column three is skipped which indicates a deletion since there is greater than 50% occupancy in that column in other sequences. At M4, even though there are three different

amino acids present, they are all polar charged amino acids. Aspartic acid comprises 50% of the height, while the other two each make up 25%.

HMMs also calculate the frequency of amino acids and convert these into probabilities. There are many different search programs implemented based on HMM approaches (18). The most widely used ones are:

- hmmscan (20) is a tool from the HMMER software suite (http://hmmer.org). This tool searches a sequence database using profile HMM libraries, such as CATH-Gene3D, Pfam and TIGRFAMs, to detect homologous sequences and identify potential family relationships within the query sequence.

- hmmsearch (20) is also a tool from the HMMER software suite. The method involves the use of a profile HMM for capturing patterns of conserved regions, insertions, and deletions. By aligning each sequence in the database against the profile HMM, hmmsearch identifies protein sequences that likely belong to the protein family or domain represented by the HMM. This method is used to find homologous proteins to the family represented by the HMM. Either hmmsearch or hmmscan can compare a set of profiles to a set of sequences. The only difference between the two tools is speed. While both programs are input-bound, hmmsearch loads less data due to its disk access patterns, making it the faster of the two tools. (http://hmmer.org, Version 3.4; Aug 2023).

- JackHMMER (20) is a tool within the HMMER suite (http://hmmer.org). JackHMMER is used for iterative sequence searching and profile building. The method begins with

a single protein sequence, which is iteratively searched against a sequence database, such as UniProt. In each iteration, JackHMMER identifies sequences that align well with the initial query and incorporates them into the profile hidden Markov model (HMM). This iterative process refines the HMM by progressively including more remote homologs with each round, enhancing the ability to detect distantly related sequences. This iterative refinement facilitates the discovery of additional protein family members and provides insights into deeper evolutionary relationships (19). In addition to their use in building a hierarchical relationship tree in GeMMA, HMMs were frequently used in Chapter 1 to construct protein profiles aimed at identifying potential metagenomic protein sequences with PETase activity.

1.5.3 HH-suite

HH-suite is an on online and widely used open-source software for sensitive sequence similarity searches to detect distant homologues and predict the function of an unknown protein based on the functions of proteins with similar sequences. It is based on the pairwise alignment of profile HMMs derived from multiple sequence alignments of homologous proteins. The HH-suite software contains in-built tools such as HHsearch which aligns a profile HMM against a database of target profile HMMs (18) and HHblits: an accelerated version of HHsearch that is fast enough to perform iterative searches through millions of profile HMMs (21) and various utilities to build databases of MSAs or profile HMMs. HH-suite tools were frequently used in Chapter 1 to perform profile–profile comparisons, to identify metagenomic proteins that may exhibit PETase-like functionality.

1.5.4 Cluster Database at High Identity with Tolerance (CD-HIT)

Cluster Database at High Identity with Tolerance (CD-HIT) clusters large protein sequence databases to reduce redundancy. This helps in analysing large datasets by selecting representative sequences of each cluster rather than the whole sequence database. Sequences are first sorted by length to find the longest sequences for rapid identification of representative sequences or seed. Subsequent sequences are compared against these seed sequences using a sliding window approach combined with a k-mer based algorithm. If a sequence meets the specified identity threshold defined by the user with any existing cluster representative, it is added to that cluster; otherwise, it will be the seed for a new cluster. A higher threshold results in clusters with high similarity, while a lower threshold allows for more diverse sequences within each cluster. By setting an appropriate threshold, CD-HIT significantly reduces redundancy, representing each cluster with usually the longest sequence. This reduction facilitates downstream analyses such as database searches by decreasing computational time (22,23). CD-HIT was used in the first step of GeMMA, a CATH algorithm for functional family classification, to cluster sequences based on sequence identity and reduce redundancy in the dataset.

1.6 Quantifying residue conservation

Conserved residues within proteins *e.g.*, preserved through evolution, are often found in critical regions of the protein necessary for proper folding, stability, and interaction with other molecules due to their functional importance (24). Many algorithms have been developed to identify conserved regions. They all involve analysis of residues in the columns of a multiple sequence alignment to detect highly conserved positions. In this study the Scorecons algorithm has been used to detect conserved residues and hence is described below.

1.6.1 Scorecons

Scorecons (25) calculates the degree of amino acid variability in each column of a multiple alignment sequence. It uses a statistical model to assign conservation scores to individual residues based on observed and expected residue frequencies in a multiple-sequence alignment. Scorecons assesses the significance of amino acid substitutions and identifies conserved residues likely to be functionally important. The method also incorporates the existence of gaps in calculating the final conservation score. A column with frequent gaps in a multiple sequence alignment indicates that deletion of that position has a negligible adverse effect on protein function. Scorecons scores range from 0 to 1. A low score means high variability of amino acids in that position indicating mutations at that position are unlikely to impact the overall function or structure of the protein. In contrast, a high score indicates low variability of various amino acids in a certain position indicating that the position is highly conserved among the homologous sequences, probably due to the functional or structural role required at that position. Scorecons provides insights into the evolutionary and functional importance of individual positions in a protein sequence. Residues with scores ≥0.7 are considered to be conserved (26). Scorecons was used in all three chapters to identify conserved residues within protein alignments, highlighting functionally or structurally important positions.

1.6.2 Diversity of positions (DOPs) score

The DOPs score is a metric employed by Scorecons that quantifies the diversity within a multiple sequence alignment by computing the average of conservation scores across all positions, taking into account the various conservation scores and their respective frequencies. The result is a value between zero for no diversity and 100 for high variability

(no alignment positions that have the same conservation score). It is used to ensure that sequences in a multiple sequence alignment are sufficiently diverse for the individual residue conservation scores to be meaningful. A multiple sequence alignment with a DOPs score of at least 70 is considered suitably diverse for identifying conserved residues (27). The DOPs score was used in all three chapters to assess the diversity within multiple sequence alignments, ensuring that the alignments were sufficiently diverse for meaningful residue conservation analysis.

1.6.3 Identification of specificity determining positions in a protein family

Proteins can gain new functions following gene duplication followed by specialisation of the paralogues by divergence of residue positions (Kimura, 1979). Therefore, proteins in a homologous family can be clustered into subgroups, called functional families (Das, et all, Mirny & Gelfand, 2002), if they differ to other relatives in the family *e.g.*, due to a different substrate, ligand or protein interaction. Algorithms to detect Specificity Determining Positions (SDPs) identify these functional groups within a protein family.

There are a number of SDP identification algorithms such as GroupSim (Capra & Singh, 2008) and SPEER (Chakraborty, Mandloi, Lanczycki, Panchenko, & Chakrabarti, 2012) with different scoring methods but all follow the same concept of identifying the conserved residues within subfamily specificity groups that are different between subgroups. GroupSim distinguishes the difference between specificity groups by employing a baseline for the comparison method that includes all pairs of amino acids within and between groups. First, GroupSim calculates the mean similarity between each pair of equivalent amino acids in a group based on a similarity matrix in the alignment. Next, the

average similarity of every residue between the groups is calculated. Each column is then given a score, which is the average within-group similarity minus the average between-group similarity. Higher scores indicate a greater probability to be an SDP (Capra & Singh, 2008).

GroupSim implements sequence alignment-based filtering levels to exclude residues that are conserved across all the specificity groups due to their functional importance to the whole family. These filters determine the level of conservation in a subfamily group and whether the residues are conserved in a similar way in the second group. An example is defined in figure 1.5. In the creation of CATH functional families, FunFamer uses GroupSim to identify differentially conserved residues between clusters.

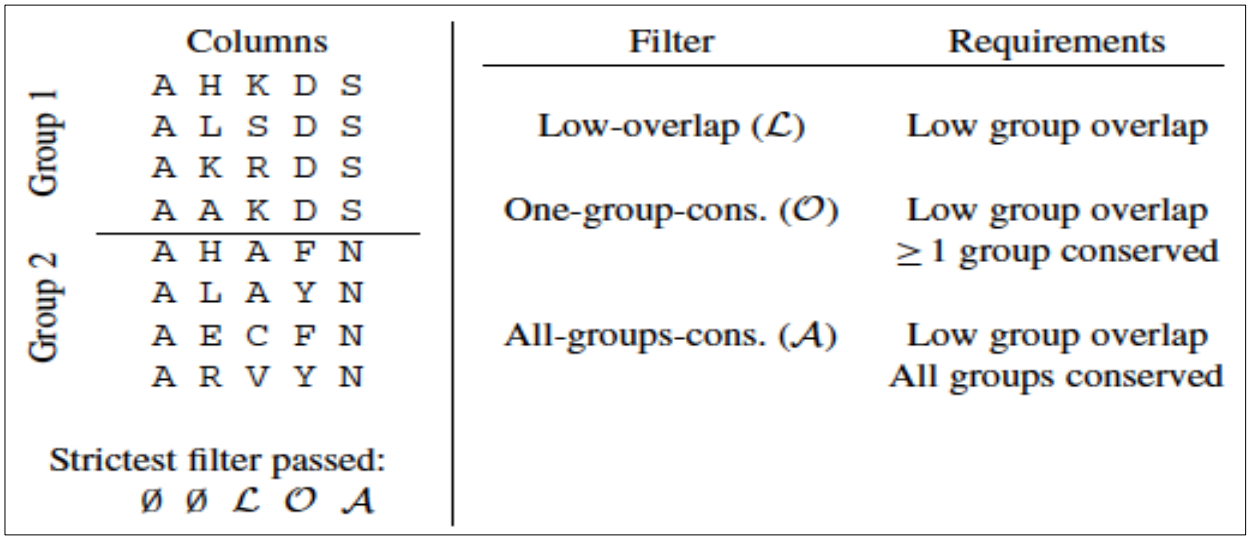

Figure 1.5. Sequence alignment-based filtering of columns by GroupSim (Capra & Singh, 2008). Columns from left to right: the first column is conserved across all specificity groups and therefore excluded (ø). There is no conserved residue between and within specificity groups, so the second column is removed (ø). The third column passes the low-overlap filter as 'K' and 'A' are conserved within group one and two, respectively (*L*). The fourth column passes the low-overlap and the one-group-conserved filter as 'D' is conserved across all sequences in group one while 'F' and 'Y' are conserved in group two sequences (*O*). The fifth column is the best example of an SDP as one amino acid within each group is completely conserved and it differs from the other subgroup (*A*).

1.7 Resources for protein classification

Protein classification is the process of grouping proteins based on their structure, function or evolutionary relationships. This classification facilitates the understanding of how protein function has evolved in protein families, as well as predicting the roles of newly discovered proteins that lack experimental data.

1.7.1 Pfam

Pfam is an open-access database of protein domain families that classifies protein sequences into families (28–30). Each Pfam family contains a set of sequences that share a conserved region, which usually corresponds to a functional or structural domain. A seed multiple sequence alignment built from a representative set of sequences in the corresponding Pfam family is the basis for creating a profile hidden Markov model (HMM) using the HMMER software (31). This profile HMM captures the conserved regions and features of the family and is then queried against the *pfamseq*, sequence database, via the HMMER software (http://hmmer.org/). The *pfamseq* sequence database is derived from the Reference Proteomes in the UniProt Knowledgebase (UniProtKB) Bateman et al. 2025). All sequence regions that meet the curated threshold are re-aligned to the profile HMM to build the full alignment. Sequences in Pfam families are annotated with available experimental data. Pfam (through InterPro) release 102 contains 26,073 families.

1.7.2 CATH

Since the protein structure is known to be highly conserved during evolution (32), structural comparison of proteins can be used to identify homologous proteins.

CATH (33) is a hierarchical protein domain classification database comprising three-dimensional (3D) structures of proteins deposited in the PDB. CATH also classifies

predicted protein domain sequences from UniProt (without experimental structures) in the CATH-Gene3D sister resources (34), in specific CATH superfamilies and incorporates supplementary functional annotations from Gene Ontology (GO), Enzyme Classification (EC) and catalytic sites (from CSA). CATH classifies protein domains into four major hierarchical categories as stated below and shown in Figure 1.6.

- Class (C-level) refers to the secondary structure content. This level classifies protein domains based on their prominent secondary-structure composition into three major classes: mainly α (Class 1), mainly β (Class 2), α/β and α + β (Class 3). Domains with few or no secondary structures (Class 4), and multi-domain proteins (Class 5) (35,36) are also grouped together in the classification.
- Architecture (A-level) refers to the general arrangement of the secondary structures in 3D irrespective of connectivity between them (*e.g.*, alpha/beta sandwich). There are currently 41 architecture groups (37).
- Topology (T-level or fold) refers to a group of domains that share similar structures and connectivity of secondary-structure elements. This level comprises 1390 topology groups.
- Homologous Superfamily (H-level) refers to a set of domains with significant structural or sequence similarity that share a conserved structural core, likely diverging from a common ancestor. There are 5,481 homologous superfamilies in groups 1 to 4 in CATH version 4.3 (38). Sub-clusters of homologous superfamilies are grouped into Functional Families (FunFams) based on predicted functional similarity (39).

Every domain receives a unique identifier that specifies its classification across multiple hierarchical levels. For example, in 3.40.50.620, the 3 refers to the class (mixed alpha-beta), the 40 refers to the architecture (3-layer alpha-beta-alpha sandwich), the 50 refers

to the topology the domain adopts (Rossmann fold) and 620 is the homologous superfamily code (HUPs).

CATH contains high-quality protein structures from the PDB (X-ray crystallography or Nuclear Magnetic Resonance (NMR) resolution equal to or below 3Å resolution (40), polypeptide length over 40 amino acids. Chop-Close (41), an automated process in Auto Chop, uses the Secondary Structure Alignment Program (SSAP) algorithm (35,42) to scan the query structure against the protein structures that have been classified into their constituent domains in CATH. If the query matches any CATH domain structures, the alignment induces the dissection of the query chain; otherwise, a manual curation procedure is applied to identify domains with the aid of several software tools including DETECTIVE, PUU and DOMAK (43,44) and information reported in the scientific literature. According to the latest update in May 2024, CATH version 4.3 consists of 536,613 structural domains classified into 6631 homologous superfamilies (Figure 1.6).

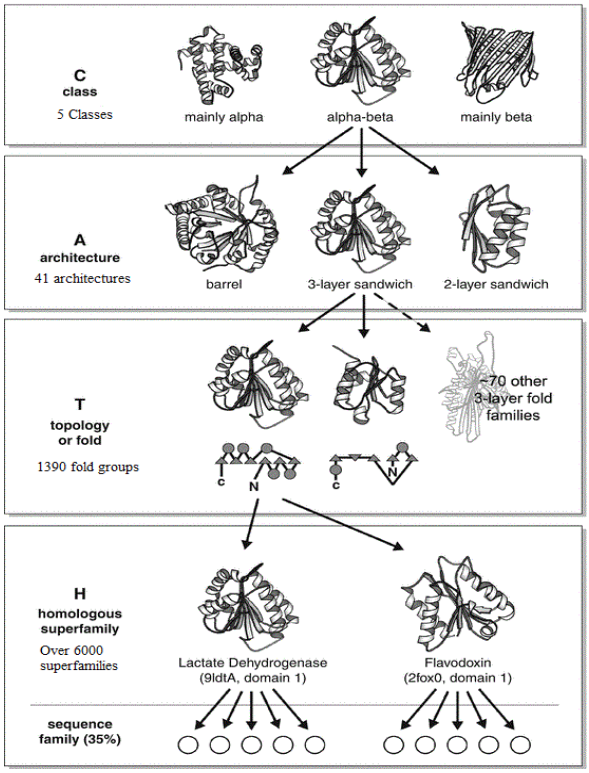

Figure 1.6. An example CATH classification of two proteins: lactate dehydrogenase and flavodoxin, according to Class, Architecture, Topology and Homology levels. Both lactate dehydrogenase and flavodoxin proteins are described as alpha-beta proteins as they are composed of both alpha helices and beta sheets (Class level). Their arrangement of secondary structures forms an alpha-beta sandwich architecture in three-dimensional space. Whilst adopting similar core folds (Topology level) they belong to different evolutionary superfamilies.

Gene3D (34), a complementary platform to CATH, assigns protein domain sequences to CATH superfamilies by using sequences from CATH domain structures to build domain superfamily-specific profile HMMs. These HMMs are then employed to identify domains in structurally uncharacterised protein sequences in UniProtKB Bateman et al. 2025) and Ensembl (45). There are currently 82,665,384 protein sequences and 151,013,797 CATH domain predictions in Gene3D v21.

1.7.2.1 Sub-classification of CATH evolutionary superfamilies into functional families

Protein superfamilies can have diverse functional relatives as a result of their evolutionary pathways and structural flexibility. Proteins within a superfamily share a common ancestor, similar folding arrangements, and a conserved structural framework. However, over time, small changes in their amino acid sequences, particularly at the binding pocket, can lead to different biochemical functions. This process, known as divergent evolution, allows proteins to adopt new roles by preserving the core structure while undergoing modifications, resulting in superfamilies with diverse functional families. CATH superfamilies are therefore subclassified into functional families (FunFams) in which relatives share significant structural and functional similarity. FunFams are identified in a two-step process. The Genome Modelling and Model Annotation (GeMMA) algorithm first performs agglomerative clustering to group relatives in a hierarchical tree. This is subsequently analysed by the FunFam algorithm to generate FunFams.

GeMMA is an automated method for classifying functional subfamilies within protein superfamilies. A key advantage of GeMMA is its capability to subclassify extremely large and diverse superfamilies without requiring an initial multiple sequence alignment. The GeMMA algorithm (Lee, Rentzsch, & Orengo, 2010) generates a hierarchical tree of sequence relationships across a superfamily by first assembling the leaves and then working inwards towards the trunk. One GeMMA run comprises the following steps (Figure 1.7):

1. CD-HIT is employed to generate a set of non-redundant representative sequences by clustering relatives at 90% identity (S90), referring to the nodes on the far left in figure 1.7.
2. MAFFT aligns the S90 non-redundant representative sequences. As GeMMA is computationally expensive, S90 clusters that, according to the UniProt Bateman et al. 2025) have no experimental GO annotations are discarded.
3. Profile hidden Markov model (profile HMMer) method (46) converts the given multiple sequence alignment into a position-specific scoring system (sequence profile). The sequence profile specifies a preference for the 20 standard amino acid residue types at each of the residue positions in the given multiple sequence alignment (47).
4. Profile HMM-HMM comparison using the HH-align (48) method to compare two adjacent nodes against one another to detect similar patterns of conservation and preferred residues for each position in the alignment. All the cluster HMMs are compared against each other using this approach.
5. The two nodes with the highest HH-align score *e.g.*, the most similar patterns, are merged.
6. Once merged, all the sequences from the two merged groups are combined and the sequence alignment of sequences in the merged nodes regenerated using MAFFT. A revised HMM is generated for the merged set of sequences. This cycle repeats until no new nodes can be merged. This gives a tree of sequence relationships across the superfamily.
7. Finally, CATH superfamilies are subclassified into FunFams by cutting the tree into subfamilies using the FunFamer algorithm (described in the next section).

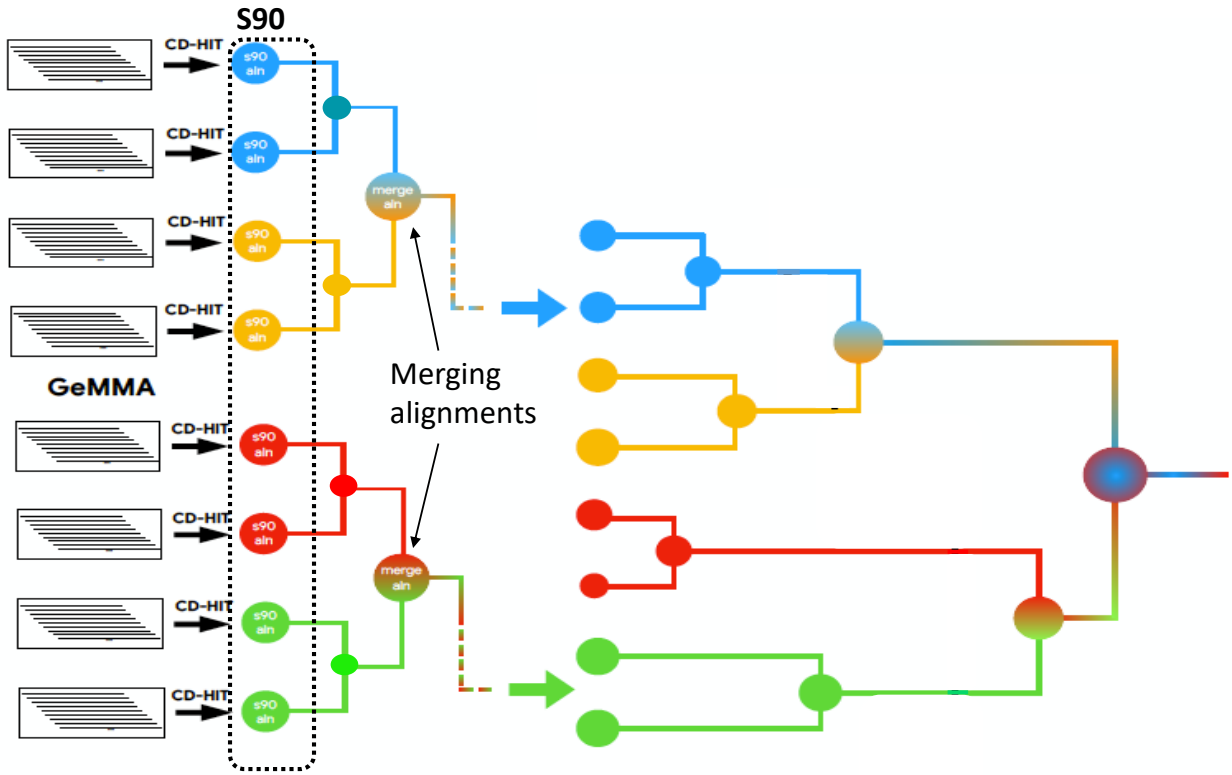

Figure 1.7. The basic GeMMA clustering pipeline algorithm.

Two different versions of GeMMA have been developed, each with different features and abilities for different purposes: the basic algorithm is called Full-Scale GeMMA (FS-GeMMA) and the faster high-throughput version is known as HT-GeMMA.

FS-GeMMA algorithm starts with an iteration of all-against-all profile-profile comparison of a set of sequence clusters. Next, it merges the highly similar clusters and realigns the newly merged clusters. Finally, a FASTA format of newly clustered subfamilies sequences and a hierarchical tree are generated. HT-GeMMA and FS-GeMMA both follow the same approach in merging similar clusters depending on the *E*-value returned from their comparison. Clusters are merged with an increasing *E*-value, for example, *E*-value cut-off from 10^{-80} for the first iteration to 10^{-30} for the last iteration. However, for analysing a large superfamily of protein domains, HT-GeMMA is employed. The main difference between

the two versions of GeMMA is that HT-GeMMA exploits multiple nodes on a compute cluster. This means that HT-GeMMA is able to analyse a large superfamily with up to 50,000 sequences much faster.

1.7.2.2 FunFamer

The FunFamer algorithm (Das et al., 2015) determines an optimal cut of the hierarchical clustering tree of sequence relatives within a given superfamily produced by the clustering algorithm, GeMMA (Lee et al., 2010). The FunFamer algorithm evaluates Diversity of Position Scores (DOPS) for multiple sequence alignments using Scorecons (Valdar, 2002), which identifies conserved positions across the superfamily. The DOPS score is 0 if all positions in the alignment have the same conservation score, and 100 if no two positions share the same conservation score. The FunFamer algorithm considers alignments with a DOPS>70 as sufficiently diverse (25). It then uses GroupSim (Capra & Singh, 2008) to predict specificity-determining positions (SDP) in a given cluster. While the conserved residues are important for the stability and folding of the protein domain, SDPs are important for function. SDP residues are conserved within a functional family but differ between the two functional families. GroupSim predicts SDP for each position by comparing two sets of MSA, one for each putative FunFam and gives a score (G_s) range between 0 to 1. Positions in the top 30% of G_s range in an alignment are considered to be SDPs (Capra & Singh, 2008). FunFamer uses the identification of differentially conserved residues between two FunFams to determine whether to merge the FunFams into a single FunFam (Figure 1.8).

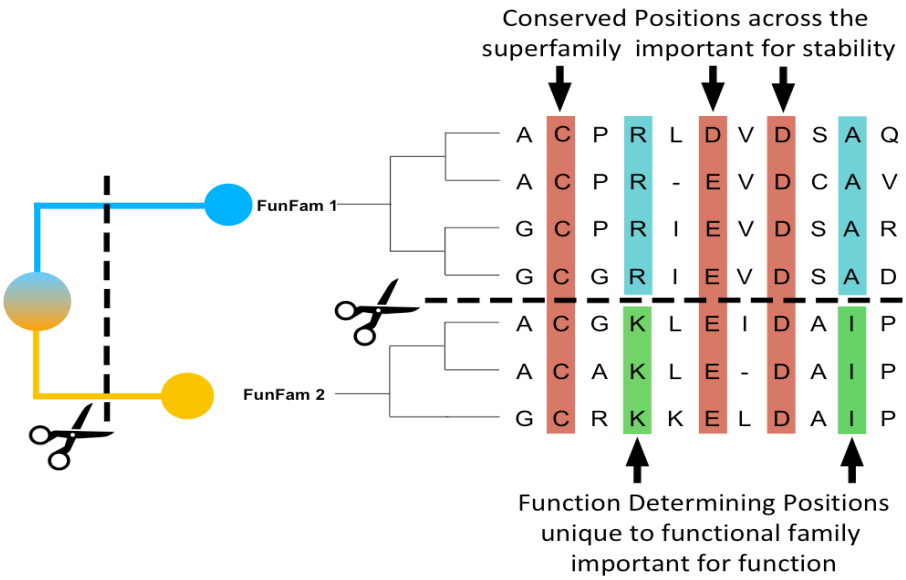

Figure 1.8. The FunFamer algorithm determines an optimal cut of the tree by incorporating Scorecons and GroupSim algorithms (Das et al., 2015).

Various methodologies of the GeMMA algorithm were introduced to accelerate the generation of the tree of relationships across the superfamily (*e.g.*, MARC and FRAN).

1.7.2.3 FunFam generation by Multidomain Architecture-based Clustering

FunFam generation by Multidomain Architecture-based Clustering (FunFam-MARC) algorithm (49) is a modification of GeMMA for superfamilies that are very large, with more than 2000 S90 clusters. FunFam-MARC significantly reduced the analysis time for the 5,481 superfamilies in CATH version 4.3 from six months to only six weeks.

The algorithm begins by partitioning the set of protein functional unit sequences into smaller sets with the same Multi-Domain Architectures (MDAs) with domains following the same order in the protein sequence. The process of constructing MDAs, which involves determining the order of domains along the protein sequence including the CATH domain superfamily being classified and additional domain partners, is accomplished using the

CATH resolve-hits protocol (50). CATH resolve-hits employs an optimisation algorithm to resolve matches to the CATH HMM libraries, resulting in a set of non-overlapping domain annotations for the sequence (Figure 1.9).

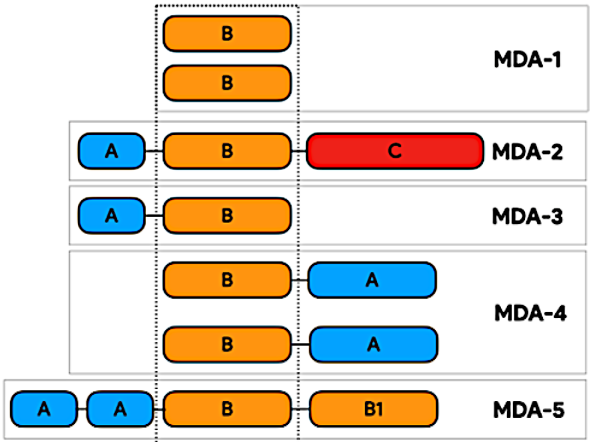

Figure 1.9. Multi-Domain Architecture (MDA) construction (49). Different MDAs reflect different domain contexts for the CATH domain superfamily (B) being sub-classified.

CD-HIT (22) clusters sequences within each MDA partition into 90% sequence identity clusters (S90). Annotated clusters are then used as the starting point (51) for a GeMMA sub-clustering of each MDA protein.

When all MDAs are processed, FunFam clusters from each MDA partition are combined to form the starting clusters for another run of GeMMA and FunFamer. The final FunFams are identified following the ultimate FunFamer algorithm run (Figure 1.10). Finally, the sequences from the uncharacterised experimental S90 clusters, which were excluded in the early stage, are added.

