Multicomponent Complexes, Structural Proteomes, Drug Discovery For Cancer Gene Census And SARS CoV-2

Ali Faham Alsulami

This thesis is submitted for the degree of Doctor of Philosophy

Department of Biochemistry

St Edmund's College

University of Cambridge

January 2022

Abstract

Nowadays, with the next-generation sequencing technique, there is abundant sequence data for humans, bacteria, and viruses compared to structural data that provides a deep understanding of function. Since experimental structural biology is expensive and time-consuming, a protein modelling algorithm such as a MODELLER can reduce the ample space between the sequence and structural annotation in less time. Thus, solving protein structures with reliable or high accuracy quality assessment can be achieved computationally.

This is a step towards an era of personalised medicine since most drug optimisation, such as selectivity and potency, is structurally based. Proteins biologically can occur as protomers or assemble as higher order states as homo- or hetero-oligomers. Furthermore, protomers can interact with other assemblies forming multi-component-complex systems. One of the challenges in computational protein modelling is to predict a particular biological state of a given protein, also modelling that state in a high order assembly that mimics the actual biological assembly. There are two main drivers for computational structural modelling: first, to reduce the colossal sequence-structure gap, and second, to understand the impact of mutations in human cancer and new variants from viruses on the protein structure. The Catalogue of Somatic Mutations in Cancer (COSMIC) curates vast amounts of human mutation data. There are 723 genes in the Cancer Gene Census; these genes are experimentally validated as drivers of cancer progression and proliferation. Unfortunately, there are only 87 genes with experimentally solved structures of gene products with more than 90% structural coverage, whereas the protein structures related to other genes are still not solved or partially solved. A comprehensive state of the art computational

structure modelling effort has been carried out to build these genes with high order assembly, i.e. homodimer, heterodimer including ligand, DNA, RNA, and intrinsically disordered regions connecting domains. In addition, predictions of the impacts of reported mutations in the COSMIC database using statistical and machine learning algorithms such as SDM and mCSM can be used to hypothesise new driver mutations with structural impact. All these data are presented in a user-friendly interface (https://cancer-3d.com/) where users can retrieve and build hypotheses. Applying the same modelling approaches to another acute infectious disease such as severe acute respiratory syndrome coronavirus 2 (SARS CoV-2) can be beneficial to our understanding of the virus proteome and selecting a new drug target. The SARS CoV-2 sequence genome was released in early 2020, and full protomer and oligomeric structures were built where there were no experimental structures. Pocket detection, mutational analysis, protein-ligand and protein-protein docking were computationally performed to gain more insights into putative SARS CoV-2 drug targets. All this information is presented in a new user-friendly interface (https://sars3d.com/) that can be accessed freely to build hypotheses and download the data. Experimental validation of the impacts of mutations and validation of drug discovery targets is essential to assess the computational approaches that have been carried out for human cancer gene census and SARS CoV-2 and described in this thesis. Therefore, GTPase NRAS frequently reported mutants such as Q61K/L, G12D, and G13D were studied experimentally to understand the impacts of these mutations on protein structural conformation and function. In addition, hypotheses are presented concerning newly identified allosteric pockets that could be used to disrupt continuously active NRAS. The SARS-CoV-2 Non-structural protein 13

(nsp13) was selected as a drug discovery target to develop a new putative lead compound using fragment-based approaches.

Acknowledgement

I would like to thank the King Abdullah scholarship program for generous funding for my education journey in the UK; Undergrad, Master, and PhD. Without this generous funding reaching this stage will be difficult.

I would like to thank my supervisor Prof Sir Tom Blundell, for giving me the opportunity to work in his lab. Thank you, Tom, for all the support, enthusiasm, and patience throughout my PhD journey. In specific, I would like to thank Tom for the weekly meeting we used to have in the last four years. I have learned so much from these meetings. In particular, I have learned that I should always seek knowledge to become multi-disciplinary and think comprehensively about scientific projects. Also, I have learned from him that gender balance and multinational colleagues are very important in science. I would say this PhD degree would not be possible without Tom's support.

I would like to thank all of Prof Sir Tom Blundell previous and current group; Dr Pedro Torres, Dr Sundeep Chaitanya Vedithi, Dr Marcin Skwark, Dr Bridget Bannerman, Dr Sheikh Mohammed Arif, Dr Amanda Chaplin, Dr Sherine Thomas, Christopher A Beaudoin, Arian Jamasb, and Liviu Copoiu. In particular. I would like to thank Pedro for all the support and the skills I have gained from him. Thank you, Pedro, for all the fun and the joke we had together. Also, I would like to thank Sundeep for all the computational support in the past four years. Finally, I would like to thank Arif for all the help and support on the experimental side. Thank you, Arif, for making everything look easy.

I would like to thank all my collaborators from the University of Cambridge and other Universities. In addition, I would like to thank Dr Ismail Moghul from the university college London (UCL) for his support. This thesis will not be possible without this

collaboration. Also, I would like to thank prof Suzanne Turner lab for collaboration and publication.

Finally, I would like to thank my mother and father for all the support, love, and care I receive. I would like to thank all my brothers and sisters for supporting me and making this life so beautiful. I would like to thank my wife for helping me and making the PhD journey so joyful. Finally, I would like to thank my friends Dr Mohammed Al shahrani, Saeed Aljabri and Ahmed Khan for the gathering we had every month in the last four years.

Declaration

- This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.
- It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.
- It does not exceed the prescribed word limit for the Biology Degree Committee. For more information on the word limits for the respective Degree Committees see Word Limits and Requirements of your Degree Committee

Ali Faham Alsulami

Department of Biochemistry

St Edmund's College

University of Cambridge

Table of Contents

Abstract ***2***

Acknowledgement **5**

Preface ***11***

Experimental and computational development in the drug discovery field overall process. **11**

Computational approaches. **17**

Comparative protein modelling 17

Thesis outline. **21**

References **23**

Chapter I: Cosmic Cancer Gene Census 3D Databases ***26***

Introduction **27**

Cancer background 27

Mutations in Cancer 31

Methods **40**

Modelling gene products in CGC. 40

Mutation data 48

Website development 49

Results **51**

Website analysis 51

Data statistics 57

COSMIC CGC 3D Modelling examples 62

Predicting the impact of Mutations 69

Comparing COSMIC CGC 3D models to Alphafold models. 78

Discussion **81**

Conclusion **83**

References **84**

Chapter II: SARS-CoV-2 3D Proteome Database ***92***

Introduction **93**

SARS CoV-2 background. 93

SARS CoV-2 genome 96

SARS CoV-2 mechanism 97

Spike protein and pathogenesis 101

SARS CoV-2 drug development strategy and technology. 104

Valuable resource developed for SARS CoV-2. 106

Methods **108**

Proteome modelling 108

Virtual screening docking 112

Protein-protein-docking 114

Pocket predictions 114

Website development 115

Results **116**

Website analysis 116

Data statistics 121

SARS CoV-2 3D Modelling protome 123

Mutational analysis example 136

Molecular docking example 139

Protein-protein docking example 140

Discussion **142**

Conclusion **144**

Reference **145**

Chapter III: Studying the impact of mutations on NRAS. Drug discovery target evaluation of SARS CoV-2 nsp13. 152