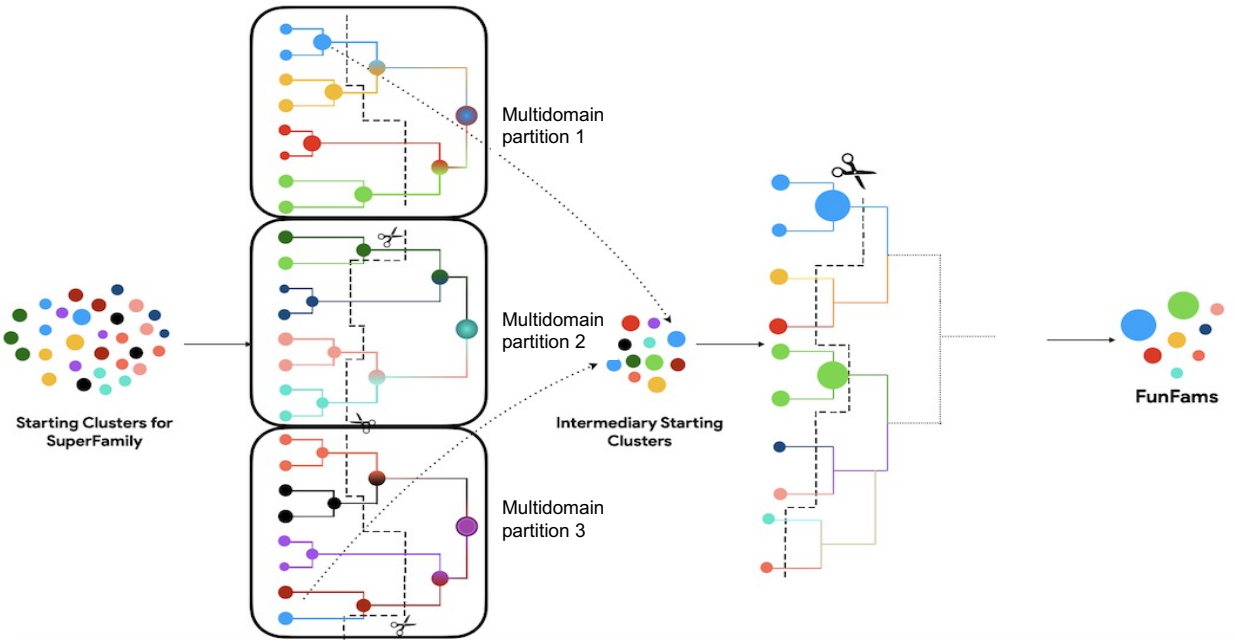

Figure 1.10. A schematic pipeline of FunFam-MARC, multiple parallel and separate GeMMA runs (49).

1.8 MGnify: resource for storing and assembling metagenomic samples

Metagenomics is a newly emerged field of molecular biology to analyse the genetic material of microorganisms acquired from a particular environment (biome) such as the abyssal waters and sediments of the oceans (Liu et al., 2019), ice and soil from the mountains (S. Kumar, Suyal, Yadav, Shouche, & Goel, 2019; Simon, Wiezer, Strittmatter, & Daniel, 2009). Unlike traditional microbiology methods that require pure cultures, metagenomics provides high-quality genomic datasets and can discover novel genes and proteins (metaproteomics) of un-cultivatable microorganisms. For example, by analysing protein sequences directly extracted from environmental samples, metagenomics has enabled the discovery of diverse microbial genes encoding PET-degrading enzymes and facilitates understanding of the genetic diversity in bacteria involved in PET degradation in natural ecosystems (52,53).

MGnify (Mitchell et al., 2020), formerly known as EBI Metagenomics, provides a series of tools for the assembly, analysis, exploration and archiving of microbiome sequencing data. MGnify in collaboration with the European Nucleotide Archive (ENA) is able to provide taxonomic and functional analysis of microbiome sequence data and provides a repository of analysed microbiome data from similar environments. Raw metagenomic reads are one of many data types that MGnify analyses and provides. Figure 1.11 outlines key steps in transforming raw reads from environmental samples into meaningful biological information.

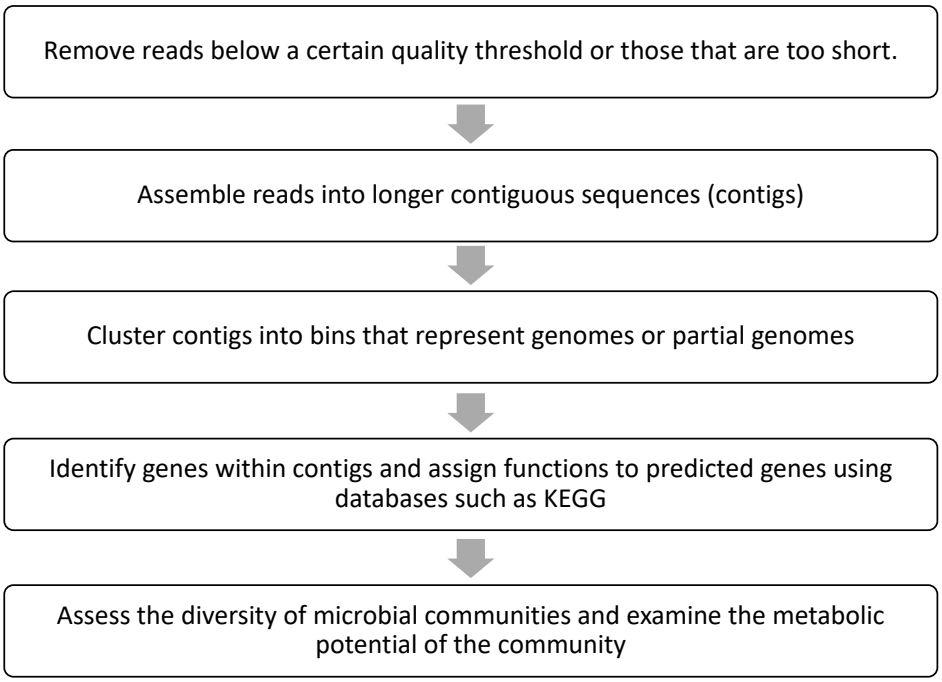

Figure 1.11. Procedure for transforming raw reads from environmental samples into meaningful biological information (54).

1.9 Protein structure prediction

Understanding the three-dimensional structures of proteins facilitates comprehension of their molecular functions and of their quaternary structure and interactions with other molecules, elucidating their relevance in biological systems (55,56). Despite continuous

advancements in structure determination technologies, such as Electron Microscopy, X-ray crystallography, or NMR spectroscopy (57,58), the rate of experimentally determined novel complex structures remains significantly lower than known sequences. However, protein structures can be predicted using a variety of approaches *e.g.*, homology modelling. This leverages the structure of a related protein (59). Recently, the development of deep learning-based methods such as AlphaFold2, from DeepMind (1) has revolutionised protein structure prediction providing 3D-models which are often of comparable quality to those determined experimentally.

1.9.1 Homology modelling

Homology modelling involves aligning the target protein sequence with the template protein sequence and applying this alignment to construct a model of the target protein's structure. The higher the similarity between the target and template sequences, the more accurate the predicted model. Several homology modelling techniques have been developed, including MODELLER (60), SWISS-MODEL (59,61), I-TASSER (62). This approach has been widely used to obtain accurate models for proteins with a close homologue (≥30% sequence identity) which has an experimentally determined structure that can be used as the template.

1.9.2 SWISS-MODEL

SWISS-MODEL (59,61) is a widely used web-based tool for protein structure homology modelling. The query sequence in the FASTA format can be uploaded onto the server's modelling workplace to build the most suitable template using BLAST and HHblits in which the selected template will be provided with cross-references to structural databases such

as PDB via the link to the SWISS-MODEL library. The search results consist of a list of 50 templates sorted in descending order according to a template score.

Before the advent of high-quality predicted structure by AlphaFold, the best model was selected according to the following criteria (importance descending order):

1. Coverage: coverage is represented as a percentage and indicates how much of the sequence of interest is covered by the chosen template.
2. Identity and resolution: higher scores for both properties would generate a better model.

SWISS-MODEL uses two quality assessment metrics to evaluate the accuracy of predicted protein structures:

1. GMQE (Global Model Quality Estimation) scores the expected quality of the model between 0 and 1 where the greater number indicates higher quality of the model. If AlphaFold DB is used as a template for building a model, GMQE is calculated by summing the per-residue plDDT values of the aligned template residues and normalizing by the target sequence length.
2. QMEAN Z-scores (63,64) calculates the degree of nativeness of the built model based on multiple factors such as:
    - Interaction potential between beta carbons or all atoms
    - The solvation potential estimates the solubility of the protein
    - The torsion angles for each amino acid and the neighbouring residues based on the Ramachandran plot.

QMEAN Z-scores range starts from zero to negative values. QMEAN Z-scores closer to

zero indicates a more reliable model and is comparable to what one would expect from experimental structures of similar size. No model with QMEAN Z-scores below -4.0 would be suitable for downstream analysis. The local quality plot gives a quick overview to find any low-quality regions of the model where predicted local similarity to target score is below 0.6 zone (the range is between 1 to 0). The QMEAN Z-scores is not computed for models that use AlphaFold DB templates.

The comparison plot helps to understand how good the model is in comparison to structures with a similar length from the PDB database. It is generated based on the normalised QMEAN Z-score (y-axis) against the experimental protein structures of similar sizes (x-axis) to represent an estimation of the quality of the model in the real life. The model will be more reliable and has similar properties to the experimentally determined PDB structures if it falls in the darker zone (Z-score will be between 0 to 1). More recently, SWISS-MODEL is selecting and displaying predicted structures from the AlphaFold database which allows users to access AlphaFold-predicted structures directly through the SWISS-MODEL interface.

Lastly, the PDB format of the selected model of interest can be downloaded to be used for further analysis (Figure 1.12). This tool was employed in the preliminary research to predict some metagenomic PET-degrading enzymes in Chapter 2, followed by the use of AlphaFold2 for structural prediction of all sequences in Round 2 in Chapter 2, and for human complexes in Chapter 3, where no experimental structures were available.

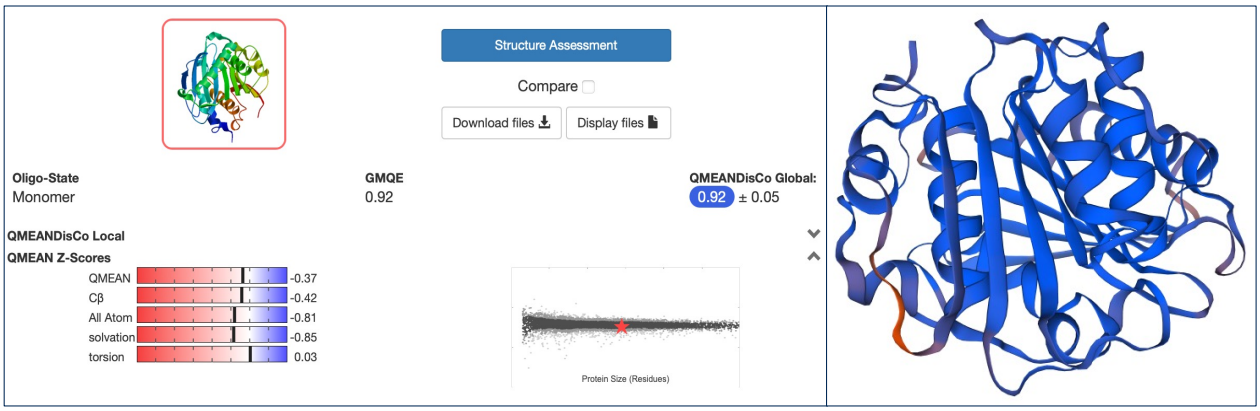

Figure 1.12. Building homology mode using SWISS-MODEL. QMEAN Z-score provides an estimation of the degree of nativeness of the structural features observed in the model on a global scale. The lower local quality of the model is coloured in red on the structure.

1.9.3 Deep learning-based modelling

Deep learning-based structure prediction predicts the three-dimensional structures of proteins from their amino acid sequences by using extensive protein sequence and structural data to learn patterns that indicate how proteins fold into their functional forms. For example, platforms like AlphaFold2 (1,65) and RoseTTAFold (66) have predicted highly accurate structures, often comparable to those obtained through experimental techniques such as X-ray crystallography or NMR spectroscopy. AlphaFold2 and RoseTTAFold use evolutionary information from multiple sequence alignments (MSAs). However, there is a limited amount of data for orphan proteins and rapidly evolving proteins, making this information less abundant than these methods require. AlphaFold2 also employs deep learning for structure prediction. Deep learning uses neural networks with multiple layers of nodes. The network is a collection of simulated nodes, linked by connections that can become stronger or weaker. RoseTTAFold integrates evolutionary information and structure modelling techniques.

AlphaFold2 was used to predict the structure of all proteins in UniProt release October 2024 lacking experimental three-dimensional structure for the analyses reported in chapters 2, 3 and 4.

1.9.4 AlphaFold2

AlphaFold2, developed by DeepMind, is a free and accessible platform for predicting protein structures with high accuracy using deep learning techniques. DeepMind collaborates with Google Colab (1) offering researchers a cloud-based platform that facilitates seamless access to powerful computational resources which exploit GPUs and TPUs to accelerate the complex calculations required for the protein structure predictions.

The process commences with the generation of multiple sequence alignments through tools like HHblits and JackHMMER, which extract evolutionary information critical for predicting the protein fold. Leveraging these alignments, AlphaFold2 predicts residue-residue contacts by using deep learning strategies to identify three-dimensional constraints crucial for guiding the folding process. The deep learning models, integrate convolutional and recurrent neural networks, to predict the three-dimensional coordinates of amino acids. These initial predictions undergo iterative refinement to improve accuracy, incorporating physical principles such as clashes and hydrogen bonding. AlphaFold2 validated its predictions against experimental data from the Protein Data Bank, optimising metrics such as Global Distance Test (GDT) and predicted local distance difference test (pLDDT) scores to ensure reliability.

GDT is a measure of global similarity between two protein structures with known amino acid correspondences. It compares the predicted protein structure with the experimentally

determined structure. The pLDDT is a per-residue estimate of local confidence, ranging from 0 (no confidence) to 100 (high confidence). A higher pLDDT score indicates that the region of the protein is modelled with high accuracy.

The Critical Assessment of Structural Prediction (CASP) is a competition platform for evaluating the accuracy of *in silico* protein structure modelling from amino acid sequences that have little to no similarity to existing structures. In December 2020, AlphaFold2 won the 14th CASP competition, significantly outperforming the other 146 participating groups due to its highly accurate predictions for 88 out of 92 targets. AlphaFold2 achieved a Global Distance Test Total Score (GDT_TS) Z-score of 244.0217, while the second-place group had a Z-score of 92.1241.

AlphaFold-Multimer (67) is an extension of AlphaFold2 developed to predict the structures of protein-protein complexes such as homomeric (proteins composed of identical subunits) and heteromeric (proteins composed of different subunits) complexes. To predict interactions between two proteins, AlphaFold-Multimer identifies these interactions and predicts the structure of the entire complex. Both AlphaFold and AlphaFold-Multimer use the same deep learning architecture. However, AlphaFold-Multimer incorporates an approach that selects subsets of residues for training, allowing the model to learn how protein chains interact with each other. Additionally, it introduces small adjustments to loss functions to refine the structure at the interface, enabling accurate interaction predictions between proteins while maintaining high intra-chain accuracy. This extension is accessible through the Google Colaboratory platform, and more information can be found on its GitHub repository at https://github.com/sokrypton/ColabFold. AlphaFold2 was used in Chapter 2 to build models of the selected putative metagenomic PETases, due to its

high accuracy in protein structure prediction, as demonstrated by its top-ranking performance in the CASP14 competition. In Chapter 3, both AlphaFold2 and AlphaFold-Multimer were employed to construct human–SARS-CoV-2 protein complexes, exploiting their capabilities for both monomeric and multimeric structure prediction.

1.10 Protein functional annotation

Protein functional annotation is the process of identifying the biochemical, biological, and cellular functions of proteins according to their activities and interactions within a biological context by using computational and experimental approaches (68). A number of resources have been established to collate and disseminate experimental and predicted data for protein functions. These are described below.

1.10.1 Enzyme commission number

The enzyme commission (EC) number provides a numerical hierarchical classification scheme of four levels describing an enzyme catalytic reaction (69). The first number (x.-.-.-) presents the enzyme class, the second number (-.x.-.-) refers to the type of bond that is acted on, the third (-.-.x.-) describes the details of reaction and the fourth number (-.-.-.x) refers to the substrate.

At the top EC level, the six main reactions are EC 1 for Oxidoreductase reactions, EC 2 for Transferase reactions, EC 3 for Hydrolase reactions, EC 4 for Lyase reactions, EC 5 for Isomerase reactions, and EC 6 for Ligase reactions (70,71).

The EC number links the genomic repertoires of enzyme genes or proteins to reactions in metabolic pathways (72,73). The Joint Commission on Biochemical Nomenclature of the

International Union of Biochemistry and Molecular Biology and the International Union of Pure and Applied Chemistry manually assigns EC numbers based on published articles reporting the full characterisation of enzymes (72,73).

One limitation of the EC system is that non-enzymatic proteins are excluded from this numeric system and since the EC number only classifies the enzymes' chemical-reactions, the classification does not provide the roles of a specific enzyme within a biological system (74). The EC number helped identify enzymes within the ABH CATH superfamily, enhancing our understanding of how similar enzymes catalyse PET degradation or other types of plastic hydrolysis.

1.10.2 Gene Ontology (GO)

GO (75) is a comprehensive framework that describes the roles of genes and gene products (*e.g.*, proteins) consistently across species and databases. The aim of the GO consortium is to standardise vocabulary for describing the gene and gene product across the multiplicity of species in the tree of life.

Each term in the ontology is assigned with a unique identifier and a definition. GO is structured as a hierarchical ontology, grouped into three ontologies (Figure 1.13):

- Molecular function: this level explains the activities of proteins at the molecular level, such as catalytic activities. (*e.g.*, kinase activity or DNA binding).
- Biological process: This category describes the biological pathway accomplished by one or more proteins (*e.g.*, cell cycle or response to stress).
- Cellular component: This level describes the location of proteins within the cell or extracellular environment (*e.g.*, nucleus, mitochondria, or cell membrane).

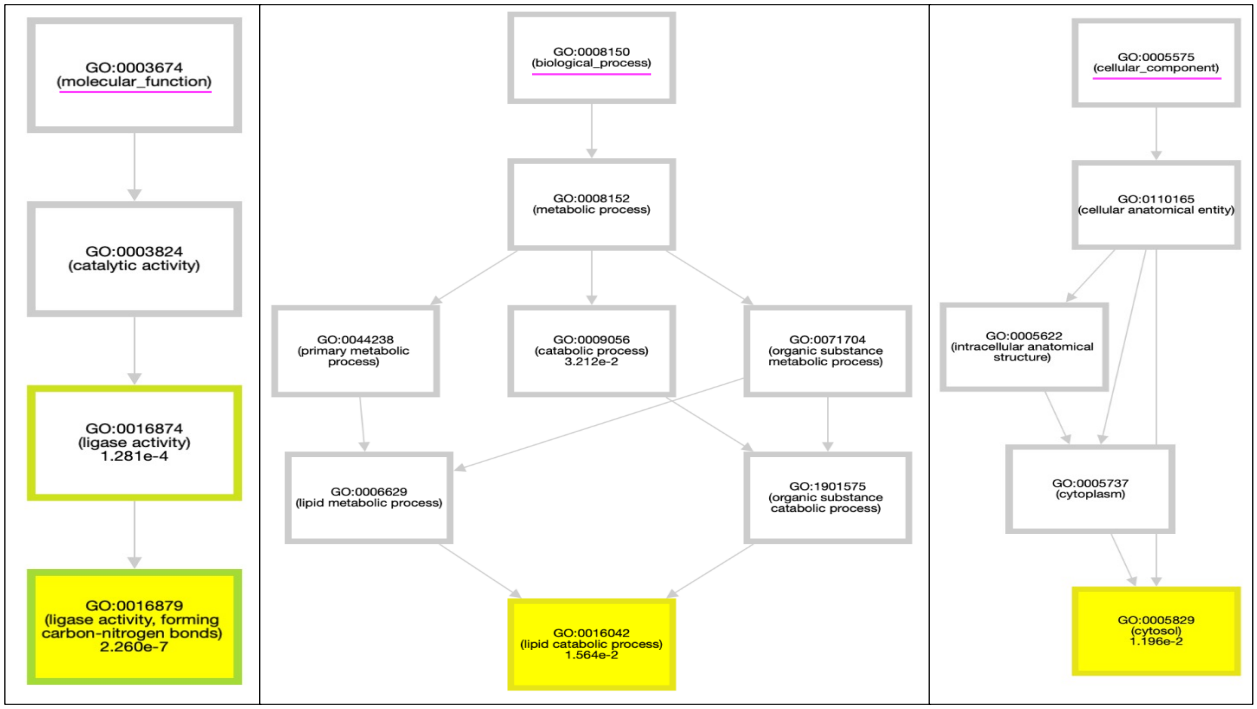

Figure 1.13. GO annotations. Showing GO annotations for human lipase family members. The molecular function (left diagram), the biological process (middle diagram), and the cellular component (right diagram). Image is taken from the GO website (75).

1.10.3 Analysing metabolic pathways

A pathway is a group of functionally associated genes that work together to carry out a specific biological process. By identifying a set of differentially expressed genes in genome-scale experiments and applying relevant statistical techniques, pathway enrichment analysis enables researchers to suggest novel biological functions associated with the particular conditions, genotype-phenotype relationships and disease mechanisms (76,77). Gene Ontology Biological Process (GOBP), which is part of the Gene Ontology resources (https://geneontology.org), a widely used platform for pathway enrichment analysis, provides curated annotations for biological processes, molecular functions, and cellular components across multiple species (78). Other biochemical pathway databases, Reactome (79) and KEGG (80) are also widely used and include multiple types of pathways, such as pathways involved in cancer and immune response.

Changes in expression of metabolic enzymes can occur in cancer tumours resulting from gene duplication or deletion. Accumulation of mutations, overexpression of oncogenes and loss of multiple tumour suppressors can lead to extreme variation in the genetic composition of human tumour cells. One advantage of this alteration is the enhancement of various metabolic pathways to survive and grow (81). When a treatment inhibits one metabolic pathway, the tumour can substitute it with another. Therefore, understanding which metabolic pathways are most exploited by a specific cancer type allows researchers to develop treatments to target the key genes in these pathways (82).

1.10.4 Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)

STRING (Szklarczyk et al. 2019) is a database and a user-friendly web resource providing information on known and predicted protein-protein interactions, represented as direct physical or indirect functional associations, along with the pathways in which they participate (Figure 1.14).

The interaction score is calculated by combining probabilities from various evidence sources, including experimental data, computational predictions, literature, and databases of known interactions (*e.g.*, Gene Ontology and KEGG), while correcting for the probability of randomly observing an interaction (83). The types of evidence include high-throughput experimental results, curated interaction databases, and computational predictions based on genomic and proteomic data. STRING also groups statistical enrichment observations for various pathways and functional subsystems, therefore enables exploring potential drug targets, understanding disease mechanisms, and identifying biological pathways relevant to specific conditions.

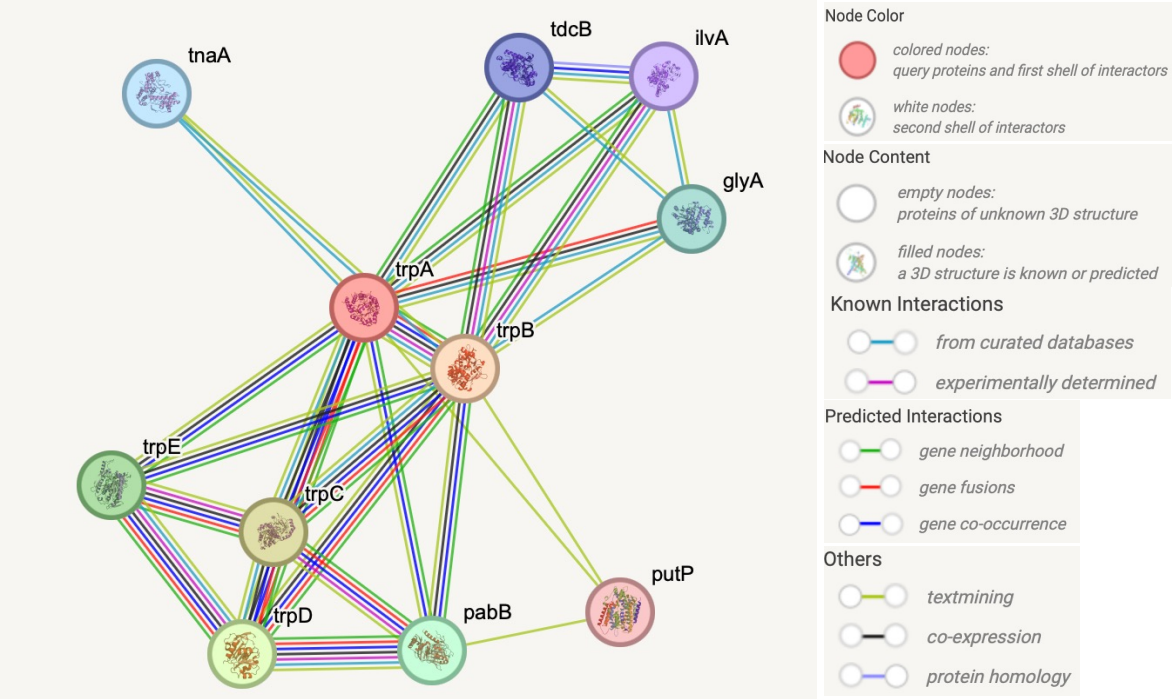

Figure 1.14. An example of STRING protein-protein interactions. In this example, Tryptophan synthase, Alpha subunit was used. The Alpha subunit is responsible for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate. Network nodes represent proteins, and each node represents all the proteins produced by a single, protein-coding gene locus. The edges represent protein-protein associations and proteins jointly contribute to a shared function (The image is taken from the STRING website (84).

1.11 Genetic variants and disease susceptibility

Studying human genetic variation is crucial for understanding susceptibility to infectious diseases such as COVID-19 or the development and progression of diseases such as cancer. Individual genetic differences can influence responses to pathogens, with variants in immune-related genes potentially affecting infection severity and treatment outcomes. This knowledge aids in identifying at-risk populations and tailoring prevention, diagnosis, and therapy strategies to genetic profiles. Genetic variations can also change gene function or regulation, which may lead to cancerous transformations in cells by promoting uncontrolled cell growth and accumulation of genetic mutations.

1.11.1 Genomic instability and cancer transformation

Cancers exhibit an abnormal phenotype due to the imbalance between DNA damage and repair which results in accumulating numerous mutations. Genomic instability, from mutations to chromosomal rearrangements and transcriptional activity, leads to the transformation of normal cells into malignant ones (85–89). This process was first reported in 1974 (90).

In acute myeloid leukaemia, unbalanced translocations result in the loss of chromosomal material and the gain of selected genes (91). Additionally, the Philadelphia chromosome, which results from a reciprocal translocation of chromosomes 9 and 22, has been associated with chronic myelogenous leukaemia and is reported in 95 percent of patients (92).

Extending this understanding to lung adenocarcinoma (LUAD), our study identifying driver mutations in LUAD focuses on the impact of mutations affecting protein function, particularly in relation to their correlation with the emergence of such mutations pre- and post-genome duplication.

1.11.2 Prediction of mutation effects on proteins

Genome-wide association studies (93) identify Single Nucleotide Polymorphisms (SNPs) that are associated with diseases by scanning genomes for variations more frequent than expected in affected individuals. Experimental functional genomics tools, including CRISPR/Cas9 (94,95), investigate SNP effects on gene function. Molecular assays such as luciferase reporter assays study SNP effects on gene regulation (96). However, due to

the high costs and the requirement for access to cutting-edge techniques, various in silico approaches have been developed to study the impact of SNPs and other genetic variations on protein functions and, consequently, human diseases.

1.12 Predicting the impact of genetic variations

Predicting genetic variations in individuals is a significant challenge in systems biology. Accurate prediction of the impact of genetic variants enhances our understanding of how genetic information influences molecular and cellular functions and contributes to clinical factors such as susceptibility to diseases, resistance to illnesses, and response to medications (97).

mCSM-PPI2 and MutPred2 are used in the work presented in chapters 3 and 4 of this thesis to predict the impact of mutations in human proteins concerning COVID-19 and lung cancer. Additional tools, including PolyPhen-2, SIFT, CADD and VarMap, were also employed in chapter 4 to assess mutation impacts in proteins in LUAD cancer.

1.12.1 mCSM-PPI2

mCSM-PPI2 (98) is a web-based machine learning tool that predicts the effect of missense mutations on protein–protein interaction binding affinity. Multiple features such as evolutionary data, inter-residue non-covalent interaction networks analysis, energetic terms and data from various other multiple outsourced tools are embedded into this algorithm to enhance its prediction accuracy.

The method exploits evolutionary information using BLAST (99) since typically some residues in the interface region are evolutionarily conserved. It also considers the

difference in the number of van der Waals', aromatic and hydrogen bond contacts of the wild-type and mutant residue using Arpeggio (100). It calculates the interaction energy between the two interacting chains using FoldX (101).

Taking into account all the above features and employing the Gibbs free energy formulation (102) the output in kcal/mol represents the changes in binding affinity with a positive or negative value for increasing or decreasing binding affinity, respectively.

An mCSM-PPI2 threshold cut-off value of $\Delta\Delta G \geq 0.5$ Kcal/mol is implemented to exclude any possible false positives driven by a minor systematic error in the predictions. This threshold is considered to be associated with a significant enhancement of the binding affinity between the two-interacting proteins in a complex. This threshold has been previously validated with experimental analyses on protein-protein complexes (103,104).

1.12.2 MutPred2

MutPred2 is a machine learning-based tool that uses multiple sequence alignments. It probabilistically assesses the pathogenicity of amino acid substitutions by analysing protein sequence conservation, physical and chemical properties of amino acid substitutions, and conservation scores to generate a comprehensive pathogenicity prediction score (105). MutPred2 scores equal to or greater than 0.5 are considered indicative of predicted pathogenicity.