GTPase NRas (NRAS) 153

RAS Introduction 153

Materials and methods 162

Protein production 162

Protein expression and purification 165

Results 168

NRAS structure annotation 168

Synthesis and purification of NRAS Q61K/L mutant. 173

Crystallisation condition 175

SARS CoV-2 Non-structural protein 13 (NSP-13) 176

Nsp-13 introduction 176

Materials and methods 178

Protein expression and purification 178

Results 180

Preliminary expression and purification tests. 180

Synthesis and purification of nsp13. 180

Structure analysis of nsp13 181

Fragment growing and linking 182

Discussion 186

Conclusion 188

References 189

Supplementary section 193

Preface

Experimental and computational development in the drug discovery field overall process.

The rapid development in science, including genome sequencing, combinatorial chemistry, structural biology, cryogenic electron microscopy (cryo-EM), and informatics, significantly accelerate drug discovery development (Renaud *et al.*, 2018). The new genome sequencing technology allows rapid sequencing of humans, bacteria and viruses to identify potential drug targets and detect new lineages and variants (Suzuki, 2020). Combinatorial chemistry extensively helps speed up the drug discovery process by creating a library of diverse molecules that can be tested against known targets using high throughput screening (HTS) (Alon *et al.*, 2021). New techniques have been developed recently, such as artificial intelligence (AI) AlphaFold, which predicts protein structure from the sequence with high accuracy (Jumper *et al.*, 2021). In addition, the revolution of cryo-EM, which predicts protein structure in an aqueous solution without the need for crystals, has helped the field dramatically to work with drug targets that were challenging in the past.

X-ray crystallography is still a powerful technique for determining protein structure with no limit in size. That is very important since small molecules such as fragments are still not easily seen by cryo-EM. Knowledge of the 3D structure, such as dynamics topologies and topographies, is very essential in the structural based drug design since the 3D structure is the starting point. This undoubtedly will help the medicinal chemist grow or link small fragments with more interactions with the target protein to gain more potency and selectivity (García-Nafría and Tate, 2020). An early

example of using a structure-based approach is the production of amprenavir and nelfinavir to combat the human immunodeficiency virus (HIV) (Surleraux *et al.*, 2005)(King *et al.*, 2012)

Comparative modelling or protein structure prediction allows solving the 3D structure from both close and distant homologues (Šali and Blundell, 1993). If the distant homologue has a proven drug or potent compound already provided, similar chemistry can be exploited as a starting point. High throughput screening (HTS) is still the leading approach for discovering new molecule hits in the industry. It relies on radioactivity or fluorescence-based assay, which is typically performed in 384-well plates (Tsuganezawa *et al.*, 2013). However, biophysical techniques such as surface plasmon resonance (SPR), isothermal titration (ITC) and nuclear magnetic resonance (NMR) can be used to obtain functional binding affinity of low molecular weight and tractable fragments that HTS campaigns will miss (Renaud *et al.*, 2016).

Sequence analysis of the globular folded domain of the target protein is essential before performing protein crystallisation. Pure 10-20 mg protein of the folded region tends to form crystals. Usually, the linker region within the globular domain has a low-complexity sequence and tends to be flexible, interfering with the crystallisation process. Therefore, removing it will help crystal formation. Multiple tools developed that detect these flexible or intrinsically disordered regions, such as DISOPRED and IUPred2A, can be used prior (Jones and Cozzetto, 2015). In addition, homology modelling can help identify surface lysine and glutamates that can be mutated if the initial crystallisation is unsuccessful.

The development of new machines and technologies helps many labs achieve parallel protein expression, purification, and crystallisation in a high-throughput manner. Our department is lucky to have an automated crystallisation sampling

method available with different commercial crystallisation plates with varieties of precipitating reagent, PH, and buffer compactions. The mosquito crystallisation robot reduces the quantity of protein required in sitting drop vapour diffusion experiments. In addition, video imaging machines allow monitoring of protein crystallisation processes.

The *Escherichia coli* expression system is the most used because it is cheap and quick (Rosano and Ceccarelli, 2014). However, optimisation can be required to avoid the production of insoluble inclusion bodies. Examples include small ubiquitin-like modifiers (SUMO) or green fluorescent protein (GFP) at the N-terminal. The expression temperature and amount of isopropyl β-D-1-thiogalactopyranoside (IPTG). Including tags such as His helps a lot in protein purification.

Protein structure determination using X-ray, cryo-EM, and NMR has become very important to understand drug targets by many companies and academia. These include Astex Technology Ltd (Cambridge, UK) and TRIAD Therapeutics (San Diego, USA). A high-resolution structural model requires designing a new compound that interacts with the protein structure. The accuracy of the high-resolution structure depends on refinement and structural restraint. However, proteins are dynamic molecules that can accommodate small conformational changes. The chemist should consider this when structural-based drugs are designed for specific interaction of active sites or small allosteric pockets.

Once the three-dimensional protein structure is identified, it is essential to identify the active site, allosteric pockets and key binding interactions (Radoux *et al.*, 2016). Many proteins undergo conformational change on ligand binding, and it is essential to identify the apo state to observe the change due to ligand binding. Protein-ligand interactions are experimentally done by soaking the ligand into a protein crystal of

interest. The CCP4 package was developed for macromolecular X-ray crystallography and used to automate the model fitting and structure minimisation into a newly generated electron density map. After the protein-ligand structure is solved, key ligand interactions can be optimised to reduce the entropic and enthalpic interactions. Increasing the ligand potency can often be achieved by adding a hydrophobic group. However, this should not be at the cost of essential ligand solubility. Lipinski summarises five rules that should guide the development of drug-like molecules to overcome the bioavailability issue that most drugs face in early development. (Lipinski *et al.*, 2012)

Computational drug discovery is as vital as experimental drug discovery. In the absence of an experimental drug-target structure, a protein with sequence similarity of more than 30-40% can be used to predict the overall protein structure topology. A more recent method, such as AlphaFold consists of three essential building blocks. The first is the pre-processing stage, where the target sequences are queried against the genetic database to look for similar evolutionary sequences, generating multiple sequence alignment (MSA). In addition, AlphaFold searches for a structural template with a similar sequence. Since template structure can have low sequence similarity or there is no template in some cases, the network is forced to learn from the MSA instead of relying on the template. All these data are integrated into pair representation and MSA representation. The second stage is the refinement of the evolutionary MSA and the special pair representation by the Evoformer, which is a (48 blocks) deep transformer-like network. The third stage is the structure module (8 blocks) that takes the abstract output from the Evoformer into three-dimensional

coordinates of the protein structure. Therefore, solving protein structure can be done computationally with a high-quality assessment score.

Nowadays, the predicted protein model can be used to generate hypotheses and find binding sites etc. Multiple protein pocket detection software has been developed, such as Fpocket, which identifies minor binding sites (Le Guilloux, Schmidtke and Tuffery, 2009). Fragment hotspot maps can be used to identify binding sites (Radoux *et al.*, 2016). The considerable cost of high-throughput screening and binding assays shifted the attention to virtual screening approaches, using either an experimentally solved structure or predicted model. Protein-ligand docking is widely used in academia and industry to study ligand-protein interactions. It predicts the relative binding affinity free energy and conformation of small drug-like compounds. User-friendly tools such as Glide-Schrödinger and MolSoft ICM have been developed to predict the relative ligand poses in the selected protein pocket (Meng, X. Y., Zhang, H. X., Mezei, M., Cui, 2011)(Halgren *et al.*, 2004). These methods are fast, and a large number of ligands ~2 million can be docked and ranked.

Another computational low-cost in silico technique predicts the impact of mutations on protein sequence and structure (Pires, Ascher and Blundell, 2014; Pandurangan *et al.*, 2017). The major challenge in this area is to differentiate driver from passenger mutations. Clearly, the most accurate way is to define the impact of mutations experimentally. However, if we have 1 million mutations, it becomes very challenging. In silico predictions help to reduce the vast number of reported mutations into a few putative mutations to be tested and validated experimentally. Many tools have been developed, including sequence-based and structure-based, using machine learning and statistical approaches, taking advantage of an experimental database that reported different free energy between wild-type and

mutant protein such as ProTherm (Nikam *et al.*, 2021) and Platinum DB (Pires, Blundell and Ascher, 2015).

Proteins exist in different homo-oligomeric states and can interact with other proteins forming a high order hetero-oligomeric assembly. In addition, many human proteins have intrinsically disordered polypeptide regions between domains (Piovesan *et al.*, 2021). Therefore, it is very important to address all these challenges computationally. Also, it is essential to predict protein pockets and compounds that bind them through virtual screening. Furthermore, predicting the impact of mutations to identify driver mutants is crucial for developing new compounds that interact with the mutant form. Therefore, the author has been involved in constructing two user-friendly computer databases: COSMIC cancer gene census 3D (https://cancer-3d.com/) for human proteins and SARS CoV-2 3D (https://sars3d.com/) for coronavirus. These two targets were selected to be tested and validated experimentally: Human NRAS from the COSMIC cancer gene census 3D and nsp13 from SARS CoV-2 3D database.

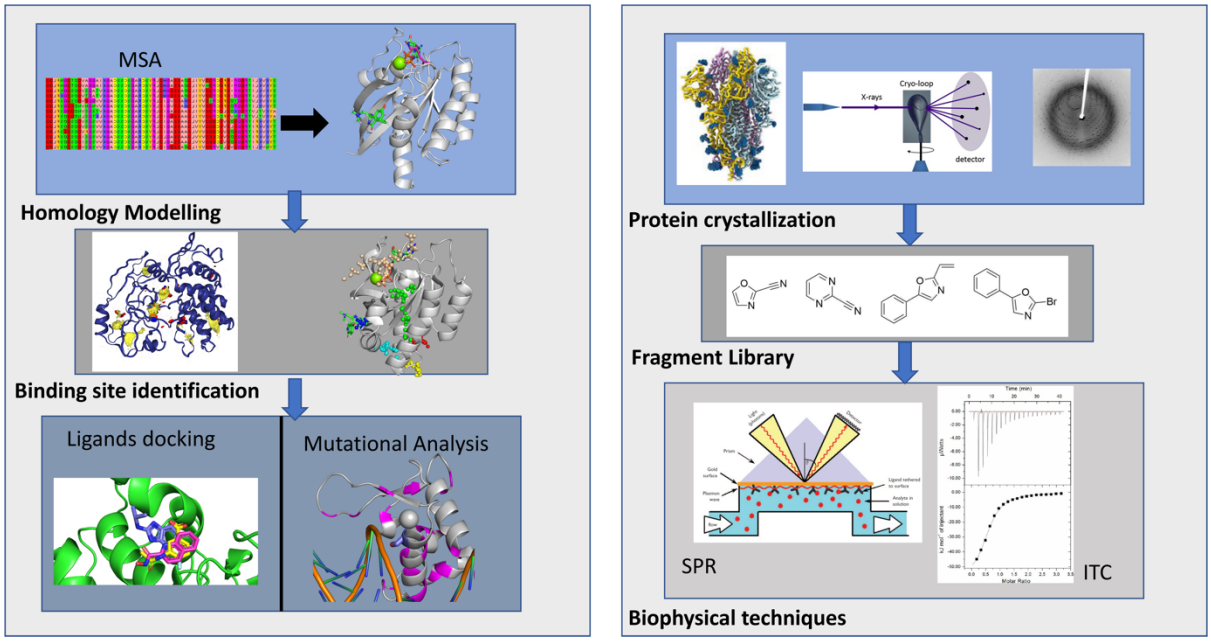


Figure 1. The overall drug discovery process in the Blundell lab. Usually, the computational process starts prior to lab work shown on the left side. Then experimental techniques carried out are shown on the right side.

Computational approaches.

Comparative protein modelling

Homology modelling, also known as comparative modelling, is a methodology to predict protein structure from its sequence to solve unknown three-dimensional protein structures from very close and distant homologues. Structures that have a similar sequence adopt the same structural fold. The increase in the 3D structures in the Protein Data Bank (PDB) has confirmed this rule (Berman *et al.*, 2000).

The number of 3D structures in the PDB has been increasing exponentially.