1.12.3 Polymorphism Phenotyping

Polymorphism Phenotyping (PolyPhen) version 2 (106) is a computational tool that employs a high-quality multiple sequence alignment pipeline (data derived from the

UniRef100 database), to detect conserved residues. It also exploits physicochemical properties, structure-based predictive features and, a probabilistic classifier based on a machine-learning method to predict the potential impact of an amino acid substitution on the structure and function of the human protein, particularly in the context of identifying disease-associated variants in the human genome. PolyPhen v2 uses a scoring system to reflect the degree of change between the wild-type and mutant amino acids:

- 0 to 0.15: Probably benign (likely not damaging)
- 0.15 to 0.85: Possibly damaging (uncertain)
- 0.85 to 1.0: Probably damaging (likely to be damaging)

1.12.4 Sorting Intolerant From Tolerant (SIFT)

SIFT (107) is a computational tool to predict the potential impact of amino acid substitutions on a protein by employing sequence homology information and evaluating the conservation of a particular amino acid position throughout evolution across different species. SIFT uses a scoring system to reflect the degree of change between the wild-type and mutant amino acids:

- A score close to zero indicates that the substitution is predicted to have a significant impact on the protein and likely to be intolerant, damaging, or deleterious.
- A SIFT score close to one suggests that the substitution is predicted to be tolerated and less likely to have an adverse impact on the protein.

1.12.5 Combined annotation dependent depletion (CADD)

CADD (108), a meta-predictor, is a scoring system that employs various human genomic annotations (such as functional annotations and conservation) within a metric framework

to assess the impact of single nucleotide variants, as well as short insertions and deletions in the human genome. CADD integrates information from diverse genomic features, gene model annotations, evolutionary constraints, epigenetic measurements, and functional predictions (109). CADD scores range from 1 to 99, with higher scores likely to be damaging or deleterious. A score of 10 is considered the threshold for the top 10% most deleterious variants, and a score of 20 corresponds to the top 1% most deleterious variants.

1.12.6 VarMap

VarMap (110) maps single nucleotide polymorphisms (SNPs) onto their respective canonical UniProt isoform sequences and corresponding protein 3D structures. It enables clinical geneticists to investigate the impact of genetic variants on the structure of the associated protein.

VarMap performs a GRCh37/CGCh38 assembly check using the Ensembl REST API (111,112) and extracts relevant information for a given SNP from various databases including Ensembl Variant Effect Predictor (VEP) (113), UniProt (114,115), SWISS-PROT (59), BioMart (116), HGNC (117), CATH (27), Pfam (30), M-CSA (118), FASTA (119), PDBsum (120), Scorecons (25), gnomAD (121), and ClinVar (122).

VarMap provides information on mutation impacts as measured by Polymorphism Phenotyping (106) or The Sorting Intolerant from Tolerant scores (107), and VEP (113) for insertions, deletions, copy number variations (CNVs), or structural variants. Initially, the corresponding transcript RefSeqs are extracted from Ensembl BioMart. Then, the UniProt canonical isoform is retrieved from the SWISS-PROT database, it is compared against all PDBe sequences to align the variant amino acid to its relevant position in the

3D structure of the protein entry in the PDB that best matches the canonical form. This comparison allows VarMap to determine if the substitution occurs at a catalytic residue, is involved in a disulfide bond, or interacts with DNA, proteins, ligands, or metals (obtained from PDBsum).

Additionally, VarMap extracts the RefSeq Select accession for each gene from HGNC, retrieves the global allele frequency of each variant from gnomAD, and calculates amino acid conservation using Scorecons (25). Information about clinically approved diseases related to amino acid substitutions are extracted from UniProt and ClinVar. CATH and Pfam provide information about domain properties.

1.13 Resources providing information of known and predicted functional sites

Several resources have been developed to predict and analyse interactions between biomolecules, such as proteins and their ligands. These tools focus on identifying protein ligand interactions, enzyme binding and catalytic sites, and protein-protein interactions, providing valuable insights into biological mechanisms. The platforms which predict molecular interactions affected by mutations in LUAD cancer in Chapter 4, are described below.

1.13.1 BioLip

BioLip (123) is a semi-manually curated database that comprehensively collects and analyses biologically significant interactions between proteins and various ligands, including small molecules, cofactors, metal ions, and other biologically active compounds. Data are sourced from protein structures deposited in the Protein Data Bank (PDB). Each entry undergoes extensive annotation to provide detailed information such as ligand-binding residues, affinity, catalytic sites, Enzyme Commission numbers, and Gene

Ontology terms pertaining to specific protein-ligand interactions. BioLip utilises hierarchical classification schemes to facilitate systematic comparisons and analyses across related protein-ligand complexes.

1.13.2 Mechanism and Catalytic Site Atlas (M-CSA)

M-CSA (118) is a database dedicated to annotating and analysing enzyme active sites and their catalytic mechanisms. It compiles data from enzyme structures deposited in the Protein Data Bank (PDB). Each enzyme structure undergoes detailed annotation to identify specific residues and regions involved in catalysis, including substrate-binding sites, catalytic residues, and essential cofactors or metal ions crucial for enzyme activity. Enzymes are categorised according to their catalytic mechanisms, which describe their roles in chemical reactions.

1.13.3 Inferred Biomolecular Interaction Server (IBIS)

IBIS (124) is a bioinformatics tool and web server dedicated to predicting and analysing protein-protein interactions (PPIs) and other molecular interactions. It utilises sequence and structural information to infer potential binding partners and interaction interfaces based on the analysis of homologous structural complexes. IBIS integrates data from protein databases and experimental interaction data to enhance the accuracy and biological relevance of predicted interactions. The web server provides tools for visualising and analysing predicted interaction networks, facilitating comprehensive exploration of molecular interaction data.

1.14 Overview of the thesis

This thesis comprises three protein research projects all of which examine the impacts of residue mutations on the protein structure and function in different contexts. A brief description of each chapter is outlined below.

Chapter 2 reports the application of a diverse set of computational algorithms to investigate the sequence and structural attributes of *Ideonella sakaiensis* PETase (*Is*PETase). This protein has been shown to have catalytic activity against polyethylene terephthalate (PET), a component of some plastics. This analysis served as the basis for the identification and analysis of novel and potent PETases from metagenome samples predicted to have enhanced substrate binding affinity when compared to the *Is*PETase enzyme. To accomplish this, Hidden Markov Model (HMM) based protocols were applied to select putative PETase-like enzymes from over a billion metagenome sequences in the EBI-EMBL MGnify database. From this subset, a total of twenty-seven putative PETase sequences (based on the variations in size, hydrophobicity and, possessing beneficial residues in the binding pocket), were shortlisted through a two-step process (sixteen in phase one and eleven in phase two). Of the 11 selected in phase two, three demonstrated PETase activity *in vitro*.

In chapter 3, with the emergence of SARS-CoV-2 and the onset of the COVID-19 pandemic, this research focused on understanding the impact of amino acid changes in SARS-CoV-2 proteins and human interactor proteins on the binding affinity between the human and viral proteins. To accomplish this, missense variants from various ethnic groups in human and SARS-CoV-2 proteins from several publicly accessible human and

viral databases respectively, were compiled. A set of twelve human:SARS-CoV-2 complexes (from the protein data bank or built using AlphaFold-multimer) were identified for analysis due to their involvement in the host cell entry, mediating immune responses and cellular translation machinery. Subsequently various computational algorithms were applied to predict the impact of missense variants on binding affinity. The analyses predicted that sixteen human missense variants present in twelve human proteins increase the binding affinity to SARS-CoV-2 interacting proteins by $\Delta\Delta G \geq 0.5$ kcal/mol. Three of these proteins were implicated in host cell entry receptors through spike-binding, mediating immune responses or cellular translation machinery. This research therefore identified putative impacts of human and virus genetic variation on cell entry and infection and, immunity responses. Variation in mutations and their impact across diverse ethnic groups was also considered.

In chapter 4, the impact of mutations affecting the function in proteins associated with lung adenocarcinoma (LUAD) was investigated together with their correlation with the emergence of such mutations, depending on whether they occurred pre- or post-genome duplication in the cancer tumour. Mutation data were compiled from the TRACERx database. Knowledge of the protein functional sites and cancer mutation clusters were derived from paralogues in CATH Functional Families. An in-house method, FunVar (https://funvar.cathdb.info/), was used to identify mutations likely to have functional impact (Functional Impact Events - FIEs) in known driver genes and predicted potential novel driver genes in LUAD. Genes harbouring FIE mutations were further analysed to identify enriched pathways. While the majority of known FIEs were found to occur pre-genome

duplication, post-duplication FIEs were observed to contribute to tumour specialisation during the evolution of LUAD.

Chapter 2: Exploration of naturally evolved polyethylene terephthalate hydrolases from metagenomic data

Chapter 3: Computational analysis of the effect of amino acid changes in SARS-CoV-2 proteins and human interactor proteins on the binding affinity between the host and viral proteins

3.1. Introduction

The difference among individuals in the world population is due to genetic variation of two to three million base pairs in the entire three billion nucleotide base pairs in the haploid human genome which accounts for 0.1% of total DNA (208,209). Advanced molecular techniques make it possible to discover the differences in genetic variation and allele frequency of individuals within and between continental groups. Analyses suggests that only about 10% of polymorphisms result in differences between individuals from continental groups (210,211). For example, large-scale efforts such as the 1000 Genomes Project have systematically catalogued human genetic diversity, identifying over 88 million variants from 26 populations across 18 countries in Africa, East and South Asia, Europe, and the Americas (212). While underlying health conditions and differences in socioeconomic status can also have an impact on how individuals respond to diseases, these genetic variations may play a key role in contributing to greater susceptibility or resistance to diseases for some racial and ethnic groups than for others (213,214).

Severe viral acute respiratory syndrome coronavirus 2 (SARS-CoV-2/ScoV2) was first reported in Wuhan, China in December 2019 and rapidly spread globally. This virus is responsible for the Coronavirus Disease-19 (COVID-19) pandemic with over 704,753,890 confirmed infections and over 7,010,681 deaths according to the

Worldometer (https://www.worldometers.info/coronavirus/), a real-time statistic on world population.

By employing data on the human genetic variation and viral genus from several web-resources, bioRxiv and literature analysis (of well-studied complexes and other less studied complexes), this study aims to predict the likely impact of human and SARS-CoV-2 genetic variation on COVID-19 disease susceptibility and severity of disease from asymptomatic to severe pneumonia and even death (215) in individuals across different ethnicities. It is crucial to comprehend how variations in human proteins affect a person's vulnerability to COVID-19 to improve diagnostics and therapeutic treatment.

3.1.1 Comparison of SARS-CoV and SARS-CoV-2

In the early days of the SARS-CoV-2 outbreak in Wuhan, the complete genome sequences from infected patients with SARS-CoV-2 were released and available from many different publicly accessible sequence resources such as ViralZone (216), NCBI Reference Sequence (www.ncbi.nlm.nih.gov/nuccore/) and CoV-GLUE-Viz (http://cov-glue-viz.cvr.gla.ac.uk/index.php). SARS-CoV-2 genome sequences share 79.5% sequence identity to SARS-CoV that caused the respiratory illness responsible for the 2002–2004 SARS outbreak (217). Both SARS-CoV-2 and SARS-CoV use the spike receptor binding domain (RBD) protein to recognise human angiotensin-converting enzyme 2 (hACE2) and infect the host cells (Figure 3.1) (218).

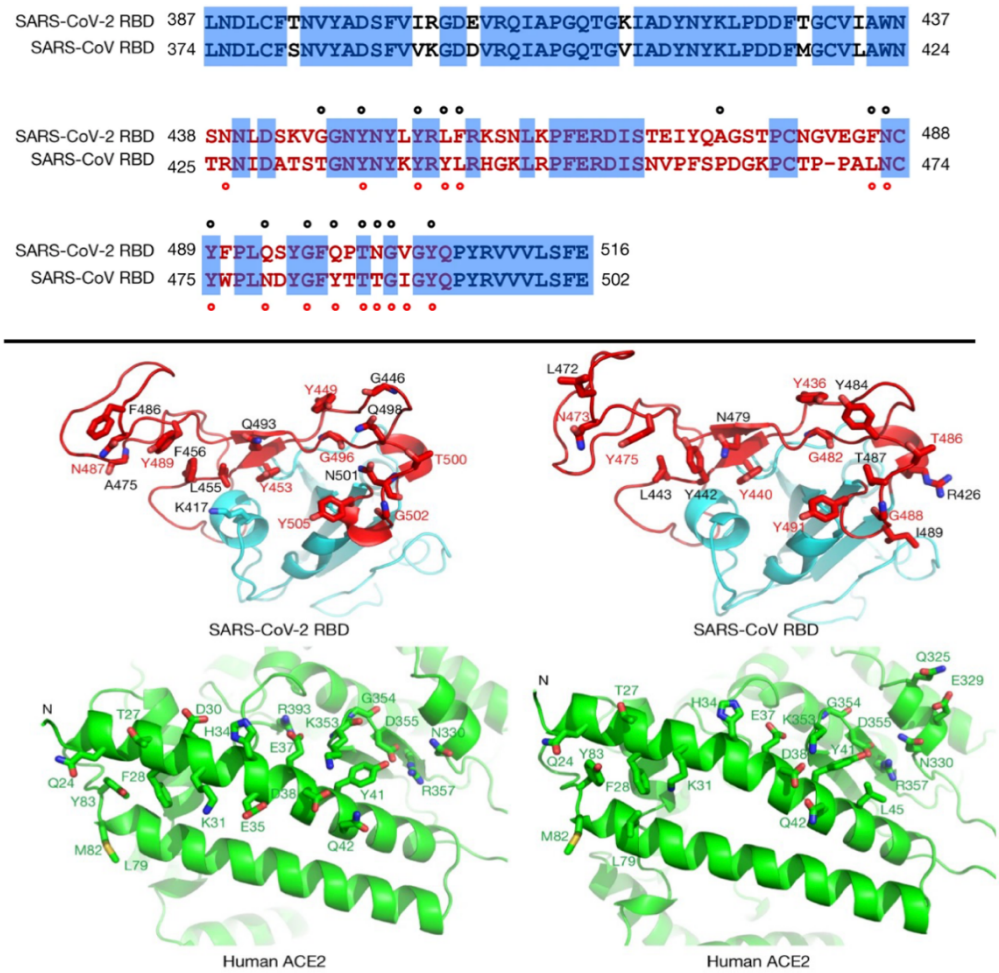

Figure 3.1. Similarities and differences in SARS-CoV-2 and SARS-CoV RBDs. Top) Sequence alignment of the SARS-CoV-2 and SARS-CoV RBDs. Contacting residues in the SARS-CoV-2 RBD are indicated by black dots; contacting residues in the SARS-CoV RBD are indicated by red dots (219). Bottom-Left: interface residues of SARS-CoV-2 RBD-ACE2. Bottom-Right) interface residues of SARS-CoV RBD-hACE2. Residues in both RBDs that are involved in hACE2 binding are indicated by red labels (220).

Despite many similarities between these two viruses, the slight differences in SARS-CoV-2 spike protein led to a 10-fold higher binding affinity to hACE2 and consequently more pathogenicity compared to SARS-CoV. While the spike proteins in both viruses are capable of “up” and “down” status, SARS-CoV-2 adopts a different conformation to SARS-CoV. SARS-CoV-2 RBD is positioned nearer to the central cavity of the spike protein trimer while RBD is packed tightly against the N-terminal domain (NTD) in SARS-CoV (220). Another difference is that different amino acids in RBD lead to a

higher binding affinity in SARS-CoV-2 (Table 3.1). In SARS-CoV-2, Lys417 forms a salt-bridge connection with Asp30 of hACE2 while in SARS-CoV a valine at the same position forms no interaction with any residues in hACE2 (219).

SARS-CoV-2/SARS-CoV	SARS-CoV-2 interaction with hACE2	SARS-CoV interaction with hACE2
K417/V404	Salt-bridge with D30	Unable to form any interaction
F486/L472	Q24, L79, M82 and Y83	The same but weaker interaction
Q493/N479	K31, H34 and E35	H34
L455/Y442	Both have similar interactions with D30, K31 and H34	
N501/Y487	Both have similar interactions with Y41, K353, G354 and D355	
Q498/Y484	Both have similar interactions with D38, Y41, Q42, L45 and K353	

Table 3.1. Summary of SARS-CoV-2 and SARS-CoV interactions with hACE2.

3.1.2. SARS-CoV-2 structure

SARS-CoV-2 belongs to the betacoronavirus 2B lineage (221), in the subfamily Coronavirinae of the family Coronaviridae (222) and the order *Nidovirales* (223). SARS-CoV-2, similar to other two members of the family: Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and Middle East Respiratory Syndrome coronavirus (MERS-CoV), has zoonotic origins and is likely to have originated in bats and camels, respectively (224,225). It has a sequence identity of 79.5% and 50% to SARS-CoV and MERS, respectively (217,226).

SARS-CoV-2 is a positive-strand single-stranded RNA virus (+ssRNA) of approximately 30 Kb. Its mutation rate compared to other RNA viruses such as influenza is lower (227) due to the proofreading activity of the viral replicative complex (228,229). SARS-CoV-2 virions are 60 - 140 nm in diameter, have four structural

proteins: the spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins (Figure 3.2) and 16 non-structural proteins (Nsp1–16) (230,231).

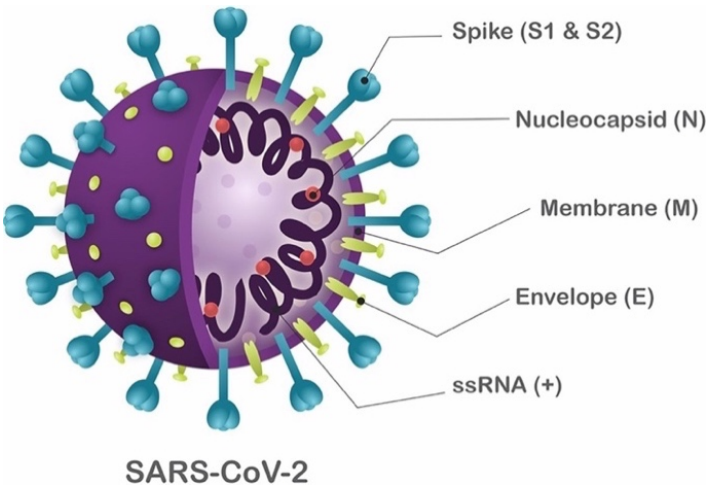

Figure 3.2. Schematic structure of SARS-CoV-2. The viral structural proteins are spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins. The origin of the lipid bilayer is the host cell membrane. The S, M, and E proteins are implanted in the envelope (232).

SARS-CoV-2 entry into human angiotensin-converting enzyme 2 (hACE2), the cell surface receptor in humans, is mediated by the spike protein, a transmembrane glycoprotein. hACE2 is a protease, responsible for blood pressure and volume regulation and expressed in a number of tissues including the lung, heart, kidneys and gastrointestinal tract (233). However, epithelial cells of the upper and lower respiratory tracts are the main tissue for SARS-CoV-2 infection (234–236).

Thus nonsynonymous mutations leading to amino acid changes in the spike protein, are of particular concern as they alter the surface spike structure which recognises the

hACE2 receptor (237) and may enhance virus transmissibility, reduce antibody binding and immune protection and vaccine efficacy (238).

3.1.3 SARS-CoV-2 spike protein structure

The spike protein is responsible for human ACE2 receptor binding and the viral fusion to the epithelial cells of the upper and lower respiratory tracts membranes. Each homotrimeric Spike (S) protein consists of two subunits (Figure 3.3): S1 and S2 which stick out from the viral surface (220,239). The smaller subunit, S1 fragment, at the membrane distal tip of the Spike, is formed of an N-terminal domain (NTD) receptor binding domain (RBD), C-terminal domain 1 (CTD1) and C-terminal domain 2. SARS-CoV-2 RBD interacts with the surface of hACE2 (219) with a high binding affinity of approximately 15 nM (239,240).

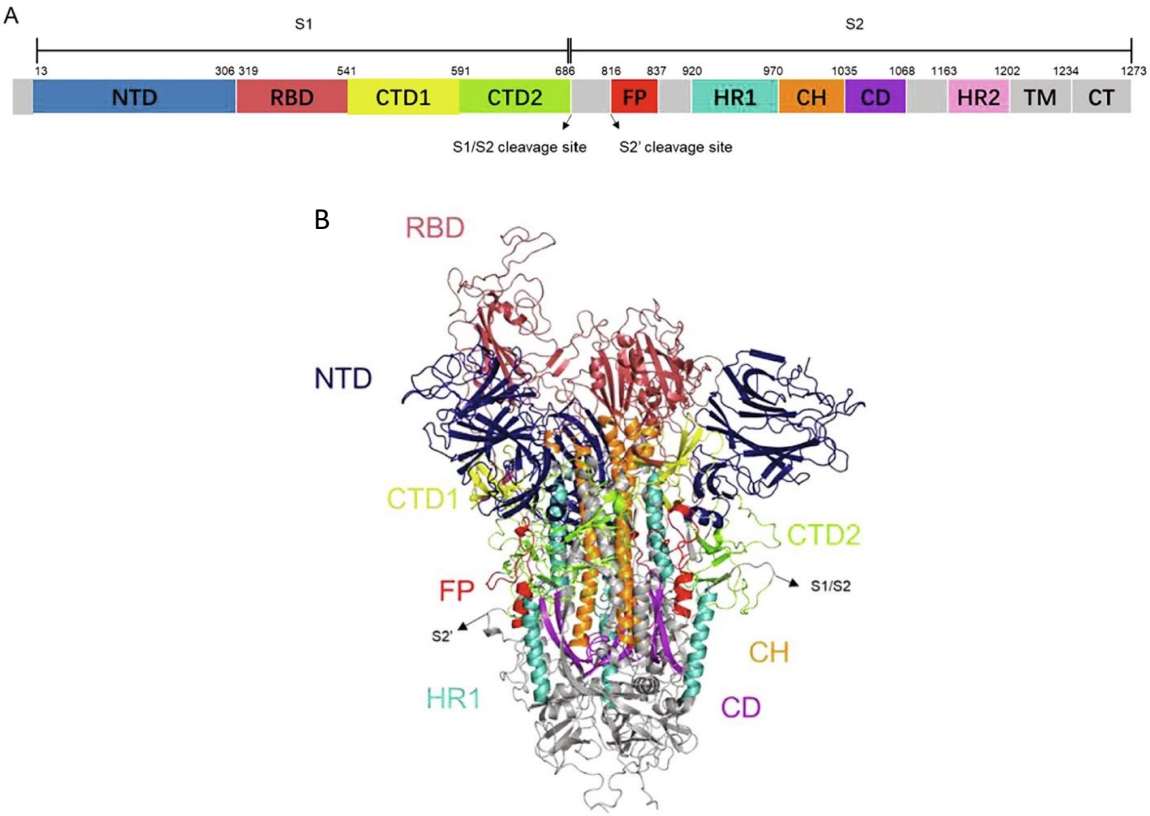

Figure 3.3. Sequence and structure of the Spike protein. A) Diagram of full-length SARS-CoV2 Spike protein; S1 receptor binding subunit; S2 membrane fusion subunit; TM, transmembrane domain; HR-N, heptad repeat-N; HR-C, heptad repeat-C. B) the schematic structure of the Spike protein. Figure is taken from Sun *et al.*, (2021) (241).

The RBD of SARS-CoV-2 is formed from two subdomains: a twisted five-stranded antiparallel β sheet (β1-β4 and β7) and the core subdomain containing the short β5 and β6 strands, α4 and α5 helices and loops which form a gently concave surface with a ridge on one side. The core is known as the Receptor Binding Motif (RBM) and this binds to the exposed outer surface of the claw-like structure of hACE2 (219,239). In RBD, of nine cysteine, eight of them form four disulfide bonds: Cys336–Cys361, Cys379–Cys432 and Cys391–Cys525 stabilise the β sheet structure in the core and Cys480–Cys488 connects the loops in the far end of the RBM (Figure 3.4).

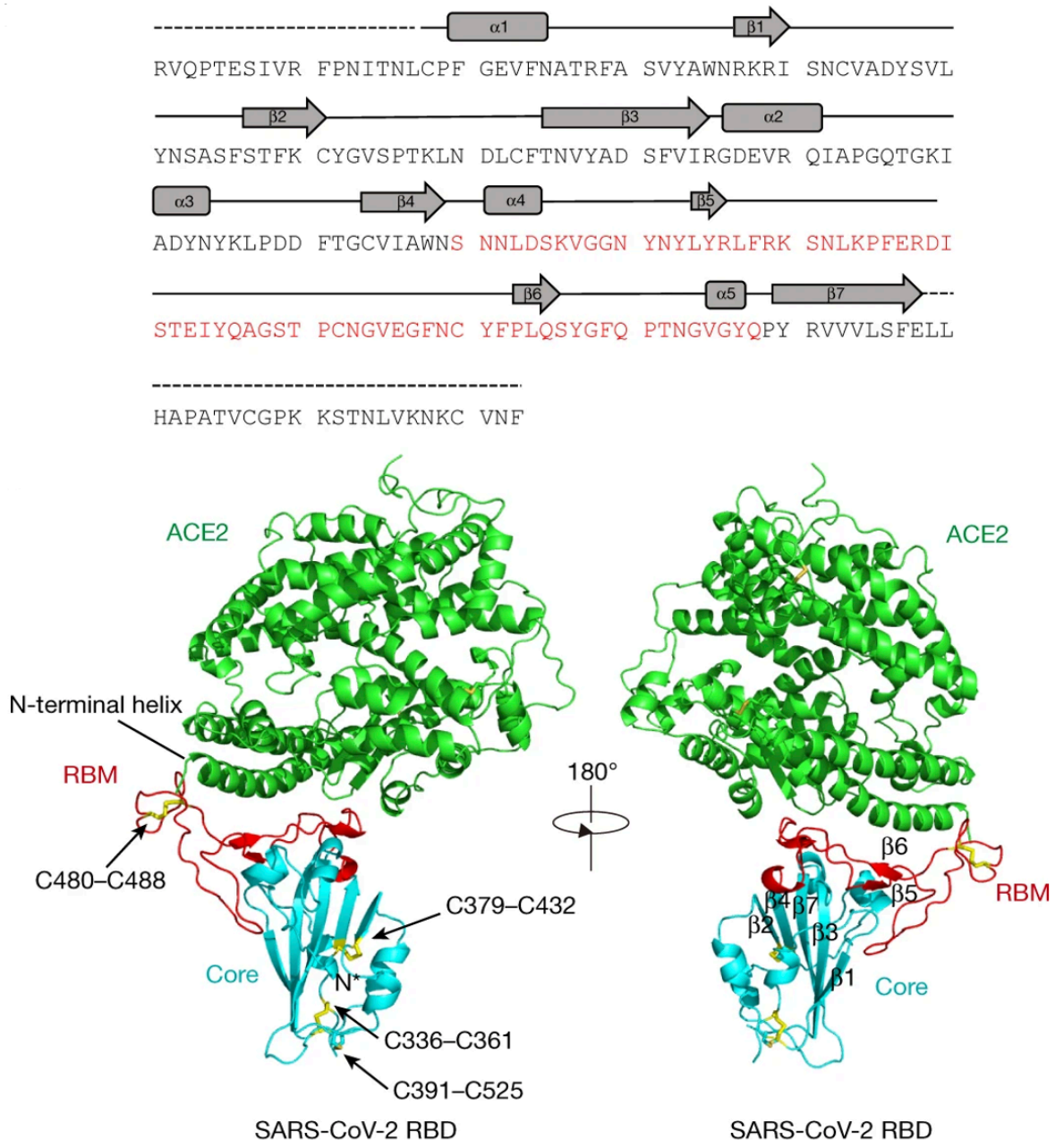

Figure. 3.4. Sequence and structure of SARS-CoV-2 RBD. Top) SARS-CoV-2 RBD sequence and secondary order of the structure, residues in red represent Receptor Binding Motif (RBM). Bottom) overall structure of SARS-CoV-2 RBD bound to ACE2 (219). SARS-CoV-2 RBD (cyan) interacts with hACE2 (green) through RBM (red). RBD's disulfide bonds are displayed in yellow.

As presented in figure 3.5, the Spike protein has two distinct structural states: prefusion and post-fusion (242).

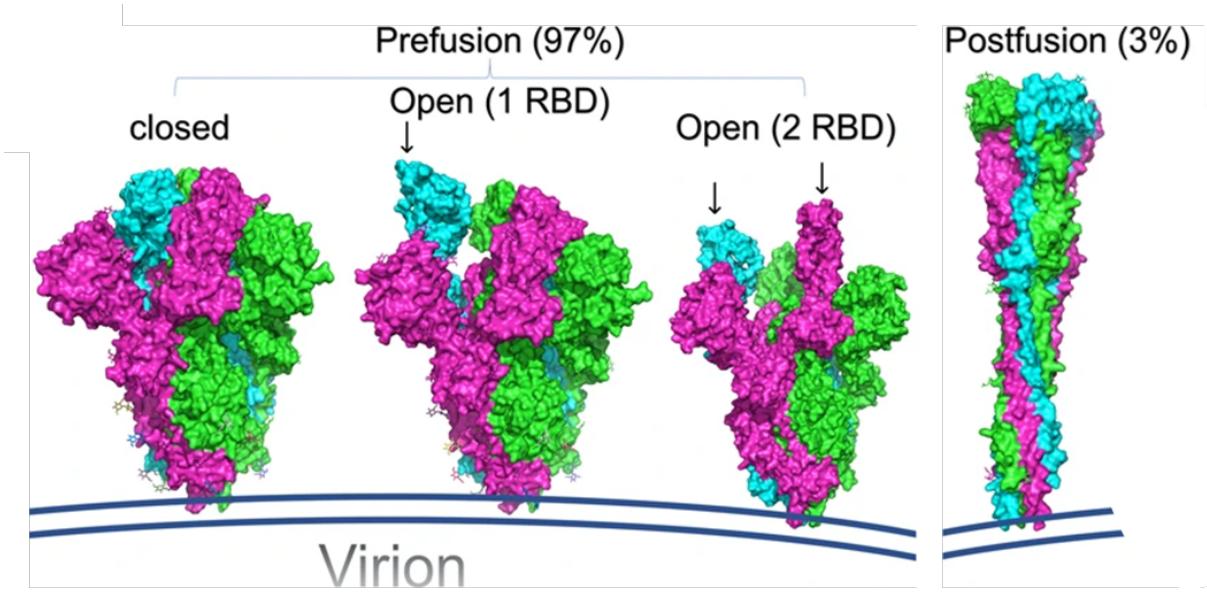

Figure 3.5. The structures of the prefusion and post-fusion trimeric spike for SARS-CoV-2. Images 1 to 3 from left to right: when the Spike is in the closed status, none of the RBDs are in 'up' conformation. During the prefusion, first, one, and then the second RBD change conformation shifts from 'down' to 'up' which makes the Spike fully open and ready for fusion. Small arrows show the open RBD in the prefusion subclasses (243). PDB structures: post-fusion (6M3W), closed prefusion (6VXX), open 1 RBD prefusion (6VYB), and open 2 RBD prefusion (6X2B).