- 161144 structures of X-ray.
- 13545 structures of NMR.
- 9210 structures of electron microscopy.

However, more than 80 million protein sequences are deposited into the UniProtKB database, creating a huge gap between sequence and structure space (Schneider and Poux, 2012).

Many deposited sequences have solved homologue 3D structures, which can be selected for comparative modelling.

Comparative modelling consists of four main methods.

- Template selection.
- Templates alignment.
- Target structure
- Model evaluation

Template identification is an essential step in comparative modelling. Generally, a template sequence with > 30% identity to the target sequence will produce a good model structure. Multiple algorithms have been developed to find a suitable template:

- BLAST
- PSI-BLAST
- Hidden Markov model (HMMs)
- Threading methods
- FUGUE

Problems in template identification occur when the templates fall into the twilight zone (below 30% identity). BLAST is a local alignment method, treating the conserved region and variable region with the same weight (Altschul *et al.*, 1990). Therefore, it can find a template from a close homologue but not a distant homologue. PSI-BLAST is a much-improved method for identifying distant homologues. It is a profile-based method that can capture evolutionary or

functionally related sequences in multiple sequence alignment using position-specific scoring matrices (PSSMs) (Ramsay *et al.*, 2000). The hidden Markov model (HMM) has enhanced features by capturing the insertions and deletions in a multiple sequence alignment (Söding, 2005). The threading method finds templates by structural similarity. It fits the target sequence to the backbone of the selected template and evaluates how well it fits (Bowie, Lüthy and Eisenberg, 1991). This method is perfect for targets with high homologue similarity but performs worse than profile-based methods such as PSI-BLAST. So far, the FUGUE program is the best algorithm for identifying templates. It combined sequence and structural information and improved the environment-specific table by considering the gap penalty and structural features such as hydrogen bonding and solvent accessibility main-chain conformation. i.e. FUGUE finds the probability of a particular amino acid being substituted by another amino acid during evolution in a specific environment.(Shi, Blundell and Mizuguchi, 2001)

Templates are selected based on coverage identity, resolution, and the experimental technique since X-ray usually has a higher resolution than cryo-EM and NMR. Furthermore, it has been observed that selecting one template for a small globular domain is better than including multiple templates, whereas multiple templates modelling enhances the modelled quality for multi-domain proteins. Therefore, this is added to the criteria of template selection.

After the templates are identified, the target sequence can be modelled using *MODELLER*, the most used software in comparative modelling. *MODELLER* works by transferring the coordinate of the template that aligns with the target sequence to construct the backbone of the model. If the amino acid is similar, the side chain

coordinate will also be transferred, whereas the backbone will only be transferred if the side chain is different. After the model is generated, a quality assessment performs to quantify the modelled target structure.

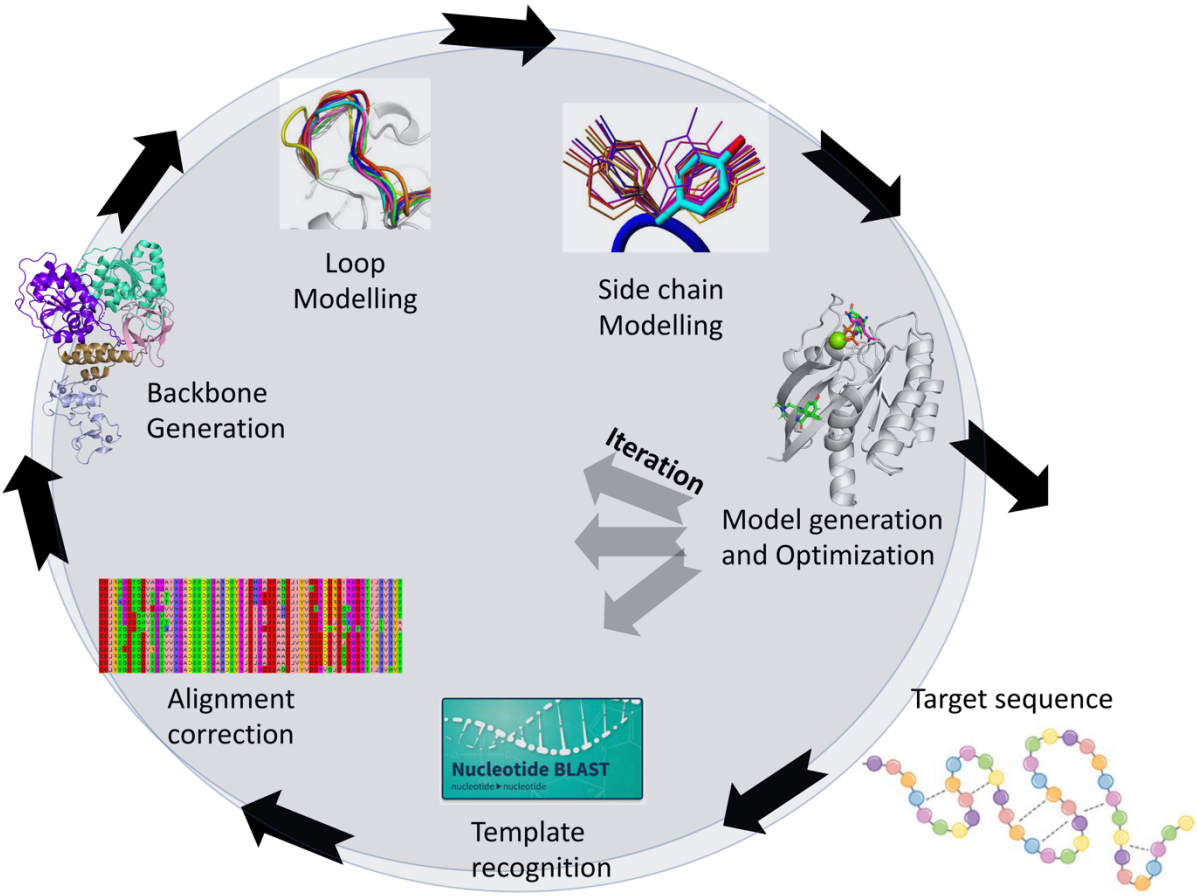


Figure 2. overall comparative modelling process. Starting from the target sequence and ending by final model generation. The iteration of these processes happens if the final generated model is not solved with a feasible quality assessment.

Thesis outline.

The thesis is divided into this introductory chapter followed by three large chapters, the first two of which are about computational structural biology databases from a drug discovery perspective, whereas the third is about experimental structural biology targets selected from the developed databases.