3.1.4 SARS-CoV-2 RBD:hACE2 interface binding and the mechanism of host infection

The Spike protein exhibits two main conformational states, including the open state or up conformation and the closed state or down conformation (Figure 3.6). In the fully open state, the three RBDs protrude from the interface formed by three spike protein protomers. In the fully closed state, the three RBD are bound *in trans* into a pocket formed by the NTD and the receptor binding site is largely occluded (220).

Although RBD is usually packed down against the top of Spike (244), binding to hACE2 initiates a down-to-up conformational change in the spike protein as the transition to open state is necessary for the fusion of the SARS-CoV-2 and the host cell membranes (Walls et al., 2020, Huo et al., 2020).

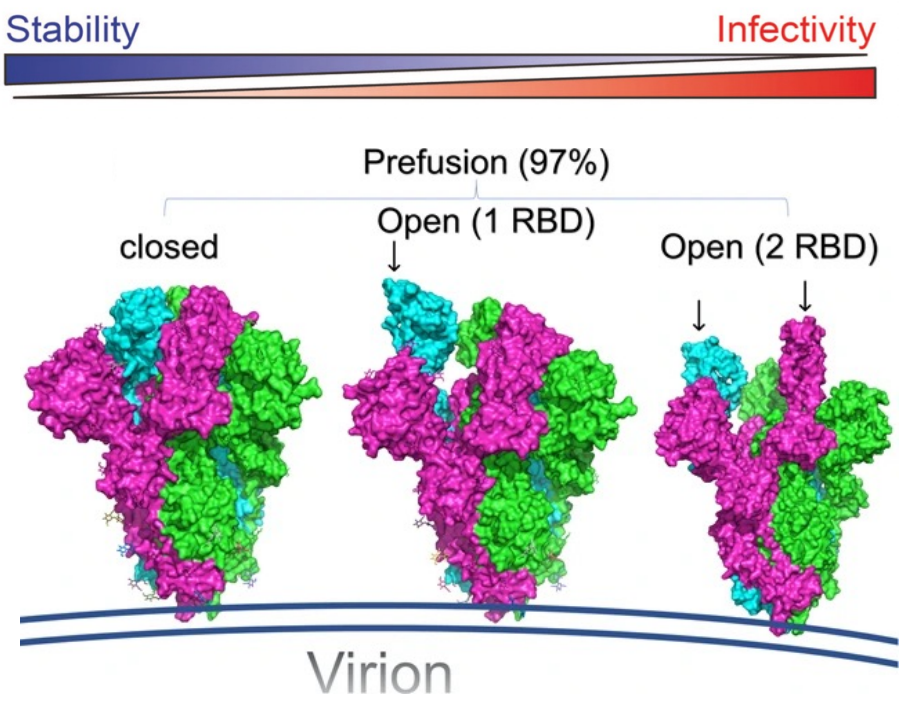

Figure 3.6. Stepwise demonstration of conformational changes of RBD. In the closed state, all three RBDs are in the 'down' conformation (stable conformation) and in the open state in the ‘up’ conformation, which can interact with hACE2 (infectious status). Stimulation factors trigger the conformation change of RBD from ‘down’ to ‘up’ in the prefusion state to make the Spike fully open and ready for fusion. Small arrows show the open RBDs (243). PDB structures: closed prefusion (6VXX), open 1 RBD prefusion (6VYB), and open 2 RBD prefusion (6X2B).

As shown in figure 3.7, hACE2-RBD binding destabilises the spike prefusion structure which leads to the S1 subunit disconnection. The S2 subunit refolds to form a stable post-fusion conformation. Next, RBD goes through conformational transitions from down to up conformation. SARS-CoV-2 RBD slightly move into the small claw-like lobe of the N-terminal of ACE2.

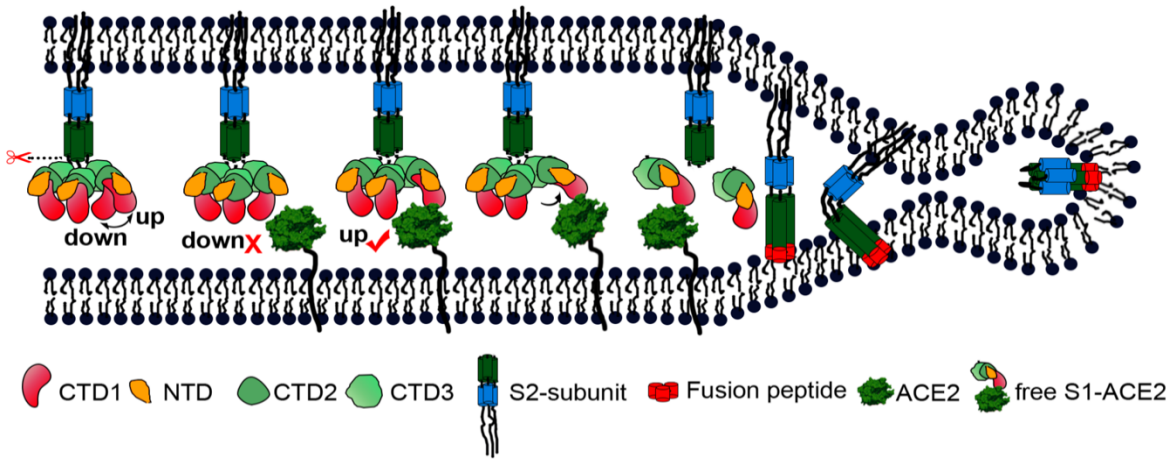

Figure 3.7. A cartoon representation showing the pre- to the post-fusion transition of the SARS-CoV-2 Spike. The binding to ACE2 induces the 'down' to 'up' transition of RBDs, promotes the disassociation of the S1-ACE2 complex from the S1/S2 cleaved S glycoprotein, induces the pre- to the post-fusion transition of the S2 subunit, and finally initiates the membrane fusion. Figure is taken from Song *et al.*, (2018) (247).

At the Phe486 position, SARS-CoV-2 interacts with hACE2 Gln24, Leu79, Met82, and Tyr83. SARS-CoV-2 Gln493 forms a hydrogen bond with Glu35. Gln493 also interacts with hACE2 Lys31 and His34 (219). Lys417 creates a positive charge patch distal (>5Å) from the main binding interface. It also forms a salt bridge with hACE2 Asp30. hACE2 glycosylated Asn90 forms a hydrogen bond with Arg408 of the RBD core. Arg408 is conserved between SARS-CoV and SARS-CoV-2; thus it is concluded that this host-pathogen glycan mechanism is one of the key factors of receptor recognition by SARS-CoV-2 (248). Figure 3.8 presents a summary of interactions between hACE2 and RBD provided by PDBSum.

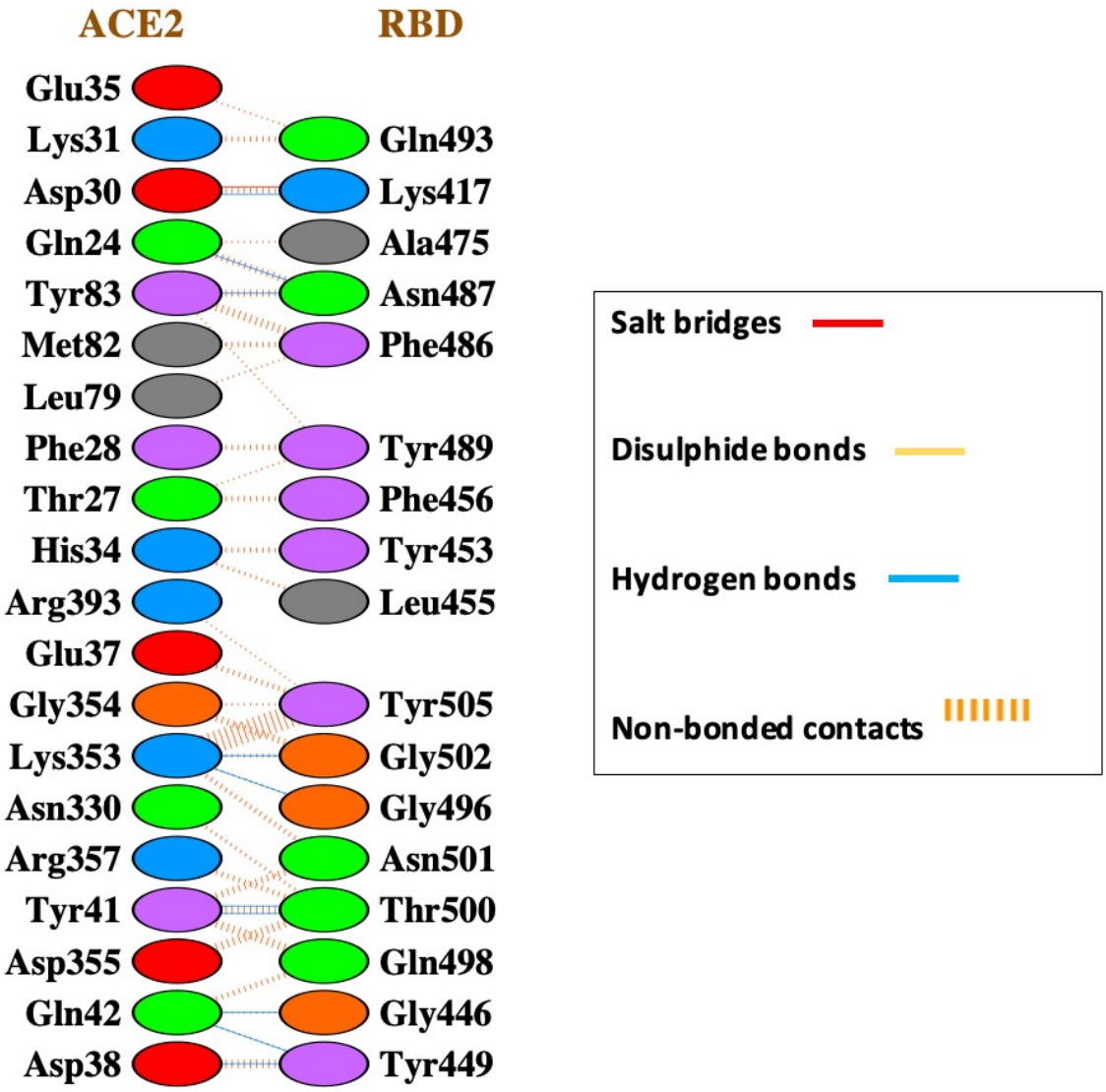

Figure 3.8. A summary of the hACE2:SCoV2 RBD interactions. One salt bridge, ten hydrogen bonds and 101 non-bonded contacts are predicted. For non-bonded contacts, the width of the striped line is proportional to the number of atomic contacts. Data collected from PDBSum (120).

RBM contact with the N-terminal helix of hACE2 is maximised due to the combination of two flexible and two bulky residues. This four-residue motif glycine-valine/glutamine-glutamate/threonine-glycine which forms the loop in RBD adopts a conformation which brings the residues closer to hACE2. As a result, an extra hydrogen bond between Ala475 of the SARS-CoV-2 RBM, and Gln24 of ACE2 is formed. Other hydrogen

bonds between the SARS-CoV-2 RBM and ACE2 are separate bonds between Gln493 of the SARS-CoV-2 RBM with Lys31 and Glu35 of ACE2 (248).

Thus, it is important to assess the impact of single amino acid changes in the RBD of the SARS-CoV-2 in the emerged variants of concern to monitor changes in the receptor recognition mechanism of the SARS-CoV-2 and the binding affinity to human hACE2, which determines the transmissibility, infectivity, pathogenesis of the virus.

3.1.5 SARS-CoV-2 variants of concern

The World Health Organisation (WHO) has been assessing the mutations that could affect SARS-CoV-2 RBD binding affinity or have other impacts on the disease (249). Those variants with a significantly higher transmissibility, severity and/or immunity are associated with increased risks to global public health and therefore are classified as Variant Of Concern (VOC). From the start of the pandemic through the first quarter of 2024, mutations in Spike proteins of variants of concern: Alpha (lineage B.1.1.7), Beta (B.1.351, B.1.351.2 and B.1.351.3), Gamma (P.1, P.1.1 and P.1.2), Delta (B.1.617.2, AY.1 and AY.2) and Omicron (BA.1, BA.2, BA.2.12.1, BA.2.75, BA.4, BA.5, XBB.1.5, BA.2.86 and JN.1) led to an enhanced binding affinity to the hACE2 receptor and escape from neutralizing antibodies; thus an increase in infectivity and transmission (249–252).

While the mutations on the RBD-ACE2 interface have a direct effect on binding affinity, non-interface mutations may have an impact on forming different hydrogen bonding patterns and back-bone torsional changes thus the preference for one RBD

conformation (‘down’ and ‘up’ state) over another. SCoV2 RBD in the ‘up’ state, leads to increased SARS-CoV-2 virulence and transmissibility. For example, the first observed single mutation in SARS-CoV-2 RBD was Asp614Gly. Although, Asp614 interacts with Lys854 and Thr859 through hydrogen bonds to retain the RBD in ‘down’ state, Gly614, works in favour of the up conformation by reducing the energy cost of the conformational transition due to lack of ability to form hydrogen bonds with Lys854 and Thr859 (Ray et al., 2021, Mansbach et al., 2021). This hypothesis was experimentally validated in vitro, as the RBD ‘up’ state was observed nearly seven times more than the full ‘down’ state while the occurrence for both conformations is equal in the wild-type strain (254).

3.1.5.1. Alpha variant

On 14 December 2020, the United Kingdom officially announced the emergence of a new SARS-CoV-2 variant and soon after, it spread world-wide. It was named lineage B.1.1.7 or Alpha variant (255). The mutation Asn501Tyr is found in RBD and Ser98Phe, Asp138His and Trp152Arg in NTD (216). The sole mutation in RBD, Asn501Tyr forms a new favourable π-π stacking interaction (Figure 3.9) with Tyr4. Tyr501 also forms a hydrogen bond with Lys353 of ACE2 (256).

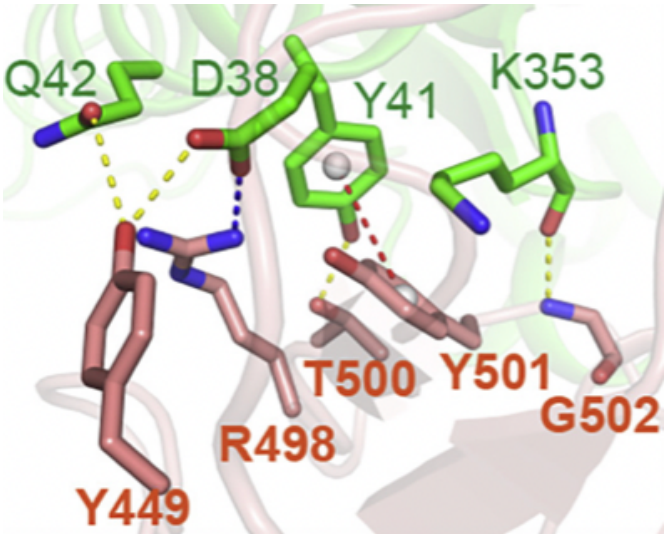

Figure 3.9. The interaction network of RBD (brown) with ACE2 (green). The mutation, Asn501Tyr in RBD forms a new π-π stacking interaction with Tyr41 in ACE2. Figure is taken from Han *et al*., (2022) (256).

Ala570Asp mutation in subdomain1 (SD1) replaces a hydrophobic amino acid with a charged one. This mutation alters the interaction of RBD with the S2 core (257) which may facilitate the RBD conformation change to the 'up' state (258). Table 3.2 presents all emerged mutations in Spike protein (NTD and RBD) in the Alpha variant.

Spike Domain	Mutations in Alpha variant
NTD	S98F, D138H, W152R
RBD	N501Y

Table 3.2. List of Spike's NTD and RBD mutations in Alpha variant. Source: viralzone.expasy.org (216).

3.1.5.2. Beta variant

The emergence of lineage B.1.351 or Beta variant was reported in South Africa on 18 December 2020. This variant carries four mutations: Phe384Leu, Lys417Asn, Glu484Lys and Asn501Tyr in RBD. Beta variant' RBD binding affinity to hACE2 is 4.62-times greater than the wild-type RBD (259). The Glu484Lys mutation on RBD of Lineage B.1.351 is the key mutant associated with a high rate of immune escape. The

replacement of the negatively-charged Glutamic acid with a positively-charged Lysine at residue 484 leads to changes in RBD's interface charge with a higher electrostatic complementarity and consequently to higher affinity binding to hACE2 (249). Table 3.3 presents mutations in Spike protein (NTD and RBD) in the Beta variant.

Spike Domain	Mutations in Beta variant
NTD	L18F, T19I, A27S, D80A, D215G
RBD	P384L, K417N, E484K, N501Y

Table 3.3. List of Spike's NTD and RBD mutations in Beta variant. Source: viralzone.expasy.org (216).

3.1.5.3. Gamma variant

The first confirmed cases with gamma variants were Brazilian tourists vising Japan on 6 January 2021. Although beta and gamma share the same mutations in RBD: Lys417Thr, Glu484Lys and Asn501Tyr, these mutations arose independently (260). While the mutation at 417 is common between the two variants, Lys417Thr makes Gamma less resistant to naturally acquired or vaccine-induced antibody responses. This unfavourable mutation in RBD leads to the loss of a salt bridge bond with Asp30 of hACE2; thereby reducing the binding affinity (261). Table 3.4 presents mutations in Spike protein (NTD and RBD) in Gamma variant.

Spike Domain	Mutations in Gamma variant
NTD	L18F, T20N, P26S, D138Y, R190S
RBD	K417T, D427N, E484K, N501Y

Table 3.4. List of Spike's NTD and RBD mutations in Gamma variant. Source: viralzone.expasy.org (216).

3.1.5.4. Delta variant

The first SARS-CoV-2 delta variant infected cases were detected in India in late 2020 (255). This variant carries fourteen mutations in the Spike glycoprotein as published

in the Centres for Disease Control and Prevention (www.cdc.gov). Asn417 of RBD forms a salt bridge with Asp30 of hACE2 which contributes to higher receptor-ligand binding affinity (256).

The Pro681Arg mutation in the S protein of delta increases the concentration of the cleaved S2 subunit. The level of cleaved S2 subunit was even higher in presence of both mutations: Asp614Gly and Pro681Arg (262). Table 3.5 presents mutations in Spike protein (NTD and RBD) in Delta variant.

Spike Domain	Mutations in Delta variant
NTD	T19R, T95I, G142D, Y145H (sublineage AY.2), R158G, A222V, W258L
RBD	K417N (sublineage AY.1), L452R, T478K, N501Y

Table 3.5. List of all mutations in Spike's Delta variant. Source: viralzone.expasy.org (216).

3.1.5.5. Omicron variant

The World Health Organisation announced the emergence of a new variant of COVID-19, B.1.1.529 (Omicron variant) in South Africa on the 24th of November 2021. Compared with other variants, omicron possesses the most mutations in the receptor-binding motif of RBD-Spike protein.

The surface charges of the omicron RBD change to be significantly more positive due to mutations in three residues in the RBD: Thr478Lys, Gln493Arg, and Gln498Arg. This leads to enhancing the binding affinity as Arg493 and Arg498 interact with Glu35 and Asp38 of hACE2, respectively. A new hydrogen bond is formed between Asn477 of RBD and Ser19 of hACE2 (256). Although, Lys478 is not in direct contact with any

residues in hACE2; it may have an impact on hACE2 binding allosterically. In contrast, Glu484Lys/Ala reduces the negative charge in RBM.

While Ser477Asn and Asn501Tyr in omicron RBD lead to higher binding affinity to hACE2 (256), other substituted residues (*e.g.,* Lys417Asn, Gly446Ser, Glu484Lys/Ala, Gly496Ser, and Tyr505His) reduced the binding affinity. E484A leads to lower binding affinity due to lack of side chains to interact with Lys31 of hACE2 (263). Tyr505His mutation is unable to form the same contacts with ACE2; thus, resulting in weaker binding affinity. It could explain that even though, the omicron lineage showed higher transmissibility; its binding affinity to hACE2 is weaker than delta strain (256). Table 3.6 presents mutations in Spike protein (NTD and RBD) in variants of concerns in Omicron sub-lineages BA.1, BA.2, BA.2.12.1, BA.2.75, BA.4, BA.5, XBB.1.5, BA.2.86 and JN.1.

Omicron stains	Spike domain	Non-synonymous mutation in BA.1, BA.2, BA.2.12.1, BA.2.75, BA.4, BA.5, XBB.1.5, BA.2.86 and JN.1
BA.1	NTD	V83A, G142D, H146Q, Q183E, V213E, G252V.
	RBD	G339H, R346T, L368I, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, V445P, G446S, N460K, S477N, T478K, E484A, F486P, F490S, Q498R, N501Y, Y505H, D614G.
BA.2	NTD	T19I, G142D, V213G.
	RBD	G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, S477N, T478K, E484A, Q493R, Q498R, N501Y, Y505H, D614G.
BA.2.12.1	NTD	T19I, LPPA24-27S, G142D, V213G.
	RBD	G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, L452Q, S477N, T478K, E484A, Q493R, Q498R, N501Y, Y505H, D614G.
BA.2.75	NTD	T19I, G142D, K147E, W152R, F157L, I210V, V213G, G257S.
	RBD	G339H, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, G446N, N460K, S477N, T478K, E484A, Q498R, N501Y, Y505H, D614G.
BA.4 and BA.5	NTD	T19I, G142D, V213G.
	RBD	G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, L452R, S477N, T478K, E484A, F486V, Q498R, N501Y, Y505H, D614G.
XBB.1.5	NTD	T19I, V83A, Q52H, G142D, H146Q, Q183E, V213E, G252V.
	RBD	G339H, R346T, L368I, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, V445P, G446S, F456L, N460K, S477N, T478K, E484A, F486P, F490S, Q498R, N501Y, Y505H, D614G.
BA.2.86	NTD	T19I, V83A, Q52H, G142D, H146Q, Q183E, V213E, G252V.
	RBD	I332V, R403K, D339H, V445H, G446S, N481K, N450D, L452W, 483del, E484K, F486P, N501Y, Y505H, D614G.
JN.1	NTD	T19I, V83A, Q52H, G142D, H146Q, Q183E, V213E, G252V.
	RBD	I332V, R403K, D339H, V445H, G446S, L455S, N481K, N450D, L452W, 483del, E484K, F486P, N501Y, Y505H, D614G.

Table 3.6. List of common Spike's NTD and RBD mutations in Omicron sub-lineages. Source: viralzone.expasy.org (216).

3.1.6 The role of other SARS-CoV-2 proteins in COVID-19 susceptibility and infection

The fourteen open-reading frames (ORFs) in the 30Kb genome of SARS-CoV-2 are translated to four structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N); 16 non-structural proteins (NSPs) and 9 accessory factors (Figure 3.10).

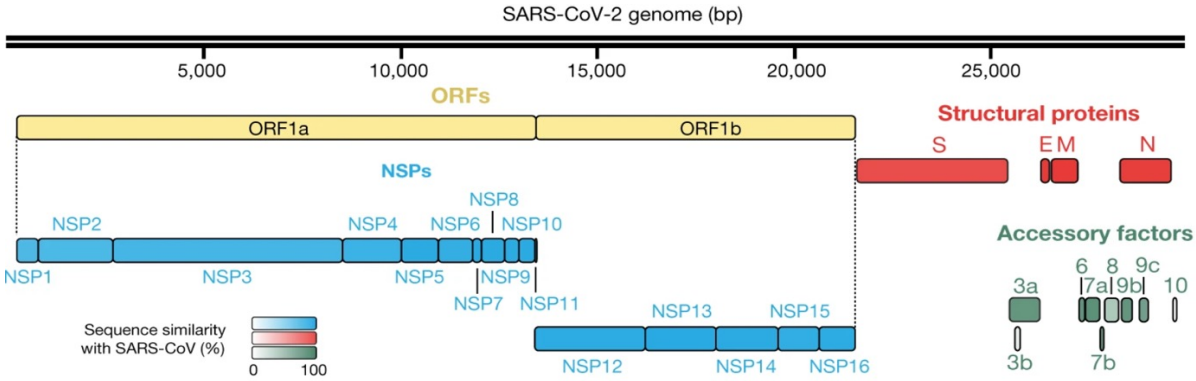

Figure 3.10. A schematic presentation of the SARS-CoV-2 genome and encoded proteins. Fourteen ORFs encode 4 structural proteins (red bars), 16 Nsps (blue bars) and nine accessory factors (green bars). Sequence similarity of each protein in SARS-CoV-2 and SARS-CoV is presented as the gradient. Figure is taken from Gordon *et al*., (2020) (264).

ORF1a and ORF1b encode the non-structural proteins (Nsp1–Nsp16) that assemble as the replicase transcriptase complex which comprises of multiple enzymes, such as the papain-like protease (Nsp3), the main protease (Nsp5), the Nsp7–NSP8 primase complex, the primary RNA-dependent RNA polymerase (Nsp12), a helicase–triphosphatase (Nsp13), an exoribonuclease (Nsp14), an endonuclease (Nsp15) and N7- and 2′*O*-methyltransferases (Nsp10 and Nsp16) (264–267).

In an early study, Gordon *et al.,* (2020) (264) found 322 high-confidence protein-protein interactions (PPIs) between the SARS-CoV-2 and human proteins in a variety of host pathways and biological processes such as:

- Nsp5, Nsp8, Nsp13 and E complex interact with the host epigenetic and gene-expression regulators.
- Nsp2, Nsp6, Nsp7, Nsp10, Nsp13, Nsp15, ORF3a, E, M and ORF8 interact with the host vesicle trafficking proteins.
- Nsp8 and N interact with the host RNA processing and regulation.
- Nsp9, Nsp15 and ORF6 interact with the host nuclear transport machinery.
- Nsp1 And Nsp13 interact with the host cytoskeleton.
- Nsp4, Nsp8 and ORF9c interact with the host mitochondria.
- Nsp9 interacts with the host the extracellular matrix.

They showed that different proteins of SARS-CoV-2 such as Nsp13, Nsp15, ORF9b and envelope proteins interact and compete with human immune-associated proteins in innate immune pathways to suppress or block triggering interferon (IFN) and NF-B pathways in response to the infection.

SARS-CoV-2 suppresses the production and activation of type I interferons, more effectively than SARS-CoV. Interferons are important messenger molecules in the innate immune response (268). It has been reported that SARS-CoV-2 VOCs are more resistant to interferons and more capable of evading innate immunity which both are key factors in the evolution of the SARS-CoV-2 virus (269).

Therefore, in this study, in addition to studying the impact of genetic variants across different human populations on the binding between human ACE2 and SARS-CoV-2 Spike protein, the changes in protein-protein interaction and binding affinity between the relevant immune-associated proteins and their partner proteins in SARS-CoV-2 were also investigated.

3.1.7 Aim of the project

Since the beginning of the pandemic in early 2020, it has been observed that certain ethnic groups, albeit being clinically healthy, experienced severe morbidity compared to other groups. Although environmental elements such as access to health care, cultural factors and lifestyle play an important role, genetic variation may also have an impact on a broad range of COVID-19 morbidity. This research project applies several computational algorithms to predict the likely impact of human missense mutations on proteins that interact with SARS-CoV-2's proteins and consequently alter human:virus complex binding affinity.

Briefly, proteomics-based studies by Gordon *et al.,* (2020) (264) and the dataset of SARS-CoV-2-interacting proteins in humans were collected from the UniProt-SARS-CoV-2 resource and IntAct's COVID-19 dataset (https://www.ebi.ac.uk/intact/home) identified the human:virus interacting proteins. Human:viral protein complexes were compiled from the PDB or modelled. Human genetic variation information and associated allele frequencies were collected from several publicly accessible databases. Subsequently, changes in binding affinity of human:virus complexes were predicted, and the functional sites and potential structural and functional impacts associated with the human mutations were investigated. In summary, this study discloses the importance of genetic variants in human proteins and their possible impact in putting particular ethnic communities at a higher risk of COVID-19 morbidity.

3.2 Methods

Genomic variation databases provide comprehensive data on genetic differences across various human populations, allowing researchers to analyse and compare patterns of genetic variation, such as single nucleotide polymorphisms (SNPs), and understand genetic diversity. In addition to exploring how populations have adapted to different environments and tracing human migration patterns, these databases serve as a fundamental resource for investigating and identifying population-specific disease susceptibilities. In this project, we used two viral and six human genomic databases, as described below.

3.2.1. Viral zone

ViralZone (216) is a publicly available data-sharing platform of viral genomic and proteomic sequences for all known viruses. The information on this website is provided by the SWISS-Prot virus annotation database. ViralZone utilises the Baltimore system (270) for its database virus classification which is based on the nature of the nucleic acids in the virion particle: dsDNA, ssDNA, dsRNA, ss(+)RNA, ss(-)RNA, ssRNA(RT) or ssDNA(RT). The data is organised according to fact sheets that hold comprehensive information about the virus's genome, replication cycle, taxonomy and epidemiology, virion organisation, genome transcription and translation program. ViralZone is constantly being updated with new information extracted from the International Committee on Taxonomy of Viruses (ICTV), and scientific publications.

3.2.2. CoV-GLUE-Viz

CoV-GLUE-Viz (http://cov-glue-viz.cvr.gla.ac.uk) maintains a database of mutations, insertions and deletions which have been observed in GISAID (271) hCoV-19 sequences sampled from the ongoing COVID-19 pandemic. The website provides tailored information according to user's preference such as displays based on one specific or all SARS-CoV-2's lineages, Open Reading Frames, mutation types, details of a specific mutation across all observed lineages and the proportion of a specific mutation.

3.2.3. The Genome Aggregation Database (gnomAD)

The Genome Aggregation Database (gnomAD) browser (https://gnomad.broadinstitute.org/) has deposited data from over 195,000 individuals with no reported disease and contains harmonised sequencing data including features such as allele frequency, per-base expression levels, constraint scores, and variant co-occurrence (272).

One of the main features in gnomAD version 2.1.1 is information on the frequency of a variant in the general population which depends on the number of heterozygous and homozygous individuals. The population frequency for a particular gene provides enrichment of variants within five continental populations (Africans/African Americans, Latino/Admixed American, East Asians, South Asians and non-Finnish European), two demographically distinct populations (Ashkenazi Jewish and Finnish), and any remaining un-categorised (other) samples (Figure 3.11). Although most pathogenic

variants are found to be infrequent, identifying them is the first step in correlating them to specific disease morbidity or the emergence of rare new diseases.

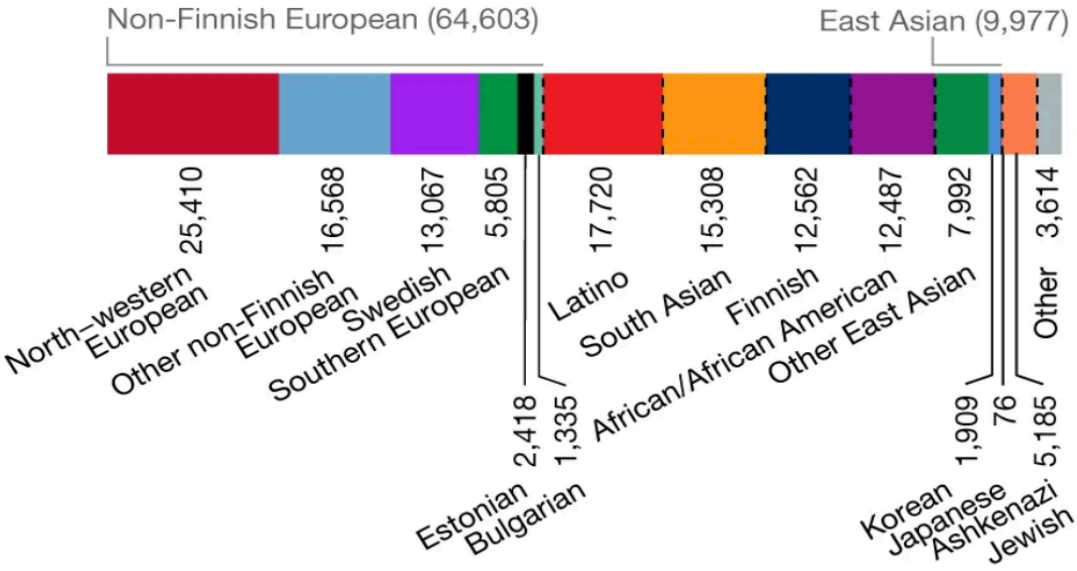

Figure 3.11. The number of people in the gnomAD database, broken down by demographic and subpopulation (272,273).

3.2.4. All of Us

The All of Us is a research program funded by the United States of America National Institutes of Health (NIH) which comprises one million individuals from all backgrounds across the USA to establish a diverse health database. It allows the researchers to understand how the individuals' biology, lifestyle, and environment affect health, prevent disease, and propose new treatments (https://allofus.nih.gov).

3.2.5. SweGen

SweGen data set (274) is a whole-genome data resource of genetic variability in the Swedish population reflecting a cohort of over 1,000 Swedish-born individuals composed of 506 males and 494 females with an average age of 65.2 years at the

time of sampling. The database provides a total of 29.2 million single-nucleotide variants with an average of 7199 individual-specific variants per sample. The database is accessible through https://swefreq.nbis.se.

3.2.6. GenomeAsia 100K

The GenomeAsia 100K Project (275) (http://www.genomeasia100k.com) comprises a whole-genome sequencing reference dataset from 1,739 individuals of 219 population groups and 64 countries across Asia. It reflects genetic variation, population structure, and disease associations. The GenomeAsia 100K consortium includes 598 genomes from India, 156 from Malaysia, 152 from South Korea, 113 from Pakistan, 100 from Mongolia, 70 from China, 70 from Papua New Guinea, 68 from Indonesia, 52 from the Philippines, 35 from Japan and 32 from Russia. About a fifth of sequencing data in GenomeAsia 100K has come from Africans, Europeans, and Americans samples to provide a comprehensive and comparative analysis. This consortium provides broader genetic diversity data of the population in the centre of Asia as they are more underrepresented in worldwide studies.

3.2.7. IndiGenomes

IndiGenomes database (276) (http://clingen.igib.res.in/indigen/) comprises whole genome sequencing of 1029 healthy Indian individuals (534 females and 495 males with a mean age of 32.96 and 41.35 years, respectively) from the different states in India. Analysis of the participants' genomic DNA generated 55,898,122 single allelic genetic variants from geographically distinct Indian genomes. The database provides

calculated allele frequency, allele count, and allele number, as well as the number of heterozygous or homozygous individuals.

3.2.8. jMorp

jMorp (277) (www.jmorp.megabank.tohoku.ac.jp) is a metabolome and proteome database of 5,093 Japanese healthy volunteers (male: 2077, female: 3016) who participated in the Tohoku Medical Megabank Project Cohort Study. This database comprises metabolome data measured by proton nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LC–MS) and proteome data obtained by nano LC–MSin plasma.

3.2.9. Dataset of human:SARS-CoV-2 interactor protein complexes

The human:SARS-CoV-2 interacting proteins were collected from the UniProt-SARS-CoV-2 resource and IntAct's COVID-19 dataset (https://www.ebi.ac.uk/intact/home). The data from both resources are compiled from experimentally proven interactions in the literature. Every interaction in IntAct is assigned with a Molecular Interaction (MI) score (a range from 0 to 1) which is mainly evaluated according to experimental evidence. MI score ≥ 0.45 is considered to be associated with high-confidence interactions (278) thus we used it as a threshold to select suitable complexes for downstream analysis. For the human:SARS-CoV-2 immunity-associated proteins, mapping UniProt IDs to the InnateDB database (http://innatedb.sahmri.com) and evidence from the literature (264) were applied to further shortlist the dataset from 650 to 94 (for example, excluding complexes where the human protein domain interacting with the SARS-CoV-2 protein had no reported mutations in gnomAD).

Models of 94 human proteins interacting with SARS-CoV-2 complexes were built using both the AlphaFold2_ptm and AlphaFold2-multimer (v1) methods, which were made available in July 2021 and March 2022, respectively. A model was then selected from one of these methods—whichever had the best interface quality (1,67). High-confidence models were chosen where models had an overall pLDDT (predicted local difference distance test) ≥ 70, as well as a pTM-Score (predicted TM-score) ≥ 70. Protein–protein models for which high-quality predictions were not achieved were excluded. Complexes were further filtered based on interface quality, including interface-pLDDT (complexes with scores < 70 were excluded) and interface-PAE (predicted alignment error > 10 were excluded). Models exhibiting overlapping or entangled interfaces, as identified through manual inspection, were excluded. Interface stability scores were evaluated using the PIZSA method, which assesses protein interaction stability; a Z-score ≥ 1.5 indicates a stable interface (279). Predicted binding affinities of the complexes were assessed using PROtein binDIng enerGY prediction (PRODIGY) for binding energy (ΔG) scores (280,281) the more negative ΔG, the stronger the binding between the receptor and the ligand. Additional filters reduced the models to 12, and based on further investigation through literature review, the 7 complexes were selected for downstream analysis (Table 3.7).

Human: SCoV2 complex	Type of interaction	Source of the human: SCoV2 complex	Interface pLDDT	Interface PAE	PIZSA Interface stability Z-score	PRODIGY ΔG (Kcal/mol)
hTOM70:ORF9b	immunity-associated	PDB: 7KDT	N/A	N/A	2.507 (stable)	-17.4
hRPS3:Nsp1	immunity-associated	PDB: 6ZMT	N/A	N/A	2.805 (stable)	-4.5
hhPASL1:E-protein	immunity-associated	PDBe ID: 7M4R	N/A	N/A	1.685 (stable)	-6.6
ISG15:PLpro	immunity-associated	PDBe ID: 7RBS	N/A	N/A	2.046 (stable)	-13.4
hACE2:RBD	receptor-binding	PDBe ID: 6M0J (WT SCoV2)	N/A	N/A	2.582 (stable)	-11.9
hIFIT2:PLpro	immunity-associated	AF2	72.82	9.86	2.46 (stable)	-11.3
hIFIH1:PLpro	immunity-associated	AF2- m	83	10	2.375 (stable)	-10.1
hTRIM25:N-protein	immunity-associated	AF2	93.15	2.87	2.24 (stable)	-9.1
hKREMEN1:RBD	receptor-binding	AF2	81.02	4.74	2.148 (stable)	-10.2
hARF6:Nsp15	immunity-associated	AF2	82.61	7.20	2.063 (stable)	-7.5
hTRIMM:Nsp14	immunity-associated	AF2	81.24	7.58	1.974 (stable)	-8.8
hAXL:NTD	receptor-binding	AF2	75.31	8.55	2.321 (stable)	-10.6

Table 3.7. Details of complexes selected for downstream human: SCoV2 interactor protein-protein analysis. PDBe ID of experimentally resolved structures are outlined. The quality of AlphaFold2 models was assessed with algorithms such as Interface pLDDT (acceptable range is ≥70), Interface PAE (acceptable range is <10), PIZSA Interface stability Z-score (acceptable range is >1.5 indicates stable interface) as well as the predicted PRODIGY binding energy (ΔG). Abbreviations: Wild-type (WT), AlphaFold2-advanced (AF2) and AlphaFold-multimer (AF-m). Abbreviations: N-protein and E-protein are SCoV2 Nucleocapsid and Envelope proteins, respectively.

3.2.10. HADDOCK

High Ambiguity Driven biomolecular DOCKing (HADDOCK) is an intuitive docking webserver (282,283) platform with the capacity to perform the docking of up to 20 macromolecules including proteins, small-molecules, nucleic acids, peptides, cyclic peptides, glycans, and glycosylated proteins (284,285). HADDOCK uses experimental and theoretical information derived from interface restraints from NMR, mutagenesis experiments, or bioinformatics predictions (286,287); shape data from small-angle X-ray scattering (288); cryo-electron microscopy experiments (289); orientations of the individual structures in the complex from NMR residual dipolar couplings (290), relaxation anisotropy (291) and pseudo-contact shifts experiments (292) to drive a data-driven docking method and assess the ligand:receptor binding affinity (284,293). The key docking steps are as follows:

i. The 3D structures of the protein and ligand are provided by the user.
ii. HADDOCK uses information about potential interaction sites (ambiguous or unambiguous) to predict the docking models.
iii. After predicting several docking models, it refines these models using energy minimisation and molecular dynamics.
iv. The models are scored based on the overall energy, including binding energy, between the protein and ligand. A lower score indicating a more thermodynamically favourable binding affinity.

3.2.11. Validating the AlphaFold2 models conformation using HADDOCK

HADDOCK was employed to verify AlphaFold2 and AlphaFold-multimers prediction of the complexes for which no experimental structures (*e.g.*, from NMR or X-ray

crystallography) are available in the Protein Data Bank (PDB), at the time of conducting this research. Prior to validating our predicted AlphaFold2/multimers models, we first used HADDOCK to predict the best human:SCoV2 complexes conformation of complexes whose experimental structures are available (according to their corresponding PDB structures). Each protein structure, whether human or SCoV2 was uploaded separately, and the binding residues for each protein in the complex were provided, treating the structures as if they were part of an unknown complex. Then, HADDOCK predicted the most favourable interactions between the two, based on these interface residues. At this stage, our objective was to determine how accurate HADDOCK's predictions are. The best HADDOCK prediction was superposed on the PDB structure of the known structures, and we measured the RMSD of the structural differences (PDB versus HADDOCK's prediction). This was performed on five human:SCoV2 complexes outlined in Table 3.8.

Several published molecular docking validation methods (294–296) have suggested that an RMSD < 2.0 Å corresponds to good docking solutions. As the RMSD values for the five human:SCOV2 experimental complexes were below this suggested RMSD threshold, the same method was applied to the complexes modelled by AlphaFold2/multimers (hARF6:Nsp15, hAXL:NTD, hIFIH1:PLpro, hhIFIT2:PLpro, KREMEN1:RBD, hTRIM25:N, hTRIMM:Nsp14). All RMSD values of the superposed best HADDOCK pose to the predicted AlphaFold2/multimers models were below the threshold and thus they could be used for downstream analysis (Table 3.8).

The outcome of the docking is a list of all possible water-refined structure conformations that are ranked according to their intermolecular energy or HADDOCK

score, the lower the energy the better-predicted conformation. In addition to the HADDOCK score, the Z-score is a statistical measure used to calculate the quality of the predicted protein-protein complex structure. A lower Z-score suggests that the predicted complex structure is more energetically favourable or closer to the likely native or experimentally observed structure; therefore, indicates a better result (284,293).

Validation of the method using the PDB structures				
Complex (human:SCoV2)	**PDB structure**	**HADDOCK score**	**HADDOCK Z-score**	**RMSD (HADDOCK vs PDB complex)**
hTOM70:SCoV2 ORF9b	7KDT	-134.2 +/- 4.0	-2.9	0.445
hISG15:SCoV2 PLpro	7RBS	-130.9 +/- 2.1	-1.7	0.404
hPALS1:SCoV2 Envelope protein	7M4R	-94.4 +/- 3.2	-2.0	0.586
hRPS3:SCoV2 Nsp1	6ZMT	-77.3 +/- 19.0	-1.8	1.137
hACE2:SCoV2 RBD	7A95	-125.4 +/- 3.8	-2.3	0.546

Application of the method on the AF2/m models				
Complex (human:SCoV2)	**AF2 model**	**HADDOCK score**	**HADDOCK Z-score**	**RMSD (HADDOCK vs AF2 complex)**
hARF6:SCoV2 Nsp15	AF2 model	-105.5 +/- 10.2	-1.7	0.459
hAXL SCoV2 NTD	AF2 model	-143.8 +/- 3.3	-2.6	0.462
hIFIH1:SCoV2 PLpro	AF2-multimer	-106.3 +/- 4.3	-1.4	0.405
hIFIT2:SCoV2 PLpro	AF2 model	-135.2 +/- 11.8	-2.3	0.517
hKREMEN1:SCoV2 RBD	AF2 model	-138.0 +/- 1.1	-2.3	0.766
hTRIM25:SCoV2 Nucleocapsid protein	AF2 model	-93.1 +/- 2.9	-1.7	0.467
hTRIMM:SCoV2 Nsp14	AF2 model	-97.6 +/- 2.5	-2.7	0.385

Table 3.8. Verification and application of conformation and interaction in AlphaFold2 modelled complexes. Top table: the PDB complexes were used to validate the method. The HADDOCK predicted docking conformation were superposed on the corresponding PDB structure and assessed using RMSD. Bottom table: following the validation, the method was used to assess the AlphaFold2/multimer modelled complexes of proteins used in this study.

3.2.12. OHM, Allostery site prediction

OHM, a protein structure network-based method, identifies and characterises allosteric communication networks within proteins. OHM's approach is independent of simulation-based methods and is based on the structure of the protein of interest (297). Upon uploading a PDB file, OHM determines all contacts from the tertiary structure of the protein and calculates the average atom-contacts matrix in the given 3D structure. A probability matrix is generated by computing the number of contacts between each pair of residues divided by the number of atoms in each residue. At the next step, the algorithm perturbs residues in the active site and the perturbation is propagated to other residues according to the probability matrix, 10,000 times. Ultimately, the allosteric coupling intensity (ACI) frequency is evaluated. ACI is a floating-point number between zero and one in which ACI $\geq$ 0.85 are considered to be allosteric residues.

3.2.13. Predicting the impact of human and viral protein variants on the binding affinity of the corresponding complexes using mCSM-PPI2

mCSM-PPI2 (method described in the introduction chapter) was employed to predict the impact of missense variants in different human populations reported in human databases: gnomAD (272), SweGen (274), GenomeAsia1000K (275), allofus (https://allofus.nih.gov) by the National Institutes of Health (NIH), jMorp (277) and IndiGenome (276) on the binding affinity of the corresponding complexes. An mCSM-PPI2 threshold cut-of value of $\Delta\Delta G \geq 0.5$ Kcal/mol is implemented to exclude any possible false positives driven by a minor systematic error in the predictions. This threshold is considered to be associated with a significant enhancement of the binding

affinity between the two-interacting proteins in a complex. This threshold has been previously validated with experimental analyses on protein-protein complexes (103,104).

Additional investigations of affinity enhancing mutations were also performed. The degree of conservation of the variant position and the conservation of neighbouring residues within 5Å was determined using Scorecons (25) if the representative protein is a member of a functional family in CATH with diversity of positions (DOPS) score greater than 70 (26,27), as discussed in the introduction chapter. Allosteric site prediction of the corresponding position and neighbouring residues within 5Å was performed using OHM (298), and the probable pathogenicity effect of the missense mutation was determined using MutPred2 (105). See chapter one for details of these methods.

Based on the position of an amino acid in the 3D structure and ability to interact with the corresponding SARS-CoV-2 protein, a mutation was labelled as a direct contact residue in the interface (DC) or a residue whose atoms are within 5Å of direct contact residues (DCEX). DC residues are compiled from PDBSum (120) or identified using a distance cut-off of 4Å (between any atoms), from each interacting chain using the Chimera built-in tool (185).

3.2.14. Predicting the impact of human protein variants on the binding affinity of the human:SCov2 complexes using PROtein binDIng enerGY prediction

PROtein binDIng enerGY prediction (PRODIGY) is a computational tool designed to predict the strength of interaction or binding affinity of proteins in a protein complex based on their 3D structures. It uses a statistical-based approach, considering various biological aspects and molecular interactions.

PRODIGY correlates the number of interfacial contacts in a protein-protein complex with experimental binding affinity. It includes the properties of non-interacting surfaces in the final calculation, as suggested by Kastritis *et al*., (2014). (299), where non-interacting surfaces may influence protein-protein binding affinity. The calculation incorporates factors such as van der Waals interactions, electrostatic forces (both attractive and repulsive) between charged particles, hydrogen bonds, and entropy changes to determine the binding energy (ΔG).

PRODIGY is trained using experimentally measured protein-protein binding affinities datasets, enabling it to identify patterns in the structural features of protein complexes associated with binding affinity. The tool predicts the binding affinity of a given 3D-protein complex based on its structural characteristics, presenting the outcome as a numerical value (ΔG, kcal/mol). A lower or more negative value implies a stronger binding affinity.

In this chapter, PRODIGY was employed to validate predictions made by mCSM-PPI2 for human mutations exhibiting enhanced binding affinity to their corresponding SARS-CoV-2 partners. This involved calculating the binding affinity (ΔG, kcal/mol) for both the wild-type and mutant versions of residues predicted to enhance affinity.

3.2.15. Other algorithms and methods

In addition to the algorithms and methods described above, four other tools were used, which are detailed in the Introduction chapter:

- Grantham matrix (190): to measure the degree of change in the physicochemical properties of wild-type and mutant residues.
- Scorecons (25): to calculate the degree of amino acid variability in each column of a multiple sequence alignment.
- AlphaFold2 (65): to predict protein structures with high accuracy using deep learning techniques, particularly for complexes where no experimental structure is available.
- MutPred2 (105): to predict the pathogenicity of amino acid substitutions based on conservation, as well as physical and chemical properties.

3.3. Results

As described in the Methods section, we collected 322 experimentally determined human protein SARS-CoV-2 interactors reported by Gorden *et al.,* (2020) (264) and after applying several criteria, the impact of amino acid mutations in twelve human proteins interacting with their likely partner in SARS-CoV-2 was evaluated (Table 3.9). For the amino acid mutations predicted to increase affinity to SARS-CoV-2, additional analyses were performed including exploring the degree of conservation using Scorecons, probability of allosteric effects using OHM (score range 0 to 1, score for a probable allosteric site ≥ 0.85) and pathogenicity using MutPred2 to better understand the impact of the mutation on the human protein function and structure.

We analysed the impact of all reported mutations in gnomAD for the selected 12 human proteins. However, in this chapter we only provide detailed summaries for mutations predicted by mCSM-PPI2 (ΔΔG ≥ 0.5 kcal/mol) to be significantly affinity-enhancing to the corresponding SARS-CoV-2 protein (Table 3.9). This threshold was implemented to exclude any possible false positives driven by a minor systematic error in the predictions. The mutations in the human genome, identified by mCSM-PPI2 as potentially enhancing the binding affinity to the corresponding SARS-CoV-2 protein, underwent further validation through an additional binding affinity prediction tool, PRODIGY. This secondary assessment was performed to reaffirm the initial predictions.

We analysed all mutations found in gnomAD or other databases, even if some of them have a low allele frequency (<0.0001) in order to increase awareness of existing

mutations. This could be beneficial, even if a low-frequency mutation was found in only one nation or region. The effects of the affinity-enhancing human mutations on their corresponding SARS-CoV-2 protein interactions were further evaluated. Three selected complexes are described and discussed below (hTOM70: SCoV2 ORF9b, hIFIH1: SCoV2 PLpro and hIFIT2: SCoV2 PLpro), while the remaining results are provided in Appendix 3.

Complex (human: SCoV2)	No. of reported human mutations in gnomAD	No. of human mutations with enhancing-affinity	Number and type of affinity-enhancing mutation to the interface	
hACE2: SCoV2 RBD	162	2	One mutation in DCEX	One mutation >5Å of DC residues
hKREMEN1: SCoV2 RBD	37	2	Two mutations in DCEX	
hAXL: SCoV2 NTD	18	1	One mutation in DC	
hTOM70: SCoV2 ORF9b	25	3	Two mutations in DCs	One mutation in DCEX
hRPS3: SCoV2 Nsp1	9	1	One mutation >5Å of DCs	
hPALS1: SCoV2 Envelope protein	13	2	One mutation in DCEX	One mutation >5Å of DCs
hTRIM25: SCoV2 Nucleocapsid protein	16	1	One mutation in DCEX	
hTRIMM: SCoV2 Nsp14	14	1	One mutation in DCEX	
hARF6: SCoV2 Nsp15	4	1	One mutation >5Å of DC residue	
hISG15: SCoV2 PLpro	15	1	One mutation in DC	
hIFIH1: SCoV2 PLpro	21	2	Two mutations in DC	
hIFIT2: SCoV2 PLpro	67	6	Two mutations in DCEX	Four mutations >5Å of DC residue

Table 3.9. A summary of reported human mutation in each human:SCoV2 complex in gnomAD and the number of affinity-enhancing mutations in the corresponding complex. Abbreviation: directly contacting residues in the interface of the complex (DC), residues within 5Å from the DC residues (DCEX).

3.3.1. Impact of hTOM70 mutations on hTOM70:SARS-CoV-2 ORF9b interaction

TOM70, an outer mitochondrial membrane receptor, comprises of 25 helices and transfers polypeptide chains from the cytosol into the mitochondria (300,301). In addition, TOM70's NTD-pocket is responsible for the activation of host antiviral immune responses through interacting with heat shock protein 90 (HSP90) to mediate TOM70-dependant IFN-I activation (302).

SCoV2 ORF9b, one of the coronavirus's accessory proteins, is encoded by an alternative open reading frame within the nucleocapsid (N) (221,303,304) and interacts with TOM70's CTD hydrophobic pocket. Occupation of TOM70 CTD by SCoV2 ORF9b reduces the binding affinity of TOM70:HSP90 by about 29-fold and consequently suppresses the host innate immunity (305–307).

Of twenty-five reported and analysed mutations for hTOM70, three are predicted to enhance the binding affinity of hTOM70:SCoV2 ORF9b by over 0.5 kcal/mol of which two are in direct contact (Val556Leu and Ala591Thr) and one within 5Å of a direct contact (Lys576Arg) residue (Figure 3.13). While Val556Leu and Lys576Arg only cause an increase in interactions with the corresponding neighbouring residues in hTOM70, Ala591Thr forms additional hydrophobic bonds with Phe69 in SCoV-2 ORF9b (Appendix 3.4). PRODIGY also predicted that these three mutations increase the binding affinity between human:SCoV2 complex. In addition, Val556Leu and Ala591Thr are predicted to be pathogenic by MutPred2 which reports that they may have an impact by altering the coiled-coil domain (Val556Leu) and a predicted metal

binding site (Ala591Thr). Each mutation is observed in a specific human population but with very low frequency (Table 3.10).

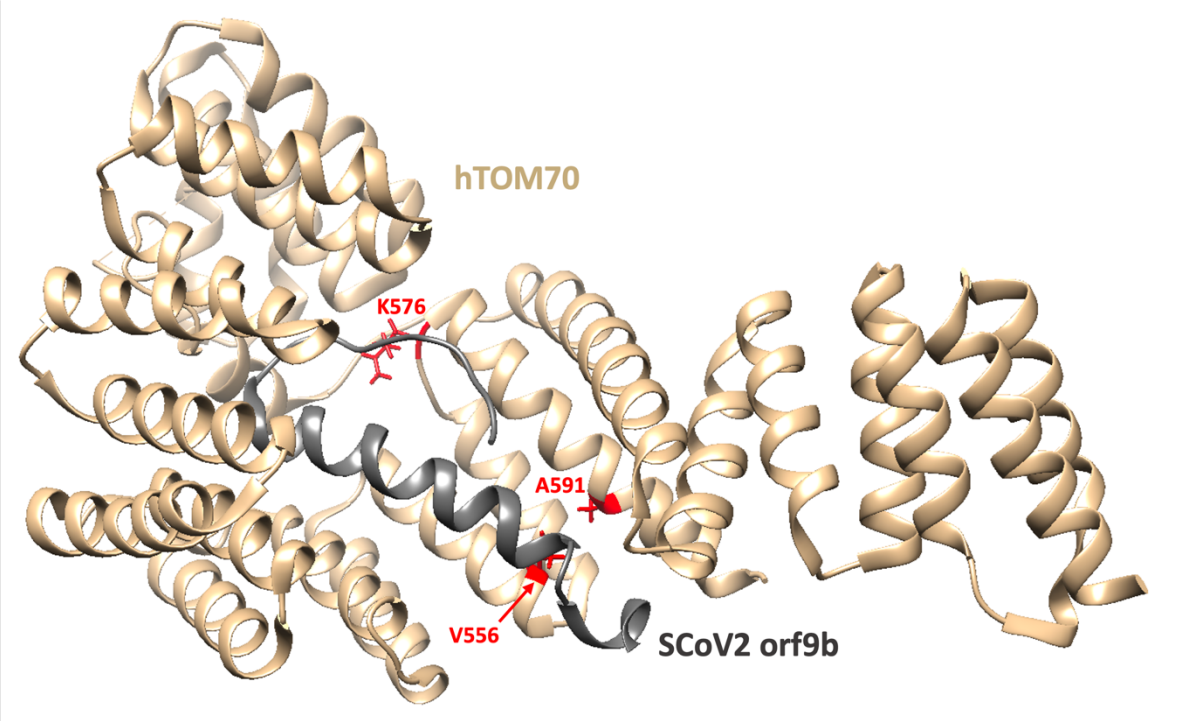

Figure 3.13. Positions of hTOM70 affinity-enhancing residues: V556, K576 and A591 (in red). Positions Val556 and Ala591 are in direct contact and Lys576 is within 5Å of a direct contact residue to SCoV2 ORF9b interface. hTOM70 (tan) and SCoV2 ORF9b (grey), PDBe ID: 7KDT.

hTOM70 variant	Distance to the interface (Å)	mCSM-PPI2 $\Delta\Delta G^{Affinity}$ (kcal/mol)	Changes in hTOM70:SCoV2 ORF9b interaction number of interactions: WT (Mut)					PRODIGY $\Delta G^{Affinity}$ (kcal/mol)	Conservation by Scorecons	Grantham score	Allosteric site (OHM)	Pathogenic (MutPred2)	Allele frequency in gnomAD
			HP	Polar	H-bond	VdW	Carbonyl/ aromatic/ ionic						
V556L	2.69 (DC)	0.905	14 (23)	9 (8)	5 (4)	0 (6)	0 (0)	-11.5	No	32 (conservative)	No	No	Latino/Admixed Americans: 0.00002917
K576R	6.01 (DCEX)	0.618	0 (0)	5 (7)	2 (3)	1 (2)	0 (0)	-11.9	No	32 (conservative)	No	(score = 0.58) altered coiled coil	South Asians: 0.00003267
A591T	3.45 (DC)	0.599	4 (8)	6 (11)	3 (6)	1 (5)	0 (0)	-11.4	No	58 (moderately conservative)	No	(score = 0.535) altered metal binding	Non-Finnish Europeans: 0.000008811

Table 3.10. Summary of three hTOM70 mutations with predicted increase in binding affinity. The distance to the interface, $\Delta\Delta G^{Affinity}$ (kcal/mol) and type of interactions by mCSM-PPI2; PRODIGY binding affinity score for wt-hTOM70:SCoV2 ORF9b is ($\Delta G^{Affinity}$ = -10.5 kcal/mol); conservation by Scorecons (conserved residue and/or within 5Å of conserved residues with Scorecons ≥ 0.9); degree of changes in physicochemical properties by Grantham score; allosteric site prediction by OHM; and predicted pathogenicity by MutPred2. Abbreviations: direct contact residue (DC), within 5Å of a direct contact residue (DCEX), hydrophobic (HP), Hydrogen bond (H-bond), van der Waals contacts (VdW), and Scorecons score (sc). Variants with allele frequency >1% are considered common variants in the population.