The first significant chapter concerns the **Cosmic Cancer Gene Census 3D Databases**. This computationally analyses the impacts of mutations in the cancer gene census. These genes are considered to be the cancer hallmarks, i.e. drive cancer progression. Many mutations have been curated for these genes. Structural appreciation of where mutations are in the protein is essential. However, 87 genes only have structures solved experimentally with structural coverage between 90-100%. The rest of the cancer gene census genes have very low structural coverage. The Cosmic Cancer Gene Census 3D Database includes the rest of the genes modelled in high order assemblies: homo-oligomer or hetero-oligomers, many with ligand DNA, RNA and co-factor, membrane, and the intrinsically disordered region between domains. SDM and mCSM were used to predict the impact of mutations using modelled oligomeric structures. These are presented in (cancer-3d.com) with other resources such as disorder prediction UniProt annotation to inform more about the queried target.

The second large chapter concerning the **SARS-CoV-2 3D Proteome Database** computationally describes the severe acute respiratory syndrome coronaviruses 2 (SARS CoV-2) proteome. The database includes other bioinformatic works and comprehensively brings all data together. Initially, after the SARS CoV-2 genome

sequence was released in March 2020, all the oligomeric models were built computationally. While other data were collected, new experimental structures started to be released into the Protein Data Bank. Comparing experimentally solved structures to the earlier constructed SARS CoV-2 built model structures gave the confidence to use the built modelled structures for other analyses such as protein-ligand docking, protein-protein docking and mutational analysis. All the data can be viewed on a fast, easily accessible, user-friendly website (sars3d.com)

The third chapter is entitled **Studying the impact of mutations on NRAS. And drug discovery target evaluation of SARS CoV-2 Nsp13** experimentally describes the impact of NRAS Q61K/L mutations on protein structure and how these frequently reported mutations could lead to continuous NRAS activation that in turn leads to activation of multiple signalling pathways. The goal is to identify the unique pocket form in the NRAS mutant not detected in the wild type of NRAS structure. Identifying these pockets is essential to design a new compound using a fragment-based approach. Another selected very promising drug target is nsp13 from the SARS CoV-2 database. nsp13 is conserved throughout the coronavirus family and is essential for forming the replication-transcription complex RTC inside the cell, which is in turn critical to the virus to attack other cells. In addition, nsp13 has multiple suspicious pockets for large and small compounds, making it more attractive as a drug discovery target.

References

Alon, A. *et al.* (2021) ‘Structures of the σ2 receptor enable docking for bioactive ligand discovery’, *Nature*, 600(December), pp. 759–764.

Altschul, S. F. *et al.* (1990) ‘Basic local alignment search tool’, *Journal of Molecular Biology*, 215(3), pp. 403–410.

Berman, H. M. *et al.* (2000) ‘The Protein Data Bank’, *Nucleic Acids Research*, 28(1), pp. 235–242.

Bowie, J. U., Lüthy, R. and Eisenberg, D. (1991) ‘A method to identify protein sequences that fold into a known three-dimensional stucture’, *Science*, 253(5016), pp. 164–170.

García-Nafría, J. and Tate, C. G. (2020) ‘Cryo-electron microscopy: Moving beyond X-ray crystal structures for drug receptors and drug development’, *Annual Review of Pharmacology and Toxicology*, 60, pp. 51–71.

Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) ‘Fpocket: An open source platform for ligand pocket detection’, *BMC Bioinformatics*, 10, pp. 1–11.

Halgren, T. A. *et al.* (2004) ‘Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening’, *Journal of Medicinal Chemistry*, 47(7), pp. 1750–1759.

Jones, D. T. and Cozzetto, D. (2015) ‘DISOPRED3: Precise disordered region predictions with annotated protein-binding activity’, *Bioinformatics*, 31(6), pp. 857–863.

Jumper, J. *et al.* (2021) ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, 596(7873), pp. 583–589.

King, N. M. *et al.* (2012) ‘Extreme entropy-enthalpy compensation in a drug-resistant variant of HIV-1 protease’, *ACS Chemical Biology*, 7(9), pp. 1536–1546.

Lipinski, C. A. *et al.* (2012) ‘Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings’, *Advanced Drug Delivery Reviews*, 64(SUPPL.), pp. 4–17.

Meng, X. Y., Zhang, H. X., Mezei, M., Cui, M. (2011) ‘Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Current Computer-Aided Drug Design.’, *Curr. Comput. Aid. Dru. Des.*, 7(2), pp. 146–157.

Nikam, R. *et al.* (2021) ‘ProThermDB: Thermodynamic database for proteins and mutants revisited after 15 years’, *Nucleic Acids Research*, 49(D1), pp. D420–D424.

Pandurangan, A. P. *et al.* (2017) ‘SDM: A server for predicting effects of mutations on protein stability’, *Nucleic Acids Research*, 45(W1), pp. W229–W235.

Piovesan, D. *et al.* (2021) ‘MobiDB: Intrinsically disordered proteins in 2021’, *Nucleic Acids Research*, 49(D1), pp. D361–D367.

Pires, D. E. V., Ascher, D. B. and Blundell, T. L. (2014) ‘MCSM: Predicting the effects of mutations in proteins using graph-based signatures’, *Bioinformatics*, 30(3), pp. 335–342.

Pires, D. E. V., Blundell, T. L. and Ascher, D. B. (2015) ‘Platinum: A database of experimentally measured effects of mutations on structurally defined protein-ligand complexes’, *Nucleic Acids Research*, 43(D1), pp. D387–D391.

Radoux, C. J. *et al.* (2016) ‘Identifying Interactions that Determine Fragment Binding at Protein Hotspots’, *Journal of Medicinal Chemistry*, 59(9), pp. 4314–4325.

Ramsay, L. *et al.* (2000) ‘A simple sequence repeat-based linkage map of Barley’, *Genetics*, 156(4), pp. 1997–2005.

Renaud, J. P. *et al.* (2016) ‘Biophysics in drug discovery: Impact, challenges and opportunities’, *Nature Reviews Drug Discovery*, 15(10), pp. 679–698.

Renaud, J. P. *et al.* (2018) ‘Cryo-EM in drug discovery: Achievements, limitations and prospects’, *Nature Reviews Drug Discovery*, 17(7), pp. 471–492.

Rosano, G. L. and Ceccarelli, E. A. (2014) ‘Recombinant protein expression in Escherichia coli: Advances and challenges’, *Frontiers in Microbiology*, 5(APR), pp. 1–17.

Šali, A. and Blundell, T. L. (1993) ‘Comparative Protein Modelling by Satisfaction of Spatial Restraints’, *Journal of Molecular Biology*, 234, pp. 779–815.