3.3.2. SARS-CoV-2 papain-like protease targets hISG15, hIFIH1 and hIFIT2 to supress the immune response

The SARS-CoV-2 papain-like protease (SCoV2 PLpro) is an essential coronavirus functional replicase complex. It enables viral spread by viral polyprotein chain processing and evades the immune system by impairing the type I interferon (IFN-1)-dependant production. The hISG15, hIFIH1 and hIFIT2 (308–310) are primary proteins triggering IFN-1 production. SCoV2 PLpro interacts with these proteins to hinder their functions and suppress the immune response. Affinity-enhancing mutations in selected complexes are discussed below for hIFIH1 and hIFIT2, while the analysis of hISG15 is included in Appendix 3.10.

3.3.3. Impact of hIFIH1 mutations on hIFIH1:SARS-CoV-2 PLpro interaction

IFIH1 (also known as Melanoma differentiation-associated protein 5 or MDA5) is a cytoplasmic innate immune receptor. It detects viral RNA with its C-terminal domain and consequently triggers activation of antiviral immunological genes transcription including IFN-Alpha, IFN-Beta and pro-inflammatory cytokines (309,311). In the case of infection with coronaviruses, SARS-COV-2 employs PLpro to bind and block the activation of an IFIH1-dependent cascade of antiviral responses (309).

Of twenty-one reported and analysed mutations in hIFIH1, mutations in two direct contact residues: Tyr13Asn and Ser16Leu (Appendix 3), are predicted to increase binding affinity to SCoV2 PLpro (ΔΔG=0.835 kcal/mol and ΔΔG=0.769 kcal/mol, respectively). PRODIGY also predicted that these two mutations increase the binding affinity between human:SCoV2 complex (ΔG_{wt} = -10.1 kcal/mol, ΔG_{Y13N}=-11.2 kcal/mol and, ΔG_{S16L} = -11.9 kcal/mol). Both Tyr13Asn and Ser16Leu mutations significantly alter the physicochemical properties of the residue (Grantham score: 143 and 145, respectively, within a range of 5 - 215). MutPred2 predicted that Tyr13Asn and Ser16Leu may result in the loss of allosteric site at the neighbouring residue Phe12 and Arg19, respectively. Loss of allosteric site at Phe12 may also alter conserved metal (*e.g.*, zinc) binding motif, predicted by MutPred2. Tyr13Asn has been identified in three databases: gnomAD (Korean), jMorp, and GenomeAsia100k, indicating that it is unique to East and South Asia, although at low frequency. Ser16Leu is recorded throughout Europe and Africa at low frequency (Table 3.12).

hIFIH1 variant, rsID	Distance to the interface (Å)	mCSM-PPI2 $\Delta\Delta G^{Affinity}$ (kcal/mol)	Changes in hIFIH1:SCoV2 PLpro interaction number of interactions: WT (Mut)					PRODIGY $\Delta G^{Affinity}$ (kcal/mol)	Conservation by Scorecons	Grantham score	Allosteric site (OHM)	Pathogenic (MutPred2)	Allele frequency in gnomAD (unless stated otherwise)
			HP	Polar	H-bond	VdW	Carbonyl/ aromatic/ ionic						
Y13N	3.28 (DC)	0.835	13 (0)	11 (10)	5 (6)	6 (4)	0 (0)	-11.2	No	143 (radical)	No	(Score: 0.64) loss of allosteric site at the neighbouring residue F12 and altered metal binding	Koreans: 0.0005238 Jmoph: 0.00262 GenomeAsia100k: 0.0006
S16L	3.26 (DC)	0.769	0 (8)	9 (4)	4 (4)	4 (2)	0 (0)	-11.9	No	145 (radical)	No	(Score: 0.65) loss of allosteric site at the neighbouring residue R19	North-western Europeans: 0.00002376 Europeans: 0.000007 (allofus) Africans: 0.00002 (allofus)

Table 3.12. Summary of two hIFIH1 mutations with predicted increase in binding affinity. The distance to the interface, $\Delta\Delta G^{Affinity}$ (kcal/mol) and type of interactions by mCSM-PPI2; PRODIGY binding affinity score for wt-hIFIH1:SCoV2 PLpro is ($\Delta G^{Affinity}$ = -10.1 kcal/mol); conservation by Scorecons (conserved residue and/or within 5Å of conserved residues with Scorecons ≥ 0.9); degree of changes in physicochemical properties by Grantham score; allosteric site prediction by OHM; predicted pathogenicity by MutPred2; max population and corresponding allele frequency is according to gnomAD unless stated otherwise; Abbreviations: direct contact residue (DC), hydrophobic (HP), Hydrogen bond (H-bond), van der Waals contacts (VdW), and Scorecons score (sc). Variants with allele frequency >1% are considered common variants in the population.

3.3.4. Impact of hIFIT2 mutations on hIFIT2:SARS-CoV-2 PLpro interaction

Human interferon-induced protein with tetratricopeptide repeats 2 (hIFIT2) is an RNA-binding protein and able to distinguish the viral RNA from the host RNA by recognising 5′ triphosphate (312) or the lack of 2′-O-methylation (313) on viral RNA, and hence, have an inhibitory effect on the expression of viral mRNA. hIFIT2 is involved in triggering IFN signalling (314) and regulation of the production of cytokines in the inflammation process. SCoV2 PLpro interacts with hIFIT2 to hinder its function and suppresses the immune response (314–316).

Of sixty-seven reported and analysed mutations for hIFIT2, six mutations are predicted to enhance the binding affinity of hIFIT2:SCoV2 PLpro by over 0.5 kcal/mol of which three (Leu373Phe, Leu343Thr and Lys221Glu) are within 5Å of direct contact residues and three (Ala239Asp, Ala319Ser and Ala319Thr) are located more than 5Å away from the directly contacting residues. PRODIGY also predicted that these six mutations increase the binding affinity between human:SCoV2 complex. Residues Ala319 and Lys221 are located in the host RNA-binding site where the entry of the host-RNA will be affected if PLpro binds hIFIT2 thus preventing hIFIT2 from triggering IFN signalling. Among these six mutations, Ala239Asp mutation significantly alters the physicochemical properties of the residue (Grantham score: 126, within a range of 5 - 215). All six residues are within 5Å of highly conserved residues (sc ≥ 90) (Figure 3.13, Appendix 3). In addition, these mutations cause an increase in interactions between the two proteins in the interface. Leu373Phe and Lys221Glu are solely reported in Asia, Leu343Phe and Ala239Asp whereas only seen in Europe. Ala319Ser

is only present in the Africans/African Americans population, whereas Ala319Thr is widespread in all four continents (Table 3.14).

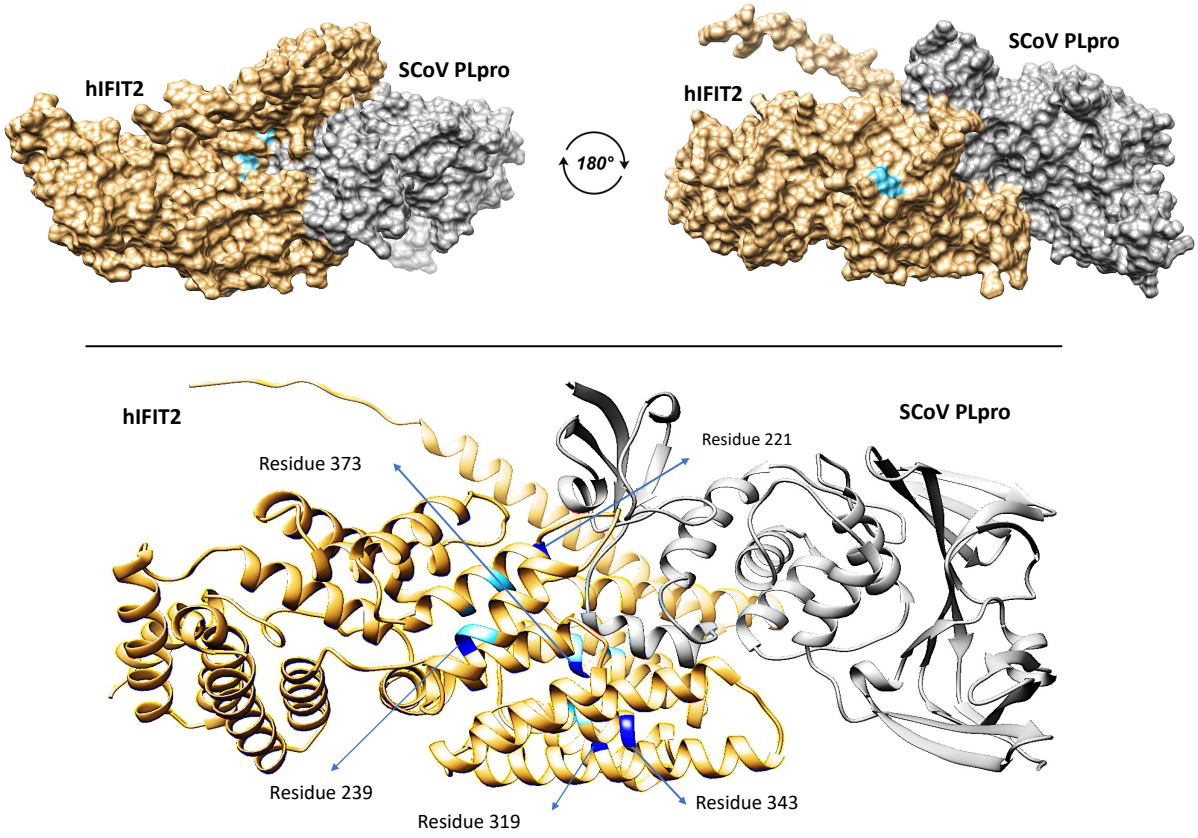

Figure 3.14. The position of affinity enhancing residues (dark blue) in hIFIT2 and their proximity to conserved residues (light blue) in space-filled (top) and ribbon (bottom) models. hIFIT2 (tan) and SCoV-2 PLpro (grey).

hIFIT2 variant	Distance to the interface (Å)	mCSM-PPI2 $\Delta\Delta G^{Affinity}$ (kcal/mol)	Changes in hIFIT2:SCoV2 PLpro interaction number of interactions: WT (Mut)					PRODIGY $\Delta G^{Affinity}$ (kcal/mol)	Conservation by Scorecons	Grantham score	Allosteric site (OHM)	Pathogenic (MutPred2)	Allele frequency in gnomAD (unless stated otherwise)
			HP	Polar	H-bond	VdW	Carbonyl/ aromatic/ ionic						
L373F	8.13 (DCEX)	0.887	23 (37)	5 (11)	1 (6)	0 (7)	0 (12) aromatic	-13.3	No but within 5Å of conserved residues: 338, 372, 375	22 (conservative)	No	No	South Asians: 0.00003268 Asians: 0.00004 (allofus)
L343F	7.38 (DCEX)	0.734	11 (26)	6 (9)	6 (6)	3 (4)	0 (0)	-12.5	No but within 5Å of conserved residues: 338, 339	22 (conservative)	No	No	North-western Europeans: 0.0001163 Europeans: 0.0001 (allofus)
A239D	21.71	0.696	3 (3)	6 (14)	5 (8)	3 (7)	0 (0)	-11.5	No but within 5Å of conserved residues: 213, 238	126 (radical)	No	No	Europeans: 0.00001 (allofus)

K221E rs13052 23950	4.66 (DCEX)	0.627	0 (0)	4 (5)	3 (4)	1 (2)	0 (0)	-12.1	No but within 5Å of conserved residues: 216	56 (moderately conservative)	No	No	South Asians: 0.00003268 Asians: 0.00002 (allofus)
A319S rs77502 7680	11.25	0.532	2 (0)	8 (10)	4 (5)	2 (2)	0 (0)	-11.6	No but within 5Å of conserved residues: 289, 339	99 (moderately conservative)	No	No	Africans/African Americans: 0.00006460
A319T rs77502 7680	11.25	0.516	2 (5)	8 (9)	4 (7)	2 (2)	0 (0)	-12.3	No but within 5Å of conserved residues: 289, 339	58 (moderately conservative)	No	No	Africans/African Americans: 0.00006460 East Asians: 0.00007130 Latino/Admixed Americans: 0.00007008 Swedish: 0.00003832

Table 3.13. Summary of six hIFIT2 mutations with predicted increase in binding affinity. The distance to the interface, $\Delta\Delta G^{Affinity}$ (kcal/mol) and type of interactions by mCSM-PPI2; PRODIGY binding affinity score for wt-hIFIT2:SCoV2 PLpro is ($\Delta G^{Affinity}$ = -11.3 kcal/mol); conservation by Scorecons (conserved residue and/or within 5Å of conserved residues with Scorecons ≥ 0.9); degree of changes in physicochemical properties by Grantham score; allosteric site prediction by OHM; predicted pathogenicity by MutPred2; max population and corresponding allele frequency is according to gnomAD unless stated otherwise; Abbreviations: within 5Å of a direct contact residue (DCEX), hydrophobic (HP), Hydrogen bond (H-bond), van der Waals contacts (VdW), and Scorecons score (sc). Variants with allele frequency >1% are considered common variants in the population.

3.3.5. Impact of SARS-CoV-2 mutations on binding affinity to human interactor proteins

Classification of SARS-CoV-2 emerged lineages as variants of concerns (VOCs) has been mostly based on mutations in the Spike protein. However higher morbidity and mortality are not solely due to a higher binding affinity between human cell receptors and the Spike protein, but also how the mutations in other SARS-CoV-2 proteins suppress and evade the host's innate immune response. A study has shown that higher expression of SCoV2 ORF9b, Nucleocapsid, and ORF6 results in lower production of interferons and impairment of the innate immune system (317).

In addition to the affinity-enhancing mutations in RBD in SCoV2 VOCs (Pro384Leu, Lys417Asn, Glu484Lys, Asn501Tyr, Asp427Asn, Leu452Arg, Thr478Lys) that were experimentally verified using surface plasmon resonance by MacGowan *et al.,* (2022) (318), we detected the impact of over 250 mutations in nine SCoV2 proteins (RBD, NTD, ORF9b, Nsp1, Nsp14, Nsp15, PLpro, envelope and nucleocapsid) on binding affinity to human interactor proteins using mCSM-PPI2. Here, mutations in SARS-CoV-2 proteins with binding affinity changes to human interactors greater than 0.5 kcal/mol are listed (Table 3.14). We identified two mutations in SCoV2 Nsp14: Leu6074Phe and Asn6054Ile predicted to enhance the binding affinity to hTRIMM by ΔΔG=0.609 and 0.53 kcal/mol, respectively. In addition, Leu1774Phe in SCoV2 PLpro is predicted to increase affinity to both hISG15 (ΔΔG=0.728 kcal/mol) and hIFIT2 (ΔΔG=0.71 kcal/mol). As of January 2024, in the SARS-CoV-2 JN.1 variant, a mutation in the Spike-RBD, Leu455Ser, leads to an enhanced binding affinity to ACE2, resulting in a ΔΔG increase of 0.475 kcal/mol (Appendix 3.13).

SCoV-2 variant	SCoV-2 protein	Mutation	mCSM-PPI2 $\Delta\Delta G^{Affinity}$ (kcal/mol)	Human interactor protein
Omicron (B.1.1.529)	Nsp14	L6074F	0.609	hTRIMM
Omicron (XBB.1.5)	Nsp14	N6054I	0.53	hTRIMM
Omicron (BA.4 and BA.5)	PLpro	L1774F	0.728	hISG15
			0.71	hIFIT2
Omicron (JN.1)	RBD	L455S	0.475	hACE2

Table 3.14. SARS-CoV-2 mutations and their effects on human protein interactions.

3.4. Conclusion and future work

In addition to differences in socio-economic and demographic factors, clinically proven host factors such as age, gender, and body-mass index in individuals with no chronic underlying health condition are correlated with a wide range of COVID-19 morbidity (319); however, the influence of host genome variability has been less studied (320).

The human protein variants that enhance the binding affinity to SARS-CoV-2 proteins, either in a receptor that facilitates viral entry into the host cell and causes infection or in a mediator of innate immune-associated pathways which interact with the virus, thereby suppressing the immune response, may play a role in COVID-19 morbidity. This study aimed to understand the impact of changes in binding affinity with SARS-CoV-2 human interactor proteins following amino acid mutations in three human cell-receptors (hACE2, hKREMEN1, and hAXL), seven human proteins mediating immune response (hTOM70, hTRIM25, hTRIMM, hARF6, hISG15, hIFIH1, and hIFIT2), one protein involved in mRNA synthesis in human cells (hRPS3) and one involved in cell polarity in human epithelial cells (hPALS1). Human genetic databases gnomAD, All of us, SweGen, GenomeAsia100K, IndiGenome and jMorp were used to collect information about populations and sub-populations' genetic variations and their allele frequencies (Table 3.15).

The CATH classification groups protein domains into superfamilies and further into Functional Families (FunFams), which consist of relatives that are highly structurally and functionally similar. We identified conserved residues in these FunFams using Scorecons. Binding affinity changes for missense mutations in both human and SARS-

CoV-2 proteins were predicted using mCSM-PPI2 and validated with PRODIGY. Of 23 affinity-enhancing mutations across 12 complexes, many occurred at conserved or contact residues, suggesting possible functional effects. MutPred2 predicted several as potentially pathogenic. Although the frequency of these enhancing-affinity mutations is low (allele frequencies 10^{-6} to 10^{-5}), this may be due to them being underrepresented populations in the five most comprehensive human variation databases. With regard to the affinity enhancing SCoV2 mutations in Nsp14 and PLpro, these two SCoV2 proteins bind to the corresponding human protein interactor to suppress the innate immune response; therefore, monitoring of emerging mutations in these proteins in addition to the Spike protein may be necessary.

The identification of affinity-enhancing mutations particularly at conserved or contact sites in human proteins interacting with SARS-CoV-2 proteins, can inform several practical areas. First, these mutations could serve as genetic biomarkers for increased susceptibility to infection or altered immune response, enabling more personalised risk assessments. Second, insights into how viral proteins like PLpro target interferon-associated proteins (*e.g.*, hISG15, hIFIH1, hIFIT2) may guide the development of antiviral therapies aimed at protecting these host interactions. Lastly, this work supports future studies in population genetics and vaccine response by identifying variants with geographic prevalence that may influence regional differences in COVID-19 severity or outcomes.

At the onset of the COVID-19 pandemic, the CATH team employed multiple computational tools—including mCSM-PPI2—to predict potential interactions between

the SARS-CoV-2 spike protein and ACE2 receptor orthologues from a broad range of mammalian species (321). Our predictions identified American mink (ΔΔG = 0.632 kcal/mol), cattle (ΔΔG = 0.56 kcal/mol), and rabbits (ΔΔG = 0.91 kcal/mol) as being at elevated risk of SARS-CoV-2 infection. Subsequent *in vivo* studies confirmed the susceptibility of both cattle and rabbits to SARS-CoV-2, supporting our computational findings. Moreover, the predictive accuracy of mCSM-PPI2 was experimentally validated by MacGowan *et al.,* (2022), who employed surface plasmon resonance (SPR) assays and adopted the same ΔΔG threshold of 0.5 kcal/mol to evaluate binding affinities (318). Building on this approach, in the present study we applied the same ΔΔG threshold of 0.5 kcal/mol to identify mutations in human proteins that interact with SARS-CoV-2 proteins, where values above this threshold are predicted to result in a stabilising interaction and may enhance viral binding or host susceptibility.

It is essential to emphasise that validating the predicted affinity-enhancing human mutations interacting with SARS-CoV-2 requires experimental mutagenesis and binding affinity assays, such as surface plasmon resonance and isothermal titration calorimetry, particularly for predictions indicating subtle or borderline changes in binding affinity ($0.5 < \Delta\Delta G^{Affinity} < 1$ kcal/mol), as these values, though classified as significant, may still represent modest effects requiring cautious interpretation. Functional cell-based assays may further reveal the biological impact of these mutations, particularly for immune-related proteins like hIFIH1 and hIFIT2. These validations are crucial not only for confirming the predictions but also for potentially repurposing drugs or establishing foundational points for the structure-based design of new medications. Additionally, population-level genetic data could be used to

explore correlations between these variants and clinical outcomes of SARS-CoV-2 infection, supporting their potential relevance in real-world contexts.

Predicting accurate protein–protein complexes remain challenging due to the difficulty in capturing conformational flexibility, which plays a critical role in real biological interactions. In this study, HADDOCK was used as an additional tool to verify the proteins' conformations, following initial complex modelling with AlphaFold2. While HADDOCK produced favourable scores suggesting close-to-native conformations, these results may not fully reflect docking performance. For example, without accounting for side-chain rearrangements, the predicted complexes may overlook critical aspects of native binding interactions. To overcome this, approaches such as ensemble docking (using flexible receptor) and molecular dynamics simulations to sample structural flexibility for a better approximate *in vivo* interaction status.

As well as the common variants, the impact of the rare non-synonymous coding variants has also been of interest since they may reveal the potential for therapeutic targets in under-represented populations. Therefore, continued efforts to increase the global representation in genetic studies, as well as the compilation of disease-relevant phenotypic information should be prioritised. Accessing relevant medical data recorded in the BioBank allows a better understanding of host genetic factors for COVID-19 susceptibility and may provide more precise insight into therapeutic development and drug repurposing (322–325).

In summary, to the best of our knowledge, this work is one of the most comprehensive studies of its kind, as we used all available human genetic databases to ensure the inclusion of populations that are under-represented in gnomAD. It allowed us to provide a more comprehensive understanding of mutation frequencies across diverse global populations. The approach in this study can be applied to analyse other experimentally determined proteins or complexes predicted by recent advanced structural prediction methods, not only for COVID-19, but also for other persistent and emerging infectious diseases.

human:SCoV2 complex	The position of affinity-enhancing mutations to the protein interface
hACE2: SCoV2 RBD	- G326E is a DCEX residue to the interface. - V447F is a conserved residue in a cluster of conserved residues.
hKREMEN1: SCoV2 RBD	- V191I is a DC residue and within 5Å of three conserved residues. - Y68H is a DCEX and conserved residue, and in a cluster of other conserved residues.
hAXL: SCoV2 NTD	- V38M is a DC residue.
hTOM70: SCoV2 ORF9b	- V556L is a DC residue. - L576R is a DCEX residue to the interface. L576R is predicted to have pathogenetic impact on the protein function. - A591T is a DC residue. A591T is predicted to have pathogenetic impact on the protein function.
hRPS3: SCoV2 Nsp1	- V91I is within 5A of one conserved and a putative allosteric position.
hPALS1: SCoV2 Envelope protein	- L484F is a conserved residue, and in a cluster of other conserved residues. - L321F is a DCEX residue to the interface. L321F is predicted to have pathogenetic impact on the protein function.
hTRIM25: SCoV2 Nucleocapsid protein	- A466T is a DCEX residue to the interface and conserved residue, and in a cluster of other conserved residues. A466T is predicted to have pathogenetic impact on the protein function.
hTRIMM: SCoV2 Nsp14	- I105F is a DCEX residue to the interface and predicted allosteric site, and in a cluster of other predicted allosteric sites.
hARF6: SCoV2 Nsp15	- L166F is a conserved residue, and in a cluster of other conserved and predicted allosteric residues. L166F is predicted to have pathogenetic impact on the protein function.
hISG15: SCoV2 PLpro	- L121Q is a DC residue and within 5Å of a conserved and probable allosteric site.
hIFIH1: SCoV2 PLpro	- Y13N is a DC residue. Y13N is predicted to have pathogenetic impact on the protein function. - S16L is a DC residue. S16L is predicted to have pathogenetic impact on the protein function.
hIFIT2: SCoV2 PLpro	- L373F is a DCEX residue to the interface and within 5Å of three conserved residues. - L343F is a DCEX residue to the interface and within 5Å of three conserved residues. - A239D is and within 5Å of three conserved residues. - K221E is a DCEX residue to the interface and within 5Å of three conserved residues. - A319S/T is and within 5Å of three conserved residues.

Table 3.15. Summary of affinity-enhancing mutations in human proteins and corresponding SCoV2 protein interactors and their associated functional features. Abbreviations: directly-contacting interface residues (DC), residues within 5Å from the DC (DCEX).

Chapter 4: Identifying driver mutations in lung adenocarcinoma

Chapter 5: Conclusions and future directions

Genes carry information that is translated into proteins, which perform numerous vital functions in biological systems. Genetic mutations, or changes in the DNA sequence, can significantly alter protein structure and function, leading to diverse outcomes. In nature, mutations contribute to genetic diversity, producing effects that range from the emergence of novel and beneficial enzymes to detrimental changes that may put the host microorganism at risk of extinction. In the context of health and disease, mutations can influence susceptibility, making some individuals more resistant or more vulnerable to specific illnesses. Additionally, genetic mutations can modify proteins in ways that cause healthy cells to become cancerous, driving disease progression and malignancy. These varied impacts highlight the central role of genetic mutations in evolution, health, and disease.

In Chapter 3, this research investigated the impact of amino acid changes in SARS-CoV-2 and human proteins on the binding affinity of human:SARS-CoV-2 protein complexes, particularly in the context of the COVID-19 pandemic. Missense variants from various ethnic groups were compiled using publicly available human and viral databases. Twelve human:SARS-CoV-2 complexes were analysed, selected for their roles in host cell entry, immune response mediation, and cellular translation machinery. These structures were sourced from the Protein Data Bank or built using AlphaFold2-multimer.

Using computational algorithms, the analysis predicted that 16 human missense variants across 12 human proteins increased binding affinity to SARS-CoV-2 proteins ($\Delta\Delta G \geq 0.5$ kcal/mol). Of these, three proteins were linked to spike-binding, immune

responses, or cellular translation. Additionally, among over 200 SARS-CoV-2 missense residues (including RBD), five viral mutations were found to enhance binding affinity with human proteins involved in immune response and cell receptor interactions. This research highlighted how genetic variation in both humans and SARS-CoV-2 may influence infection and immune responses, providing insights into COVID-19 susceptibility across diverse ethnic groups.

The field of bioinformatics is advancing rapidly, driven by the exponential growth of metagenomic data, a method for analysing genetic material directly from environmental samples, which has significantly expanded our understanding of microbial communities. Additionally, breakthroughs in high-throughput technologies, such as AlphaFold2, have set new benchmarks in protein structure prediction. AlphaFold2 empowers researchers to study proteins whose structures are challenging to determine using conventional in vitro imaging techniques.

These advancements, coupled with the dynamic progress of machine learning and deep learning algorithms, are unlocking the potential of large-scale datasets by revealing patterns and correlations previously undetectable. Such innovations are especially crucial for understanding mutations, which have long been recognised as fundamental drivers of evolution, genetic variation across populations, and biodiversity. Mutations also play a pivotal role in human health and disease, contributing to the development of conditions like cancer, genetic disorders, and infectious diseases.

AlphaFold3 (395), released by DeepMind in May 2024, elevates protein structure prediction by enabling the modelling of interactions between proteins and other

biomolecules, including DNA, RNA, small molecules, ions, and modified residues. This advancement is likely to significantly enhance our understanding of complex biological processes and accelerates progress in drug discovery, molecular biology research, and areas like personalised medicine and synthetic biology. As bioinformatics continues to evolve, these technological strides are poised to unravel the complexities of biology, profoundly influencing science, medicine, and technology.