Schneider, M. and Poux, S. (2012) ‘UniProtKB amid the turmoil of plant proteomics research’, *Frontiers in Plant Science*, 3(DEC), pp. 1–7.

Shi, J., Blundell, T. L. and Mizuguchi, K. (2001) ‘FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties’, *Journal of Molecular Biology*, 310(1), pp. 243–257.

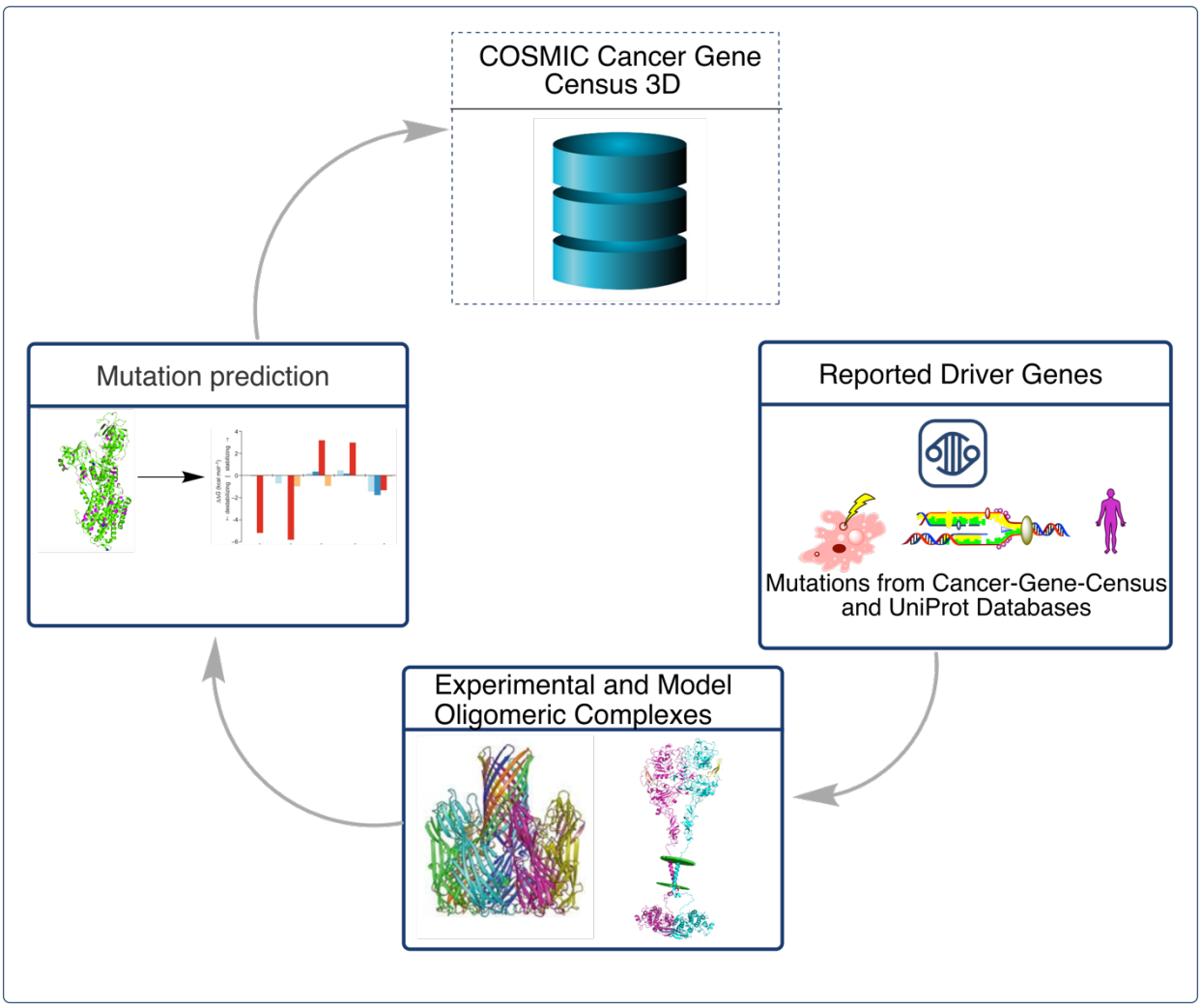
Söding, J. (2005) ‘Protein homology detection by HMM-HMM comparison’, *Bioinformatics*, 21(7), pp. 951–960.

Surleraux, D. L. N. G. *et al.* (2005) ‘Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor’, *Journal of Medicinal Chemistry*, 48(6), pp. 1813–1822.

Suzuki, Y. (2020) ‘Advent of a new sequencing era: long-read and on-site sequencing’, *Journal of Human Genetics*, 65(1), p. 10038.

Tsuganezawa, K. *et al.* (2013) ‘A fluorescent-based high-throughput screening assay for small molecules that inhibit the interaction of MdmX with p53’, *Journal of Biomolecular Screening*, 18(2), pp. 191–198.

Chapter I: Cosmic Cancer Gene Census 3D Databases



Introduction

Cancer background

According to the World Health Organization, cancer is the second leading cause of death, with an estimate of 10 million per year (Bray *et al.*, 2018). Cancer is a heterogeneous disease that can be found in different organs or tissues. The most common cancers are lungs, breasts, colorectum, prostate, stomach and liver, which account for more than 50% of all cases. The term “cancer” generically refers to a complex set of diseases that can be classified as follows (Wu *et al.*, 2018):

- **Malignant**: when healthy cells proliferate uncontrollably and invade other tissues, such as carcinoma, sarcoma, leukaemia and lymphoma.
- **Neoplastic** is abnormal tissue mass due to high cell proliferation, generally referred to as a tumour.
- **Metastasising** or **Advanced:** The tumour enters the lymphatic system or bloodstream and spreads to different tissues and organs, leading to a high mortality rate.

Multiple risk factors lead to cancer, such as smoking tobacco, alcohol abuse, unhealthy nutrition, lack of physical activity and chronic infections such as human papillomavirus (HPV) and hepatitis C. However, 30-50% of cancer deaths could be prevented by avoiding these risk factors (de Martel *et al.*, 2020). The neoplastic cell phenotype takes time to develop, which is why cancers are most often detected in elderly people; however, cancer is not age-specific and sometimes occurs in younger adults and children. In addition, the early detection of cancers leads to an enhanced response to therapies, examples of which are breast and colorectal

cancer, which have a very high success rate for treatment upon early detection (Laconi, Marongiu and DeGregori, 2020).

All cancers are driven by mutations in the deoxyribonucleic acid (DNA) sequence. Therefore, cancer is most often a corporeal disease caused by somatic mutations arising in multiple genes. The number of mutations accumulated in cells is between 1000-10000 somatic mutations, including a few deletions and insertions. Genes are programmed to control cell proliferation (Franco *et al.*, 2019). However, the accumulation of mutations in these genes induces increased cell multiplication and failure of DNA-repair mechanisms (Takeshima and Ushijima, 2019). The most common mutation types are base substitution, indel, rearrangements and amplification. The contributions of these mutations to cancer progression vary between gene and cancer type, resulting in a gain of function when mutations arise in oncogenes or loss of function when mutations arise in tumour suppressor genes (Li *et al.*, 2019).

Signal proliferation is one of the fundamental aspects of cancer development. It alters cell activity throughout the cell cycle. Healthy cells have multiple signalling pathways, tightly regulated in normal tissues, but these are generally defective in neoplastic cells. The signals are transmitted through transmembrane receptors that bind growth factors and other ligands, activating intracellular signalling pathways and cell proliferation (Sever and Brugge, 2015). Important examples of these are the tyrosine kinase receptors, which are often targeted for cancer treatment. Signalling pathways are usually controlled by a negative feedback mechanism that retro-inhibits the signalling cascade and maintains cell homeostasis. An example of negative feedback is PTEN, which regulates PI3-kinase activity by degrading PIP3 (phosphatidylinositol (3,**4**,**5**)-trisphosphate) to PIP2 (phosphatidylinositol (4,5)

bisphosphate), where the loss of PTEN function amplifies signalling and hence promotes tumour growth (Carracedo and Pandolfi, 2008).

It is known that many biochemical and cellular traits are shared by more than one cancer type. Hanahan and Weinberg established ten essential characteristic features known as hallmarks that are shared by most cancers arising from changes in cell physiology and collectively leading to the growth of malignant cells (figure 1) (Hanahan and Weinberg, 2000). However, this characterization of features is still subject to debate. For example, (Lazebnik, 2010) argues that five hallmarks are shared between benign and cancer cells and do not distinguish cancer features. The emphasis on hallmarks has shifted the focus of drug development primarily to hallmark targets, which may not always be clinically efficacious (Hanahan and Weinberg, 2011).

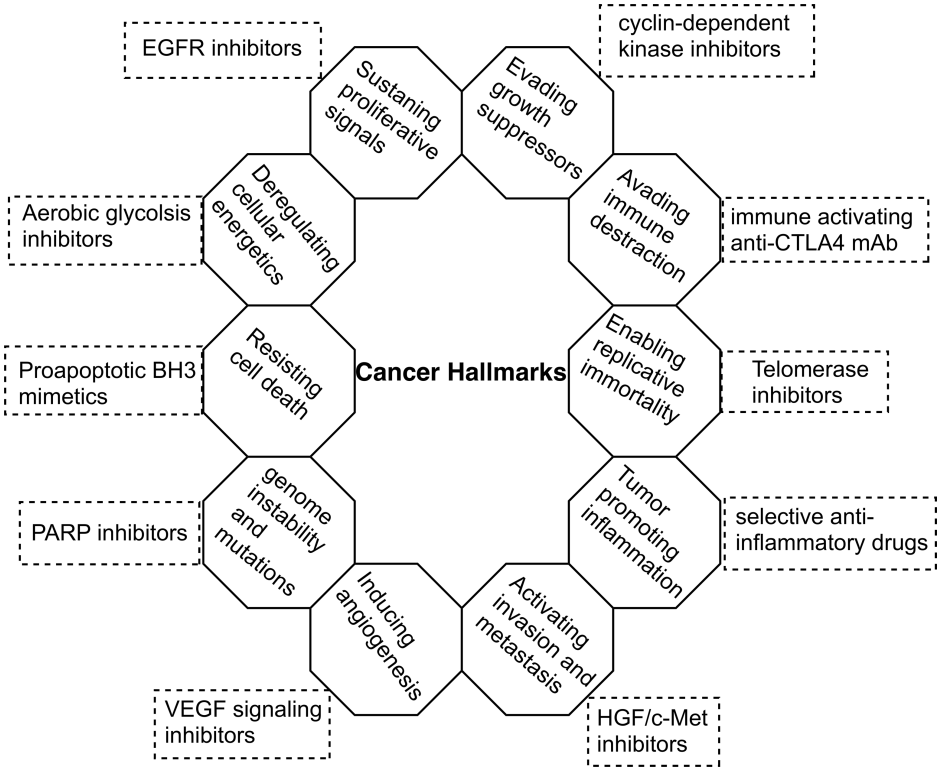


Figure 1 The ten hallmarks of cancer, a cornerstone concept in cancer biology, were first outlined by Hanahan and Weinberg in 2006 and expanded in 2010. These characteristics

are present in most neoplastic tissue cells and correlate with disease progression and prognosis. The octagons present the hallmark type, and the dash squares represent the therapeutic approaches developed toward each hallmark.

Charles Darwin 1859 (Darwin, 1859) described the evolution by natural selection based on diverse phenotypes among survival. At that time, mechanisms supporting phenotypes diversity by mutation and genetic recombination were unknown and still not entirely known today. Cancer is considered an evolution process developed through time and space (Merlo *et al.*, 2006). Cancer variation acts as a substrate of this evolution process and can be functional, known as a driver and non-functional, known as a passenger. Mutation accumulation results in tumour heterogeneity, which describes the diversity of tumours of the same cancer type. Identifying functional (driver mutations) that give cancer the advantage to grow from non-functional variation (passenger mutations) has remained challenging until now. This indeed raised an open question is cancer evolutionary developed through time and space or raised from the accumulation of mutations followed by a driver event (Matias and Degregori, 2011).

Nowadays, more is known about the type of mutations in cancer. However, little is known about the time it takes for the lesions to develop. Generally, mutations in cancer occur in three classes of genes. **Proto-oncogenes** mutations in these genes cause normal cells to become cancerous, also known as an enhancer of cancer. Typically, these genes are associated with stimulating cell differentiation and usually encode for a gene in the cell surface, for example, Epidermal growth factor receptor (EGFR) and encode intercellular genes such as KRAS. **Oncogene** is described as a result of a mutant version of Proto-oncogenes. In contrast, **Tumour suppressor** genes are associated with DNA repair, cell growth, cell division, and cell death to

control cells balance (Kontomanolis *et al.*, 2020). Most mutations in tumour suppressor are causing loss of function, whereas mutations in Oncogene result in a gain of function.