References

1. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679–82.
2. Taylor WR. The classification of amino acid conservation. J Theor Biol. 1986 Mar 21;119(2):205–18.
3. Skipper L. PROTEINS I Overview. In: Encyclopedia of Analytical Science [Internet]. Elsevier; 2005 [cited 2025 Jan 18]. p. 344–52. Available from: https://linkinghub.elsevier.com/retrieve/pii/B0123693977004933
4. Marnett LJ, Goodwin DC, Rowlinson SW, Kalgutkar AS, Landino LM. Structure, Function, and Inhibition of Prostaglandin Endoperoxide Synthases. In: Comprehensive Natural Products Chemistry [Internet]. Elsevier; 1999 [cited 2025 Jan 18]. p. 225–61. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780080912837001156
5. Riederer JM, Tiso S, van Eldijk TJB, Weissing FJ. Capturing the facets of evolvability in a mechanistic framework. Trends Ecol Evol. 2022 May;37(5):430–9.
6. Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Zool. 1970 Jun;19(2):99.
7. Studer RA, Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? Trends Genet TIG. 2009 May;25(5):210–6.
8. Hardison RC. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med. 2012 Dec 1;2(12):a011627.
9. Opazo JC, Hoffmann FG, Storz JF. Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals. Proc Natl Acad Sci. 2008 Feb 5;105(5):1590–5.
10. Lecomte JTJ, Vuletich DA, Lesk AM. Structural divergence and distant relationships in proteins: evolution of the globins. Curr Opin Struct Biol. 2005 Jun;15(3):290–301.
11. Hoffmann FG, Opazo JC, Storz JF. Whole-Genome Duplications Spurred the Functional Diversification of the Globin Gene Superfamily in Vertebrates. Mol Biol Evol. 2012 Jan;29(1):303–12.
12. Tamura M, D'haeseleer P. Microbial genotype–phenotype mapping by class association rule mining. Bioinformatics. 2008 Jul 1;24(13):1523–9.
13. Trivedi R, Nagarajaram HA. Amino acid substitution scoring matrices specific to intrinsically disordered regions in proteins. Sci Rep. 2019 Nov 8;9(1):16380.
14. Katoh K, Misawa K, Kuma K ichi, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059–66.
15. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr 1;30(4):772–80.
16. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2010/03/02 ed. 2009 Sep;10(6):402–15.
17. Lyngso RB, Pedersen CN, Nielsen H. Metrics and similarity measures for hidden Markov models. Proc Int Conf Intell Syst Mol Biol. 2000/04/29 ed. 1999;178–86.

18. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2004/11/09 ed. 2005 Apr 1;21(7):951–60.
19. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018/06/16 ed. 2018 Jul 2;46(W1):W200–4.
20. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998 Oct 1;14(9):755–63.
21. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011/12/27 ed. 2011 Dec 25;9(2):173–5.
22. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma Oxf Engl. 2012 Dec 1;28(23):3150–2.
23. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006 Jul 1;22(13):1658–9.
24. Dogan T, Karacali B. Automatic identification of highly conserved family regions and relationships in genome wide datasets including remote protein sequences. PLoS One. 2013/09/27 ed. 2013;8(9):e75458.
25. Valdar WS. Scoring residue conservation. Proteins. 2002/07/12 ed. 2002 Aug 1;48(2):227–41.
26. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinforma Oxf Engl. 2016 Sep 15;32(18):2889.
27. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017 Jan 4;45(D1):D289–95.
28. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997 Jul;28(3):405–20.
29. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D412–9.
30. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.
31. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011/11/01 ed. 2011 Oct;7(10):e1002195.
32. Xie L, Bourne PE. Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models. Thornton J, editor. PLoS Comput Biol. 2005 Aug 19;1(3):e31.
33. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature. 1994 Dec 15;372(6507):631–4.
34. Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, et al. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. Nucleic Acids Res. 2014 Jan;42(Database issue):D240-245.
35. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol. 1996;266:617–35.
36. Michie AD, Orengo CA, Thornton JM. Analysis of Domain Structural Class Using an Automated Class Assignment Protocol. J Mol Biol. 1996 Sep;262(2):168–85.

37. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L, et al. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res. 1999 Jan 1;27(1):275–9.
38. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2021 Jan 8;49(D1):D266–73.
39. Das S, Sillitoe I, Lee D, Lees JG, Dawson NL, Ward J, et al. CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic Acids Res. 2015 Jul 1;43(W1):W148-153.
40. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Struct Lond Engl 1993. 1997 Aug 15;5(8):1093–108.
41. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, et al. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 2007 Jan;35(Database issue):D291-297.
42. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol. 1989 Jul;208(1):1–22.
43. Wu X. [Biological characteristics, relationship between antigen and serology of tupaial adenoviruses (type I and II)]. Zhongguo Yi Xue Ke Xue Yuan Xue Bao. 1990 Apr;12(2):153–6.
44. Popov I, Nenov A, Petrov P, Vassilev D. Bioinformatics in Proteomics: A Review on Methods and Algorithms. Biotechnol Biotechnol Equip. 2009 Jan;23(1):1115–20.
45. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D884–91.
46. Eddy SR. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol. 1995/01/01 ed. 1995;3:114–20.
47. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. Protein Sci Publ Protein Soc. 2004 Apr;13(4):1071–87.
48. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019 Dec;20(1):473.
49. Adeyelu T, Bordin N, Waman VP, Sadlej M, Sillitoe I, Moya-Garcia AA, et al. KinFams: De-Novo Classification of Protein Kinases Using CATH Functional Units. Biomolecules. 2023 Feb 2;13(2):277.
50. Lewis TE, Sillitoe I, Lees JG. cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. Hancock J, editor. Bioinformatics. 2019 May 15;35(10):1766–7.
51. Lee DA, Rentzsch R, Orengo C. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res. 2009/11/20 ed. 2010 Jan;38(3):720–37.
52. Richter PK, Blázquez-Sánchez P, Zhao Z, Engelberger F, Wiebeler C, Künze G, et al. Structure and function of the metagenomic plastic-degrading polyester hydrolase PHL7 bound to its product. Nat Commun. 2023 Apr 5;14(1):1905.
53. Jahanshahi DA, Ariaeenejad S, Kavousi K. A metagenomic catalog for exploring the plastizymes landscape covering taxa, genes, and proteins. Sci Rep. 2023 Sep 25;13(1):16029.

54. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 2019/11/07 ed. 2020 Jan 8;48(D1):D570–8.
55. Nim S, Jeon J, Corbi-Verge C, Seo MH, Ivarsson Y, Moffat J, et al. Pooled screening for antiproliferative inhibitors of protein-protein interactions. Nat Chem Biol. 2016 Apr;12(4):275–81.
56. Fuller JC, Burgoyne NJ, Jackson RM. Predicting druggable binding sites at the protein-protein interface. Drug Discov Today. 2009 Feb;14(3–4):155–61.
57. Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. Annu Rev Biochem. 2015;84:551–75.
58. Dutta S, Berman HM. Large macromolecular complexes in the Protein Data Bank: a status report. Struct Lond Engl 1993. 2005 Mar;13(3):381–8.
59. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296–303.
60. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 Dec 5;234(3):779–815.
61. Waterhouse AM, Studer G, Robin X, Bienert S, Tauriello G, Schwede T. The structure assessment web server: for proteins, complexes and more. Nucleic Acids Res. 2024 Jul 5;52(W1):W318–23.
62. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008 Dec;9(1):40.
63. Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo—distance constraints applied on model quality estimation. Elofsson A, editor. Bioinformatics. 2020 Mar 1;36(6):1765–71.
64. Studer G, Biasini M, Schwede T. Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinformatics. 2014 Sep 1;30(17):i505–11.
65. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.
66. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021 Aug 20;373(6557):871–6.
67. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer [Internet]. 2021 [cited 2025 Jan 17]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034
68. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. Proteins. 2007;69 Suppl 8:194–207.
69. Barrett AJ. Enzyme Nomenclature. Recommendations 1992: Supplement 2: Corrections and Additions (1994). Eur J Biochem. 1995 Aug;232(1):1–1.
70. Egelhofer V, Schomburg I, Schomburg D. Automatic assignment of EC numbers. PLoS Comput Biol. 2010 Jan 29;6(1):e1000661.
71. Hu QN, Zhu H, Li X, Zhang M, Deng Z, Yang X, et al. Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. PloS One. 2012;7(12):e52901.

72. Yamanishi Y, Hattori M, Kotera M, Goto S, Kanehisa M. E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinforma Oxf Engl. 2009 Jun 15;25(12):i179-186.
73. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J Am Chem Soc. 2004 Dec 22;126(50):16487–98.
74. McDonald AG, Tipton KF. Fifty-five years of enzyme classification: advances and difficulties. FEBS J. 2014 Jan;281(2):583–92.
75. Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Res. 2008 Jan;36(Database issue):D440-444.
76. Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011 Feb 10;470(7333):187–97.
77. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015 Jul;13(7):e1002195.
78. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.
79. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2016 Jan 4;44(D1):D481-487.
80. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(D1):D353–61.
81. Martinez-Outschoorn UE, Peiris-Pagés M, Pestell RG, Sotgia F, Lisanti MP. Cancer metabolism: a therapeutic perspective. Nat Rev Clin Oncol. 2017 Jan;14(1):11–31.
82. Méndez-Lucas A, Lin W, Driscoll PC, Legrave N, Novellasdemunt L, Xie C, et al. Identifying strategies to target the metabolic flexibility of tumours. Nat Metab. 2020 Apr 21;2(4):335–50.
83. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D433-437.
84. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–13.
85. Prindle MJ, Fox EJ, Loeb LA. The mutator phenotype in cancer: molecular mechanisms and targeting strategies. Curr Drug Targets. 2010 Oct;11(10):1296–303.
86. Coates PJ, Lorimore SA, Wright EG. Cell and tissue responses to genotoxic stress. J Pathol. 2005 Jan;205(2):221–35.
87. Maser RS, DePinho RA. Connecting chromosomes, crisis, and cancer. Science. 2002 Jul 26;297(5581):565–9.
88. Rouse J, Jackson SP. Interfaces between the detection, signaling, and repair of DNA damage. Science. 2002 Jul 26;297(5581):547–51.
89. Kolodner RD, Putnam CD, Myung K. Maintenance of genome stability in Saccharomyces cerevisiae. Science. 2002 Jul 26;297(5581):552–7.

90. Loeb LA, Springgate CF, Battula N. Errors in DNA replication as a basis of malignant changes. Cancer Res. 1974 Sep;34(9):2311–21.
91. Alvarez S, Cigudosa JC. Gains, losses and complex karyotypes in myeloid disorders: a light at the end of the tunnel. Hematol Oncol. 2005 Mar;23(1):18–25.
92. Kurzrock R, Kantarjian HM, Druker BJ, Talpaz M. Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics. Ann Intern Med. 2003 May 20;138(10):819–30.
93. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nat Rev Methods Primer. 2021 Aug 26;1(1):59.
94. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. 2012 Aug 17;337(6096):816–21.
95. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science. 2013 Feb 15;339(6121):819–23.
96. McAfee JC, Bell JL, Krupa O, Matoba N, Stein JL, Won H. Focus on your locus with a massively parallel reporter assay. J Neurodev Disord. 2022 Dec;14(1):50.
97. Hu Z, Yu C, Furutsuki M, Andreoletti G, Ly M, Hoskins R, et al. VIPdb, a genetic Variant Impact Predictor Database. Hum Mutat. 2019 Sep;40(9):1202–14.
98. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic Acids Res. 2019 Jul 2;47(W1):W338–44.
99. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan 4;44(D1):D7–19.
100. Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J Mol Biol. 2017 Feb 3;429(3):365–71.
101. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005 Jul 1;33(Web Server):W382–8.
102. Thiltgen G, Goldstein RA. Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. Deane CM, editor. PLoS ONE. 2012 Oct 29;7(10):e46084.
103. Barton MI, MacGowan SA, Kutuzov MA, Dushek O, Barton GJ, van der Merwe PA. Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. eLife. 2021 Aug 26;10:e70658.
104. Thorne LG, Bouhaddou M, Reuschl AK, Zuliani-Alvarez L, Polacco B, Pelin A, et al. Evolution of enhanced innate immune evasion by SARS-CoV-2. Nature. 2022 Feb 17;602(7897):487–95.
105. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020 Nov 20;11(1):5918.
106. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

107. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W452-457.
108. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–5.
109. Zhong Y, Ye Q, Chen C, Wang M, Wang H. Ezh2 promotes clock function and hematopoiesis independent of histone methyltransferase activity in zebrafish. Nucleic Acids Res. 2018 Apr 20;46(7):3382–99.
110. Stephenson JD, Laskowski RA, Nightingale A, Hurles ME, Thornton JM. VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. Elofsson A, editor. Bioinformatics. 2019 Nov 1;35(22):4854–6.
111. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D745–51.
112. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D988–95.
113. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Dec;17(1):122.
114. Bateman A, Martin MJ, Orchard S, Magrane M, The UniProt Consortium, Adesina A, et al. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2025 Jan 6;53(D1):D609–17.
115. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.
116. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011 Jul 23;2011(0):bar030–bar030.
117. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D786–92.
118. Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res. 2018 Jan 4;46(D1):D618–23.
119. Pearson WR. BLAST and FASTA similarity searching for multiple sequence alignment. Methods Mol Biol Clifton NJ. 2014;1079:75–101.
120. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci Publ Protein Soc. 2018 Jan;27(1):129–34.
121. Lek M, Exome Aggregation Consortium, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.
122. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062–7.
123. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 2013 Jan;41(Database issue):D1096-1103.

124. Shoemaker BA, Zhang D, Tyagi M, Thangudu RR, Fong JH, Marchler-Bauer A, et al. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. Nucleic Acids Res. 2012 Jan;40(Database issue):D834-840.
125. Sinha V, Patel MR, Patel JV. Pet Waste Management by Chemical Recycling: A Review. J Polym Environ. 2010 Mar;18(1):8–25.
126. Soong YHV, Sobkowicz MJ, Xie D. Recent Advances in Biological Recycling of Polyethylene Terephthalate (PET) Plastic Wastes. Bioeng Basel Switz. 2022 Feb 27;9(3):98.
127. Eriksen M, Lebreton LCM, Carson HS, Thiel M, Moore CJ, Borerro JC, et al. Plastic Pollution in the World's Oceans: More than 5 Trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea. Dam HG, editor. PLoS ONE. 2014 Dec 10;9(12):e111913.
128. Jambeck JR, Geyer R, Wilcox C, Siegler TR, Perryman M, Andrady A, et al. Plastic waste inputs from land into the ocean. Science. 2015 Feb 13;347(6223):768–71.
129. Sui B, Wang T, Fang J, Hou Z, Shu T, Lu Z, et al. Recent advances in the biodegradation of polyethylene terephthalate with cutinase-like enzymes. Front Microbiol. 2023 Oct 2;14:1265139.
130. Hopewell J, Dvorak R, Kosior E. Plastics recycling: challenges and opportunities. Philos Trans R Soc B Biol Sci. 2009 Jul 27;364(1526):2115–26.
131. Kawai F. Emerging Strategies in Polyethylene Terephthalate Hydrolase Research for Biorecycling. ChemSusChem. 2021 Oct 5;14(19):4115–22.
132. Skoczinski P, Krause L, Raschka A, Dammer L, Carus M. Current status and future development of plastics: Solutions for a circular economy and limitations of environmental degradation. In: Methods in Enzymology [Internet]. Elsevier; 2021 [cited 2025 Jan 18]. p. 1–26. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0076687920303633
133. Ribitsch D, Herrero Acero E, Greimel K, Dellacher A, Zitzenbacher S, Marold A, et al. A New Esterase from Thermobifida halotolerans Hydrolyses Polyethylene Terephthalate (PET) and Polylactic Acid (PLA). Polymers. 2012 Feb 21;4(1):617–29.
134. Sulaiman S, Yamato S, Kanaya E, Kim JJ, Koga Y, Takano K, et al. Isolation of a Novel Cutinase Homolog with Polyethylene Terephthalate-Degrading Activity from Leaf-Branch Compost by Using a Metagenomic Approach. Appl Environ Microbiol. 2012 Mar;78(5):1556–62.
135. Carniel A, Valoni É, Nicomedes J, Gomes ADC, Castro AMD. Lipase from Candida antarctica (CALB) and cutinase from Humicola insolens act synergistically for PET hydrolysis to terephthalic acid. Process Biochem. 2017 Aug;59:84–90.
136. Tanasupawat S, Takehana T, Yoshida S, Hiraga K, Oda K. Ideonella sakaiensis sp. nov., isolated from a microbial consortium that degrades poly(ethylene terephthalate). Int J Syst Evol Microbiol. 2016 Aug 1;66(8):2813–8.
137. Ribitsch D, Guebitz GM. Tuning of adsorption of enzymes to polymer. In: Methods in Enzymology [Internet]. Elsevier; 2021 [cited 2025 Jan 18]. p. 293–315. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0076687920303761

138. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, et al. MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D753–9.
139. Elmi F, Lee HT, Huang JY, Hsieh YC, Wang YL, Chen YJ, et al. Stereoselective esterase from Pseudomonas putida IFO12996 reveals alpha/beta hydrolase folds for D-beta-acetylthioisobutyric acid synthesis. J Bacteriol. 2005 Dec;187(24):8470–6.
140. Carr PD, Ollis DL. Alpha/beta hydrolase fold: an update. Protein Pept Lett. 2009;16(10):1137–48.
141. Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, et al. The alpha/beta hydrolase fold. Protein Eng. 1992 Apr;5(3):197–211.
142. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol. 1996;266:617–35.
143. Dimitriou PS, Denesyuk A, Takahashi S, Yamashita S, Johnson MS, Nakayama T, et al. Alpha/beta-hydrolases: A unique structural motif coordinates catalytic acid residue in 40 protein fold families: DIMITRIOU et al. Proteins Struct Funct Bioinforma. 2017 Oct;85(10):1845–55.
144. Hotelier T, Renault L, Cousin X, Negre V, Marchot P, Chatonnet A. ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D145-147.
145. Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). Science. 2016 Mar 11;351(6278):1196–9.
146. Li T, Menegatti S, Crook N. Breakdown of POLYETHYLENE THEREPTHALATE microplastics under saltwater conditions using engineered *Vibrio natriegens*. AIChE J. 2023 Dec;69(12):e18228.
147. Joo S, Cho IJ, Seo H, Son HF, Sagong HY, Shin TJ, et al. Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation. Nat Commun. 2018/01/28 ed. 2018 Jan 26;9(1):382.
148. Austin HP, Allen MD, Donohoe BS, Rorrer NA, Kearns FL, Silveira RL, et al. Characterization and engineering of a plastic-degrading aromatic polyesterase. Proc Natl Acad Sci U S A. 2018 May 8;115(19):E4350–7.
149. Matak MY, Moghaddam ME. The role of short-range Cys171-Cys178 disulfide bond in maintaining cutinase active site integrity: a molecular dynamics simulation. Biochem Biophys Res Commun. 2009 Dec 11;390(2):201–4.
150. Robles-Martín A, Amigot-Sánchez R, Fernandez-Lopez L, Gonzalez-Alfonso JL, Roda S, Alcolea-Rodriguez V, et al. Sub-micro- and nano-sized polyethylene terephthalate deconstruction with engineered protein nanopores. Nat Catal. 2023 Oct 19;6(12):1174–85.
151. Lykidis A, Mavromatis K, Ivanova N, Anderson I, Land M, DiBartolo G, et al. Genome Sequence and Analysis of the Soil Cellulolytic Actinomycete *Thermobifida fusca* YX. J Bacteriol. 2007 Mar 15;189(6):2477–86.
152. Kleeberg I, Welzel K, Vandenheuvel J, Muller RJ, Deckwer WD. Characterization of a new extracellular hydrolase from Thermobifida fusca degrading aliphatic-aromatic copolyesters. Biomacromolecules. 2005/01/11 ed. 2005 Jan;6(1):262–70.
153. Furukawa M, Kawakami N, Tomizawa A, Miyamoto K. Efficient Degradation of Poly(ethylene terephthalate) with Thermobifida fusca Cutinase Exhibiting

Improved Catalytic Activity Generated using Mutagenesis and Additive-based Approaches. Sci Rep. 2019/11/07 ed. 2019 Nov 5;9(1):16038.
154. Son HF, Joo S, Seo H, Sagong HY, Lee SH, Hong H, et al. Structural bioinformatics-based protein engineering of thermo-stable PETase from Ideonella sakaiensis. Enzyme Microb Technol. 2020 Nov;141:109656.
155. Buchholz PCF, Feuerriegel G, Zhang H, Perez-Garcia P, Nover L, Chow J, et al. Plastics degradation by hydrolytic enzymes: The PLASTICS-ACTIVE enzymes database— PAZY. Proteins Struct Funct Bioinforma. 2022 Jul;90(7):1443–56.
156. Wallace NE, Adams MC, Chafin AC, Jones DD, Tsui CL, Gruber TD. The highly crystalline PET found in plastic water bottles does not support the growth of the PETASE -producing bacterium *Ideonella sakaiensis*. Environ Microbiol Rep. 2020 Oct;12(5):578–82.
157. Gregory MR, Andrady AL. Plastics in the Marine Environment. In: Andrady AL, editor. Plastics and the Environment [Internet]. 1st ed. Wiley; 2003 [cited 2025 Jan 18]. p. 379–401. Available from: https://onlinelibrary.wiley.com/doi/10.1002/0471721557.ch10
158. Fukushima K, Coulembier O, Lecuyer JM, Almegren HA, Alabdulrahman AM, Alsewailem FD, et al. Organocatalytic depolymerization of poly(ethylene terephthalate). J Polym Sci Part Polym Chem. 2011 Mar;49(5):1273–81.
159. Araújo R, Silva C, O'Neill A, Micaelo N, Guebitz G, Soares CM, et al. Tailoring cutinase activity towards polyethylene terephthalate and polyamide 6,6 fibers. J Biotechnol. 2007 Mar;128(4):849–57.
160. Silva C, Da S, Silva N, Matamá T, Araújo R, Martins M, et al. Engineered *Thermobifida fusca* cutinase with increased activity on polyester substrates. Biotechnol J. 2011 Oct;6(10):1230–9.
161. Wei R, Oeser T, Schmidt J, Meier R, Barth M, Then J, et al. Engineered bacterial polyester hydrolases efficiently degrade polyethylene terephthalate due to relieved product inhibition. Biotechnol Bioeng. 2016 Aug;113(8):1658–65.
162. Zumstein MT, Rechsteiner D, Roduner N, Perz V, Ribitsch D, Guebitz GM, et al. Enzymatic Hydrolysis of Polyester Thin Films at the Nanoscale: Effects of Polyester Structure and Enzyme Active-Site Accessibility. Environ Sci Technol. 2017 Jul 5;51(13):7476–85.
163. Mueller RJ. Biological degradation of synthetic polyesters—Enzymes as potential catalysts for polyester recycling. Process Biochem. 2006 Oct;41(10):2124–8.
164. Brott S, Pfaff L, Schuricht J, Schwarz J, Böttcher D, Badenhorst CPS, et al. Engineering and evaluation of thermostable *Is* PETase variants for PET degradation. Eng Life Sci. 2022 Mar;22(3–4):192–203.
165. Wei R, Breite D, Song C, Gräsing D, Ploss T, Hille P, et al. Biocatalytic Degradation Efficiency of Postconsumer Polyethylene Terephthalate Packaging Determined by Their Polymer Microstructures. Adv Sci. 2019 Jul;6(14):1900491.
166. Son HF, Cho IJ, Joo S, Seo H, Sagong HY, Choi SY, et al. Rational Protein Engineering of Thermo-Stable PETase from *Ideonella sakaiensis* for Highly Efficient PET Degradation. ACS Catal. 2019 Apr 5;9(4):3519–26.
167. Zhong-Johnson EZL, Voigt CA, Sinskey AJ. An absorbance method for analysis of enzymatic degradation kinetics of poly(ethylene terephthalate) films. Sci Rep. 2021 Jan 13;11(1):928.

168. Bell EL, Smithson R, Kilbride S, Foster J, Hardy FJ, Ramachandran S, et al. Directed evolution of an efficient and thermostable PET depolymerase. Nat Catal. 2022 Aug 11;5(8):673–81.
169. Cui Y, Chen Y, Liu X, Dong S, Tian Y, Qiao Y, et al. Computational Redesign of a PETase for Plastic Biodegradation under Ambient Condition by the GRAPE Strategy. ACS Catal. 2021 Feb 5;11(3):1340–50.
170. Lu H, Diaz DJ, Czarnecki NJ, Zhu C, Kim W, Shroff R, et al. Machine learning-aided engineering of hydrolases for PET depolymerization. Nature. 2022 Apr 28;604(7907):662–7.
171. Tournier V, Topham CM, Gilles A, David B, Folgoas C, Moya-Leclair E, et al. An engineered PET depolymerase to break down and recycle plastic bottles. Nature. 2020 Apr;580(7802):216–9.
172. Cui Y, Chen Y, Liu X, Dong S, Tian Y, Qiao Y, et al. Computational Redesign of a PETase for Plastic Biodegradation under Ambient Condition by the GRAPE Strategy. ACS Catal. 2021 Feb 5;11(3):1340–50.
173. Cui Y, Chen Y, Sun J, Zhu T, Pang H, Li C, et al. Computational redesign of a hydrolase for nearly complete PET depolymerization at industrially relevant high-solids loading. Nat Commun. 2024 Feb 15;15(1):1417.
174. Sulaiman S, You DJ, Kanaya E, Koga Y, Kanaya S. Crystal structure and thermodynamic and kinetic stability of metagenome-derived LC-cutinase. Biochemistry. 2014 Mar 25;53(11):1858–69.
175. Chen S, Tong X, Woodard RW, Du G, Wu J, Chen J. Identification and characterization of bacterial cutinase. J Biol Chem. 2008 Sep 19;283(38):25854–62.
176. Erickson E, Gado JE, Avilán L, Bratti F, Brizendine RK, Cox PA, et al. Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. Nat Commun. 2022 Dec 21;13(1):7850.
177. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan 4;44(D1):D7-19.
178. Chen IMA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. 2019 Jan 8;47(D1):D666–77.
179. Eiamthong B, Meesawat P, Wongsatit T, Jitdee J, Sangsri R, Patchsung M, et al. Discovery and Genetic Code Expansion of a Polyethylene Terephthalate (PET) Hydrolase from the Human Saliva Metagenome for the Degradation and Bio-Functionalization of PET. Angew Chem Int Ed. 2022 Sep 12;61(37):e202203061.
180. Kim HT, Kim JK, Cha HG, Kang MJ, Lee HS, Khang TU, et al. Biological Valorization of Poly(ethylene terephthalate) Monomers for Upcycling Waste PET. ACS Sustain Chem Eng. 2019 Dec 16;7(24):19396–406.
181. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010 Jan 30;31(2):455–61.
182. Liu B, He L, Wang L, Li T, Li C, Liu H, et al. Protein Crystallography and Site-Direct Mutagenesis Analysis of the Poly(ethylene terephthalate) Hydrolase PETase from *Ideonella sakaiensis*. ChemBioChem. 2018 Jul 16;19(14):1471–5.

183. Hantani Y, Imamura H, Yamamoto T, Senga A, Yamagami Y, Kato M, et al. Functional characterizations of polyethylene terephthalate-degrading cutinase-like enzyme Cut190 mutants using bis(2-hydroxyethyl) terephthalate as the model substrate. AIMS Biophys. 2018;5(4):290–302.
184. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8;49(D1):D1388–95.
185. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605–12.
186. Sasidharan S, Saudagar P. Prediction, validation, and analysis of protein structures: A beginner's guide. In: Advances in Protein Molecular and Structural Biology Methods [Internet]. Elsevier; 2022 [cited 2025 Jan 18]. p. 373–85. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780323902649000234
187. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: combining techniques to model RNA-small molecule complexes. RNA N Y N. 2009 Jun;15(6):1219–30.
188. Rudd MF. The Predicted Impact of Coding Single Nucleotide Polymorphisms Database. Cancer Epidemiol Biomarkers Prev. 2005 Nov 1;14(11):2598–604.
189. Li WH, Wu CI, Luo CC. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol. 1984;21(1):58–71.
190. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 1974 Sep 6;185(4154):862–4.
191. Hon J, Marusiak M, Martinek T, Kunka A, Zendulka J, Bednar D, et al. SoluProt: prediction of soluble protein expression in *Escherichia coli*. Xu J, editor. Bioinformatics. 2021 Apr 9;37(1):23–8.
192. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10(1):168.
193. Avise JC, Nicholson TH. Evolutionary pathways in nature: a phylogenetic approach. Cambridge: Cambridge University Press; 2006. 286 p.
194. Jai Ram Rideout, Greg Caporaso, Evan Bolyen, Daniel McDonald, Yoshiki Vázquez Baeza, Jorge Cañardo Alastuey, et al. scikit-bio/scikit-bio: scikit-bio 0.6.3 [Internet]. Zenodo; 2025 [cited 2025 Jan 18]. Available from: https://zenodo.org/doi/10.5281/zenodo.593387
195. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019 Jul 2;47(W1):W256–9.
196. Musil M, Konegger H, Hon J, Bednar D, Damborsky J. Computational Design of Stable and Soluble Biocatalysts. ACS Catal. 2019 Feb 1;9(2):1033–54.
197. Helen M. Berman MJG. Protein Structure Initiative - Targettrack 2000-2017 - All Data Files [Internet]. Zenodo; 2017 [cited 2025 Jan 18]. Available from: https://zenodo.org/record/821654
198. Ivanov DK, Bostelmann G, Lan-Leung B, Williams J, Partridge L, Escott-Price V, et al. A novel computational approach for predicting complex phenotypes in Drosophila (starvation-sensitive and sterile) by deriving their gene expression signatures from public data. PloS One. 2020;15(10):e0240824.

199. PDBe-KB consortium, Varadi M, Berrisford J, Deshpande M, Nair SS, Gutmanas A, et al. PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res. 2020 Jan 8;48(D1):D344–53.
200. Bordin N, Scholes H, Rauer C, Roca-Martínez J, Sillitoe I, Orengo C. Clustering protein functional families at large scale with hierarchical approaches. Protein Sci Publ Protein Soc. 2024 Sep;33(9):e5140.
201. Knott BC, Erickson E, Allen MD, Gado JE, Graham R, Kearns FL, et al. Characterization and engineering of a two-enzyme system for plastics depolymerization. Proc Natl Acad Sci. 2020 Oct 13;117(41):25476–85.
202. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002/07/24 ed. 2002 Jul 15;30(14):3059–66.
203. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004/07/21 ed. 2004 Oct;25(13):1605–12.
204. Meng X, Yang L, Liu H, Li Q, Xu G, Zhang Y, et al. Protein engineering of stable IsPETase for PET plastic degradation by Premuse. Int J Biol Macromol. 2021 Jun;180:667–76.
205. Samak NA, Jia Y, Sharshar MM, Mu T, Yang M, Peh S, et al. Recent advances in biocatalysts engineering for polyethylene terephthalate plastic waste green recycling. Environ Int. 2020 Dec;145:106144.
206. Han X, Liu W, Huang JW, Ma J, Zheng Y, Ko TP, et al. Structural insight into catalytic mechanism of PET hydrolase. Nat Commun. 2017 Dec 13;8(1):2106.
207. Charupanit K, Tipmanee V, Sutthibutpong T, Limsakul P. In Silico Identification of Potential Sites for a Plastic-Degrading Enzyme by a Reverse Screening through the Protein Sequence Space and Molecular Dynamics Simulations. Molecules. 2022 May 23;27(10):3353.
208. Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, Salisbury BA, et al. DNA variability of human genes. Mech Ageing Dev. 2003 Jan;124(1):17–25.
209. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928–33.
210. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, et al. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet. 2000 Mar;66(3):979–88.
211. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. An apportionment of human DNA diversity. Proc Natl Acad Sci U S A. 1997 Apr 29;94(9):4516–9.
212. Auton A, Abecasis GR, Steering committee, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.
213. Tang H. Confronting ethnicity-specific disease risk. Nat Genet. 2006 Jan;38(1):13–5.
214. Jorde LB, Wooding SP. Genetic variation, classification and 'race'. Nat Genet. 2004 Nov;36(S11):S28–33.
215. Lee YH, Hong CM, Kim DH, Lee TH, Lee J. Clinical Course of Asymptomatic and Mildly Symptomatic Patients with Coronavirus Disease Admitted to

Community Treatment Centers, South Korea. Emerg Infect Dis. 2020 Oct;26(10):2346–52.
216. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 2011 Jan 1;39(suppl_1):D576–82.
217. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Mar 12;579(7798):270–3.
218. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020 Apr;181(2):271-280.e8.
219. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020 May 14;581(7807):215–20.
220. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020 Apr;181(2):281-292.e6.
221. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents. 2020 Mar;55(3):105924.
222. Brian DA, Baric RS. Coronavirus Genome Structure and Replication. In: Enjuanes L, editor. Coronavirus Replication and Reverse Genetics [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005 [cited 2022 Feb 23]. p. 1–30. (Current Topics in Microbiology and Immunology; vol. 287). Available from: https://link.springer.com/10.1007/3-540-26765-4_1
223. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019 Mar;17(3):181–92.
224. Lau SKP, Woo PCY, Li KSM, Huang Y, Tsoi HW, Wong BHL, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci U S A. 2005 Sep 27;102(39):14040–5.
225. Annan A, Baldwin HJ, Corman VM, Klose SM, Owusu M, Nkrumah EE, et al. Human betacoronavirus 2c EMC/2012-related viruses in bats, Ghana and Europe. Emerg Infect Dis. 2013 Mar;19(3):456–9.
226. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020 Feb;395(10224):565–74.
227. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ. 2020 Jul 1;98(7):495–504.
228. Dolan PT, Whitfield ZJ, Andino R. Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution. Annu Rev Virol. 2018 Sep 29;5(1):69–92.
229. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, et al. Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. Mol Cell. 2020 Sep;79(5):710–27.
230. Bar-On YM, Flamholz A, Phillips R, Milo R. SARS-CoV-2 (COVID-19) by the numbers. eLife. 2020 Apr 2;9:e57309.

231. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020 Jan 1;9(1):313–9.
232. Santos I de A, Grosche VR, Bergamini FRG, Sabino-Silva R, Jardim ACG. Antivirals Against Coronaviruses: Candidate Drugs for SARS-CoV-2 Treatment? Front Microbiol. 2020 Aug 13;11:1818.
233. Samavati L, Uhal BD. ACE2, Much More Than Just a Receptor for SARS-COV-2. Front Cell Infect Microbiol. 2020 Jun 5;10:317.
234. Li F, Li W, Farzan M, Harrison SC. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. Science. 2005 Sep 16;309(5742):1864–8.
235. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020 May;368(6490):489–93.
236. Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature. 2003 Nov;426(6965):450–4.
237. Ray D, Le L, Andricioaei I. Distant residues modulate conformational opening in SARS-CoV-2 spike protein. Proc Natl Acad Sci. 2021 Oct 26;118(43):e2100943118.
238. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021 Jul;19(7):409–24.
239. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020 Mar 13;367(6483):1260–3.
240. Kirchdoerfer RN, Wang N, Pallesen J, Wrapp D, Turner HL, Cottrell CA, et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. Sci Rep. 2018 Dec;8(1):15701.
241. Sun G, Xue L, He Q, Zhao Y, Xu W, Wang Z. Structural insights into SARS-CoV-2 infection and therapeutics development. Stem Cell Res. 2021 Apr;52:102219.
242. Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol. 2016 Sep 29;3(1):237–61.
243. Ismail AM, Elfiky AA. SARS-CoV-2 spike behavior in situ: a Cryo-EM images for a better understanding of the COVID-19 pandemic. Signal Transduct Target Ther. 2020 Dec;5(1):252.
244. Roy S, Jaiswar A, Sarkar R. Dynamic Asymmetry Exposes 2019-nCoV Prefusion Spike. J Phys Chem Lett. 2020 Sep 3;11(17):7021–7.
245. Huo J, Zhao Y, Ren J, Zhou D, Duyvesteyn HME, Ginn HM, et al. Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. Cell Host Microbe. 2020 Sep;28(3):445-454.e6.
246. Walls AC, Tortorici MA, Snijder J, Xiong X, Bosch BJ, Rey FA, et al. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. Proc Natl Acad Sci. 2017 Oct 17;114(42):11157–62.
247. Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. Heise MT, editor. PLOS Pathog. 2018 Aug 13;14(8):e1007236.

248. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020 May 14;581(7807):221–4.
249. Salleh MZ, Derrick JP, Deris ZZ. Structural Evaluation of the Spike Glycoprotein Variants on SARS-CoV-2 Transmission and Immune Evasion. Int J Mol Sci. 2021 Jul 10;22(14):7425.
250. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife. 2020 Oct 28;9:e61312.
251. Wang X, Lu L, Jiang S. SARS-CoV-2 Omicron subvariant BA.2.86: limited potential for global spread. Signal Transduct Target Ther. 2023 Nov 30;8(1):439.
252. Wang TH, Shao HP, Zhao BQ, Zhai HL. Molecular Insights into the Variability in Infection and Immune Evasion Capabilities of SARS-CoV-2 Variants: A Sequence and Structural Investigation of the RBD Domain. J Chem Inf Model. 2024 Apr 22;64(8):3503–23.
253. Mansbach RA, Chakraborty S, Nguyen K, Montefiori DC, Korber B, Gnanakaran S. The SARS-CoV-2 Spike variant D614G favors an open conformational state. Sci Adv. 2021 Apr 16;7(16):eabf3671.
254. Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Cell. 2020 Oct;183(3):739-751.e8.
255. Campbell F, Archer B, Laurenson-Schafer H, Jinnai Y, Konings F, Batra N, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Eurosurveillance [Internet]. 2021 Jun 17 [cited 2022 Feb 24];26(24). Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.24.2100509
256. Han P, Li L, Liu S, Wang Q, Zhang D, Xu Z, et al. Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. Cell. 2022 Feb;185(4):630-640.e10.
257. Henderson R, Edwards RJ, Mansouri K, Janowska K, Stalls V, Gobeil SMC, et al. Controlling the SARS-CoV-2 spike glycoprotein conformation. Nat Struct Mol Biol. 2020 Oct;27(10):925–33.
258. Verkhivker GM, Di Paola L. Dynamic Network Modeling of Allosteric Interactions and Communication Pathways in the SARS-CoV-2 Spike Trimer Mutants: Differential Modulation of Conformational Landscapes and Signal Transmission via Cascades of Regulatory Switches. J Phys Chem B. 2021 Jan 28;125(3):850–73.
259. Ramanathan M, Ferguson ID, Miao W, Khavari PA. SARS-CoV-2 B.1.1.7 and B.1.351 spike variants bind human ACE2 with increased affinity. Lancet Infect Dis. 2021 Aug;21(8):1070.
260. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil. MedRxiv Prepr Serv Health Sci. 2021 Mar 3;2021.02.26.21252554.
261. Han P, Su C, Zhang Y, Bai C, Zheng A, Qiao C, et al. Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants. Nat Commun. 2021 Dec;12(1):6103.

262. Saito A, Irie T, Suzuki R, Maemura T, Nasser H, Uriu K, et al. Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. Nature. 2022 Feb 10;602(7896):300–6.
263. Huang K, Zhang Y, Hui X, Zhao Y, Gong W, Wang T, et al. Q493K and Q498H substitutions in Spike promote adaptation of SARS-CoV-2 in mice. EBioMedicine. 2021 May;67:103381.
264. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020 Jul 16;583(7816):459–68.
265. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar;579(7798):265–9.
266. Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect. 2020;9(1):221–36.
267. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol Clifton NJ. 2015;1282:1–23.
268. Xia H, Cao Z, Xie X, Zhang X, Chen JYC, Wang H, et al. Evasion of Type I Interferon by SARS-CoV-2. Cell Rep. 2020 Oct 6;33(1):108234.
269. Guo K, Barrett BS, Morrison JH, Mickens KL, Vladar EK, Hasenkrug KJ, et al. Interferon resistance of emerging SARS-CoV-2 variants. Proc Natl Acad Sci U S A. 2022 Aug 9;119(32):e2203760119.
270. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971 Sep;35(3):235–41.
271. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall Hoboken NJ. 2017 Jan;1(1):33–46.
272. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2022 Aug;43(8):1012–30.
273. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May 28;581(7809):434–43.
274. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet. 2017 Nov;25(11):1253–60.
275. Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, et al. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019 Dec 5;576(7785):106–11.
276. Jain A, Bhoyar RC, Pandhare K, Mishra A, Sharma D, Imran M, et al. IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. Nucleic Acids Res. 2020 Oct 23;gkaa923.
277. Tadaka S, Saigusa D, Motoike IN, Inoue J, Aoki Y, Shirota M, et al. jMorp: Japanese Multi Omics Reference Panel. Nucleic Acids Res. 2018 Jan 4;46(D1):D551–7.
278. Villaveces JM, Jiménez RC, Porras P, Del-Toro N, Duesbury M, Dumousseau M, et al. Merging and scoring molecular interactions utilising existing community

standards: tools, use-cases and a case study. Database J Biol Databases Curation. 2015;2015:bau131.
279. Roy AA, Dhawanjewar AS, Sharma P, Singh G, Madhusudhan MS. Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein-protein interactions. Nucleic Acids Res. 2019 Jul 2;47(W1):W331–7.
280. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. eLife. 2015 Jul 20;4:e07454.
281. Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. Bioinformatics. 2016 Dec 1;32(23):3676–8.
282. Van Zundert GCP, Bonvin AMJJ. DisVis: quantifying and visualizing accessible interaction space of distance-restrained biomolecular complexes. Bioinformatics. 2015 Oct 1;31(19):3222–4.
283. De Vries SJ, Van Dijk M, Bonvin AMJJ. The HADDOCK web server for data-driven biomolecular docking. Nat Protoc. 2010 May;5(5):883–97.
284. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc. 2003 Feb 19;125(7):1731–7.
285. Honorato RV, Koukos PI, Jiménez-García B, Tsaregorodtsev A, Verlato M, Giachetti A, et al. Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. Front Mol Biosci. 2021 Jul 28;8:729513.
286. De Vries SJ, Bonvin AMJJ. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. Fernandez-Fuentes N, editor. PLoS ONE. 2011 Mar 25;6(3):e17695.
287. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife. 2014 Sep 25;3:e03430.
288. Karaca E, Bonvin AMJJ. On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. Acta Crystallogr D Biol Crystallogr. 2013 May;69(Pt 5):683–94.
289. van Zundert GCP, Melquiond ASJ, Bonvin AMJJ. Integrative Modeling of Biomolecular Complexes: HADDOCKing with Cryo-Electron Microscopy Data. Struct Lond Engl 1993. 2015 May 5;23(5):949–60.
290. Van Dijk ADJ, Fushman D, Bonvin AMJJ. Various strategies of using residual dipolar couplings in NMR-driven protein docking: Application to Lys48-linked di-ubiquitin and validation against^{15} N-relaxation data. Proteins Struct Funct Bioinforma. 2005 Aug 15;60(3):367–81.
291. Van Dijk ADJ, Kaptein R, Boelens R, Bonvin AMJJ. Combining NMR Relaxation with Chemical Shift Perturbation Data to Drive Protein–protein Docking. J Biomol NMR. 2006 Apr;34(4):237–44.
292. Schmitz C, Bonvin AMJJ. Protein–protein HADDocking using exclusively pseudocontact shifts. J Biomol NMR. 2011 Jul;50(3):263–6.
293. van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016 Feb;428(4):720–5.

294. Tsai HHG, Tsai CJ, Ma B, Nussinov R. In silico protein design by combinatorial assembly of protein building blocks. Protein Sci Publ Protein Soc. 2004 Oct;13(10):2753–65.
295. Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SW, et al. Validation of molecular docking programs for virtual screening against dihydropteroate synthase. J Chem Inf Model. 2009 Feb;49(2):444–60.
296. Ramírez D, Caballero J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? Molecules. 2018 Apr 28;23(5):1038.
297. Wang J, Jain A, McDonald LR, Gambogi C, Lee AL, Dokholyan NV. Mapping allosteric communications within individual proteins. Nat Commun. 2020 Dec;11(1):3862.
298. Dokholyan NV. Controlling Allosteric Networks in Proteins. Chem Rev. 2016 Jun 8;116(11):6463–87.
299. Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface. J Mol Biol. 2014 Jul;426(14):2632–52.
300. Young JC, Hoogenraad NJ, Hartl FU. Molecular chaperones Hsp90 and Hsp70 deliver preproteins to the mitochondrial import receptor Tom70. Cell. 2003 Jan 10;112(1):41–50.
301. Li X, Straub J, Medeiros TC, Mehra C, den Brave F, Peker E, et al. Mitochondria shed their outer membrane in response to infection-induced stress. Science. 2022 Jan 14;375(6577):eabi4343.
302. Liu XY, Wei B, Shi HX, Shan YF, Wang C. Tom70 mediates activation of interferon regulatory factor 3 on mitochondria. Cell Res. 2010 Sep;20(9):994–1011.
303. Meier C, Aricescu AR, Assenberg R, Aplin RT, Gilbert RJC, Grimes JM, et al. The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. Struct Lond Engl 1993. 2006 Jul;14(7):1157–65.
304. Gordon DE, Hiatt J, Bouhaddou M, Rezelj VV, Ulferts S, Braberg H, et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. Science. 2020 Dec 4;370(6521):eabe9403.
305. Jiang H wei, Zhang H nan, Meng Q feng, Xie J, Li Y, Chen H, et al. SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. Cell Mol Immunol. 2020 Sep;17(9):998–1000.
306. Gao X, Zhu K, Qin B, Olieric V, Wang M, Cui S. Crystal structure of SARS-CoV-2 Orf9b in complex with human TOM70 suggests unusual virus-host interactions. Nat Commun. 2021 May 14;12(1):2843.
307. Ayinde KS, Pinheiro GMS, Ramos CHI. Binding of SARS-CoV-2 protein ORF9b to mitochondrial translocase TOM70 prevents its interaction with chaperone HSP90. Biochimie. 2022 Sep;200:99–106.
308. Swaim CD, Canadeo LA, Monte KJ, Khanna S, Lenschow DJ, Huibregtse JM. Modulation of Extracellular ISG15 Signaling by Pathogens and Viral Effector Proteins. Cell Rep. 2020 Jun;31(11):107772.
309. Liu G, Lee JH, Parker ZM, Acharya D, Chiang JJ, van Gent M, et al. ISG15-dependent activation of the sensor MDA5 is antagonized by the SARS-CoV-2 papain-like protease to evade host innate immunity. Nat Microbiol. 2021 Apr;6(4):467–78.

310. Armstrong LA, Lange SM, Dee Cesare V, Matthews SP, Nirujogi RS, Cole I, et al. Biochemical characterization of protease activity of Nsp3 from SARS-CoV-2 and its inhibition by nanobodies. PloS One. 2021;16(7):e0253364.
311. Wu XM, Zhang J, Li PW, Hu YW, Cao L, Ouyang S, et al. NOD1 Promotes Antiviral Signaling by Binding Viral RNA and Regulating the Interaction of MDA5 and MAVS. J Immunol Baltim Md 1950. 2020 Apr 15;204(8):2216–31.
312. Pichlmair A, Lassnig C, Eberle CA, Górna MW, Baumann CL, Burkard TR, et al. IFIT1 is an antiviral protein that recognizes 5′-triphosphate RNA. Nat Immunol. 2011 Jul;12(7):624–30.
313. Daffis S, Szretter KJ, Schriewer J, Li J, Youn S, Errett J, et al. 2'-O methylation of the viral mRNA cap evades host restriction by IFIT family members. Nature. 2010 Nov 18;468(7322):452–6.
314. Li Y, Li C, Xue P, Zhong B, Mao AP, Ran Y, et al. ISG56 is a negative-feedback regulator of virus-triggered signaling and cellular antiviral response. Proc Natl Acad Sci U S A. 2009 May 12;106(19):7945–50.
315. Berchtold S, Manncke B, Klenk J, Geisel J, Autenrieth IB, Bohn E. Forced IFIT-2 expression represses LPS induced TNF-alpha expression at posttranscriptional levels. BMC Immunol. 2008 Dec;9(1):75.
316. Stawowczyk M, Van Scoy S, Kumar KP, Reich NC. The interferon stimulated gene 54 promotes apoptosis. J Biol Chem. 2011 Mar 4;286(9):7257–66.
317. Reuschl AK, Thorne LG, Whelan MVX, Ragazzini R, Furnon W, Cowton VM, et al. Evolution of enhanced innate immune suppression by SARS-CoV-2 Omicron subvariants [Internet]. 2022 [cited 2025 Jan 18]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.07.12.499603
318. MacGowan SA, Barton MI, Kutuzov M, Dushek O, Van Der Merwe PA, Barton GJ. Missense variants in human ACE2 strongly affect binding to SARS-CoV-2 Spike providing a mechanism for ACE2 mediated genetic risk in Covid-19: A case study in affinity predictions of interface variants. Deane CM, editor. PLOS Comput Biol. 2022 Mar 2;18(3):e1009922.
319. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol. 2021 Mar;19(3):141–54.
320. Niemi MEK, Daly MJ, Ganna A. The human genetic epidemiology of COVID-19. Nat Rev Genet. 2022 Sep;23(9):533–46.
321. Lam SD, Bordin N, Waman VP, Scholes HM, Ashford P, Sen N, et al. SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. Sci Rep. 2020 Oct 5;10(1):16471.
322. The Severe Covid-19 GWAS Group. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020 Oct 15;383(16):1522–34.
323. Casanova JL, Su HC, Abel L, Aiuti A, Almuhsen S, Arias AA, et al. A Global Effort to Define the Human Genetics of Protective Immunity to SARS-CoV-2 Infection. Cell. 2020 Jun;181(6):1194–9.
324. The GenOMICC Investigators, The ISARIC4C Investigators, The COVID-19 Human Genetics Initiative, 23andMe Investigators, BRACOVID Investigators, Gen-COVID Investigators, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021 Mar 4;591(7848):92–8.
325. Asgari S, Pousaz LA. Human genetic variants identified that affect COVID susceptibility and severity. Nature. 2021 Dec 16;600(7889):390–1.

326. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020 Feb 6;578(7793):122–8.
327. Ciriello G, Magnani L. The many faces of cancer evolution. iScience. 2021 May 21;24(5):102403.
328. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non–Small-Cell Lung Cancer. N Engl J Med. 2017 Jun;376(22):2109–21.
329. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009 Apr;458(7239):719–24.
330. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546–58.
331. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. Trends Mol Med. 2023 Jul;29(7):554–66.
332. Salgia R, Abidoye O. ONCOGENES AND PROTO-ONCOGENES | Overview. In: Encyclopedia of Respiratory Medicine [Internet]. Elsevier; 2006 [cited 2025 Jan 18]. p. 236–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/B0123708796002787
333. Brown G. Oncogenes, Proto-Oncogenes, and Lineage Restriction of Cancer Stem Cells. Int J Mol Sci. 2021 Sep 7;22(18):9667.
334. Kugoh H, Ohira T, Oshimura M. Studies of Tumor Suppressor Genes via Chromosome Engineering. Cancers. 2015 Dec 30;8(1):4.
335. Bunz F. Tumor Suppressor Genes. In: Principles of Cancer Genetics [Internet]. Dordrecht: Springer Netherlands; 2016 [cited 2025 Jan 18]. p. 75–134. Available from: http://link.springer.com/10.1007/978-94-017-7484-0_3
336. Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. N Engl J Med. 2016 Aug 4;375(5):443–53.
337. Torgovnick A, Schumacher B. DNA repair mechanisms in cancer development and therapy. Front Genet. 2015;6:157.
338. Li L, Li X, Tang Y, Lao Z, Lei J, Wei G. Common cancer mutations R175H and R273H drive the p53 DNA-binding domain towards aggregation-prone conformations. Phys Chem Chem Phys. 2020;22(17):9225–32.
339. Dela Cruz CS, Tanoue LT, Matthay RA. Lung Cancer: Epidemiology, Etiology, and Prevention. Clin Chest Med. 2011 Dec;32(4):605–44.
340. Guarga L, Ameijide A, Marcos-Gragera R, Carulla M, Delgadillo J, Borràs JM, et al. Trends in lung cancer incidence by age, sex and histology from 2012 to 2025 in Catalonia (Spain). Sci Rep. 2021 Dec 2;11(1):23274.
341. DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB. The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation. Cell Metab. 2008 Jan;7(1):11–20.
342. Courtnay R, Ngo DC, Malik N, Ververis K, Tortorella SM, Karagiannis TC. Cancer metabolism and the Warburg effect: the role of HIF-1 and PI3K. Mol Biol Rep. 2015 Apr;42(4):841–51.
343. Taylor-Weiner A, Zack T, O'Donnell E, Guerriero JL, Bernard B, Reddy A, et al. Genomic evolution and chemoresistance in germ-cell tumours. Nature. 2016 Nov 30;540(7631):114–8.

344. Johnson C, Warmoes MO, Shen X, Locasale JW. Epigenetics and cancer metabolism. Cancer Lett. 2015 Jan 28;356(2 Pt A):309–14.
345. Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, et al. Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. Cell Rep. 2018 Apr 3;23(1):255-269.e4.
346. Sinkala M, Mulder N, Patrick Martin D. Metabolic gene alterations impact the clinical aggressiveness and drug responses of 32 human cancers. Commun Biol. 2019 Nov 14;2(1):414.
347. Patel S, Ahmed S. Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. J Pharm Biomed Anal. 2015 Mar 25;107:63–74.
348. Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab. 2016 Jan 12;23(1):27–47.
349. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. Nat Commun. 2018 Dec 14;9(1):5330.
350. López S, Lim EL, Horswell S, Haase K, Huebner A, Dietzen M, et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat Genet. 2020 Mar;52(3):283–93.
351. Frankell AM, Dietzen M, Al Bakir M, Lim EL, Karasaki T, Ward S, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature. 2023 Apr 20;616(7957):525–33.
352. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019 Jan 8;47(D1):D590–5.
353. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.
354. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W193-200.
355. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 2019 Jan 8;47(D1):D280–4.
356. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. 2017 Jan 23;9(1):4.
357. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet. 2016 Aug;48(8):827–37.
358. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016 Jul 1;76(13):3719–31.
359. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018 Apr 5;173(2):371-385.e18.
360. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941–7.

361. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308–11.
362. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan;42(Database issue):D980-985.
363. Ashford P, Pang CSM, Moya-García AA, Adeyelu T, Orengo CA. A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations. Sci Rep. 2019 Jan 22;9(1):263.
364. Patani H, Bunney TD, Thiyagarajan N, Norman RA, Ogg D, Breed J, et al. Landscape of activating cancer mutations in FGFR kinases and their differential responses to inhibitors in clinical use. Oncotarget. 2016 Apr 26;7(17):24252–68.
365. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J. 1948 Jul;27(3):379–423.
366. Hill MO. Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology. 1973 Mar;54(2):427–32.
367. Pearson K. X. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*. Lond Edinb Dublin Philos Mag J Sci. 1900 Jul;50(302):157–75.
368. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000 Jan;154(1):459–73.
369. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938–50.
370. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010 Oct;20(10):1313–26.
371. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nat Rev Genet. 2013 Sep;14(9):645–60.
372. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014 Nov;15(11):734–48.
373. Chao A, Henderson PA, Chiu C, Moyes F, Hu K, Dornelas M, et al. Measuring temporal change in alpha diversity: A framework integrating taxonomic, phylogenetic and functional diversity and the iNEXT.3D standardization. Methods Ecol Evol. 2021 Oct;12(10):1926–40.
374. Jelinic P, Mueller JJ, Olvera N, Dao F, Scott SN, Shah R, et al. Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. Nat Genet. 2014 May;46(5):424–6.
375. Romero OA, Setien F, John S, Gimenez-Xavier P, Gómez-López G, Pisano D, et al. The tumour suppressor and chromatin-remodelling factor BRG1 antagonizes Myc activity and promotes cell differentiation in human cancer. EMBO Mol Med. 2012 Jul;4(7):603–16.
376. Bultman SJ, Herschkowitz JI, Godfrey V, Gebuhr TC, Yaniv M, Perou CM, et al. Characterization of mammary tumors from Brg1 heterozygous mice. Oncogene. 2008 Jan 17;27(4):460–8.
377. Zhang L, Xiao W, Wu F, Peng R, Shi J, Li C, et al. SMARCA4-mutated lung adenocarcinoma, a distinctive non-small cell lung cancer with worse prognosis. J Clin Oncol. 2021 May 20;39(15_suppl):e20548–e20548.

378. Derynck R, Zhang Y, Feng XH. Smads: transcriptional activators of TGF-beta responses. Cell. 1998 Dec 11;95(6):737–40.
379. Massagué J, Seoane J, Wotton D. Smad transcription factors. Genes Dev. 2005 Dec 1;19(23):2783–810.
380. Levy L, Hill CS. Alterations in components of the TGF-beta superfamily signaling pathways in human cancer. Cytokine Growth Factor Rev. 2006;17(1–2):41–58.
381. Boone BA, Sabbaghian S, Zenati M, Marsh JW, Moser AJ, Zureikat AH, et al. Loss of SMAD4 staining in pre-operative cell blocks is associated with distant metastases following pancreaticoduodenectomy with venous resection for pancreatic cancer. J Surg Oncol. 2014 Aug;110(2):171–5.
382. Tan X, Tong L, Li L, Xu J, Xie S, Ji L, et al. Loss of Smad4 promotes aggressive lung cancer metastasis by de-repression of PAK3 via miRNA regulation. Nat Commun. 2021 Aug 11;12(1):4853.
383. Tang C, Mo X, Niu Q, Wahafu A, Yang X, Qui M, et al. Hypomorph mutation-directed small-molecule protein-protein interaction inducers to restore mutant SMAD4-suppressed TGF-β signaling. Cell Chem Biol. 2021 May;28(5):636-647.e5.
384. Lanauze CB, Sehgal P, Hayer K, Torres-Diz M, Pippin JA, Grant SFA, et al. Colorectal Cancer-Associated Smad4 R361 Hotspot Mutations Boost Wnt/β-Catenin Signaling through Enhanced Smad4-LEF1 Binding. Mol Cancer Res MCR. 2021 May;19(5):823–33.
385. Albig W, Doenecke D. The human histone gene cluster at the D6S105 locus. Hum Genet. 1997 Dec 11;101(3):284–94.
386. Marzluff WF, Gongidi P, Woods KR, Jin J, Maltais LJ. The human and mouse replication-dependent histone genes. Genomics. 2002 Nov;80(5):487–98.
387. Li SC, Goto NK, Williams KA, Deber CM. Alpha-helical, but not beta-sheet, propensity of proline is determined by peptide environment. Proc Natl Acad Sci U S A. 1996 Jun 25;93(13):6676–81.
388. Leipe DD, Koonin EV, Aravind L. STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. J Mol Biol. 2004 Oct 8;343(1):1–28.
389. Denu JM, Dixon JE. Protein tyrosine phosphatases: mechanisms of catalysis and regulation. Curr Opin Chem Biol. 1998 Oct;2(5):633–41.
390. Paul S, Lombroso PJ. Receptor and nonreceptor protein tyrosine phosphatases in the nervous system. Cell Mol Life Sci CMLS. 2003 Nov;60(11):2465–82.
391. Flanagan CA, Schnieders EA, Emerick AW, Kunisawa R, Admon A, Thorner J. Phosphatidylinositol 4-kinase: gene structure and requirement for yeast cell viability. Science. 1993 Nov 26;262(5138):1444–8.
392. Troche JR, Mayne ST, Freedman ND, Shebl FM, Abnet CC. The Association Between Alcohol Consumption and Lung Carcinoma by Histological Subtype. Am J Epidemiol. 2016 Jan 15;183(2):110–21.
393. Wang GX, Zhao XY, Lin JD. The brown fat secretome: metabolic functions beyond thermogenesis. Trends Endocrinol Metab TEM. 2015 May;26(5):231–7.
394. Pace L, Nicolai E, Basso L, Garbino N, Soricelli A, Salvatore M. Brown Adipose Tissue in Breast Cancer Evaluated by [18F] FDG-PET/CT. Mol Imaging Biol. 2020 Aug;22(4):1111–5.

395. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 Jun 13;630(8016):493–500.