Mutations in Cancer

Nowadays, the evolution of DNA sequencing machines allows DNA sequences to be defined from many cancer samples with high accuracy and speed (Pareek, Smoczynski and Tretyn, 2011). However, over the past years, not much significant work has been done on how mutations lead to abnormal genes that drive cancer progression in multicellular organisms. Cancer mutations can be either somatic or germline (Figure 2).

- **Somatic mutations** involve altering the DNA nucleotide sequence, which can occur in any cell except the germ cell. Therefore, it cannot be inherited.
- **Germline mutations** are less common, occurring in the egg cell. The child can inherit them, i.e. the entire organism will be affected

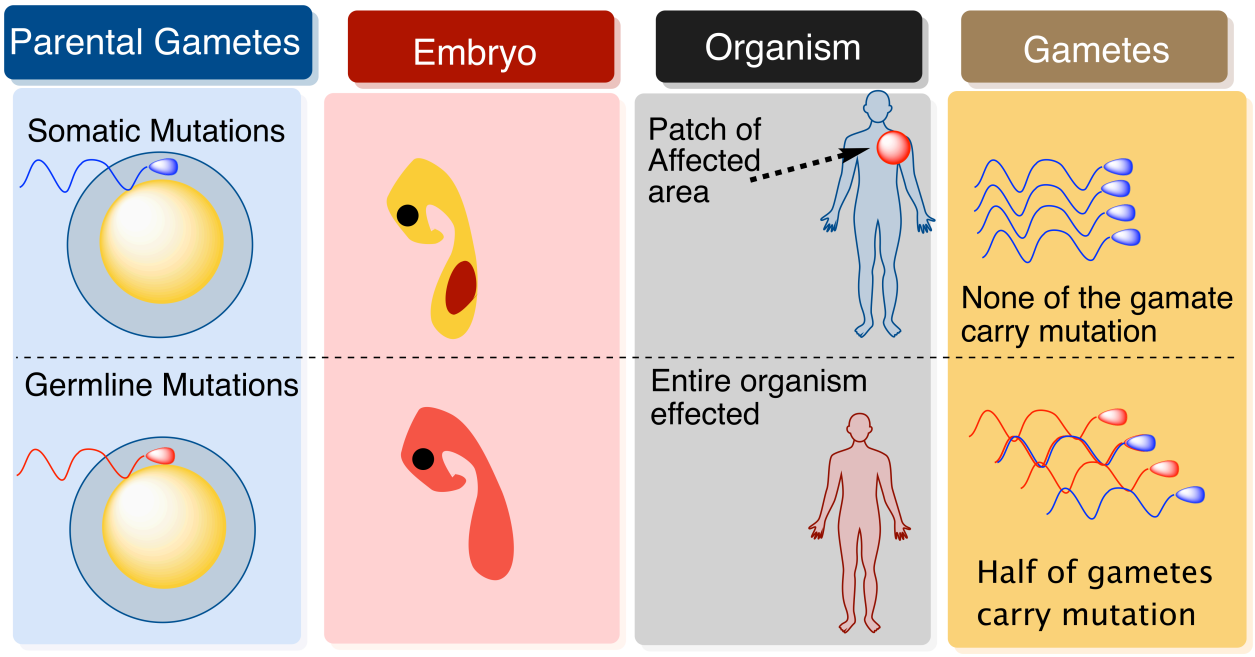


Figure 2 The effect of somatic mutations and germline mutation on organisms. The red area in the embryo indicates that the entire organism will be affected by germline mutations, whereas somatic mutations only affect specific tissues.

80% of mutations in cancer are somatic, 10% are germline, and 10% are both somatic and germline (Futreal *et al.*, 2004). The term “somatic mutation” encompasses several types. These include base substitution, also known as point mutation, which can be:

- **Silent** (synonymous) changes in the DNA codon with another codon, resulting in the same amino acid.
- **Missense** (non-synonymous) changes the DNA codon with another codon, resulting in different amino acids.
- **Nonsense,** where a change in the DNA codon results in a stop codon, usually cause a truncated non-functional protein.

Other somatic mutations are:

- **insertion** and **deletion**, which usually lead to frameshifts and a variety of outcomes.
- **Rearrangement,** when the DNA is broken and rejoining to another segment of DNA in the genome, results in an increased number of copy genes. This is also known as gene amplification.

Passenger and driver mutations

Data for thousands of mutations have been deposited in The Cancer Genome Atlas (TCGA) (Campbell *et al.*, 2020), Cancer Genome Project (CGP) (Weinstein *et al.*, 2013) and Catalogue of Somatic Mutations in Cancer (COSMIC)(Tate *et al.*, 2019). However, not all these mutations are drivers and contribute to malignant growth and

metathesis. Differentiating driver from passenger mutations is critical to understanding tumour development and targeted therapy. Driver mutations are defined as mutations that provide a selective advantage for tumour growth and cancer development. In contrast, passenger mutations are defined as mutations that do not directly provide tumour initiation and progression (Stratton, Campbell and Futreal, 2009). All cancer driver genes from TCGA, CGP, and the literature are listed in the COSMIC cancer gene census (723 genes) (Sondka *et al.*, 2018). Driver genes can be: *oncogene,* containing mutations that result in activation or a new gene function; *tumour suppressor;* mutations lead to the inactivation of the gene. Driver mutations in oncogenes usually arise from focal amplification or missense mutation, whereas driver mutations in tumour suppressor genes are generally caused by frameshift, nonsense mutations and focal deletions (Kern and Winter, 2006). There are, however, exceptions, such as the TP53 tumour suppressor gene, which is affected by missense mutations. Driver genes may not be recognised unless predominant mutations appear, such as BRAF V600E mutation (Kern and Winter, 2006). Mutations vary between different cancer types, and samples from the same cancer type, the frequency of copy number alteration (CNAs) (somatic changes in the chromosomal structure) and single nucleotide variant has been observed to correlate inversely across 12 cancer types, i.e. both mutations cannot be detected in the same cancer sample (Ciriello *et al.*, 2013). The CNAs mutations are dominant in breast and carcinomas samples, whereas the single nucleotide variants were dominant in acute myeloid leukaemia and glioblastoma (Bunting and Nussenzweig, 2013). Comprehensive identification of driver mutations should consider all mutations discussed above.

Estimates of 5-8 driver mutations are required for a cancer cell to develop (Stratton, Campbell and Futreal, 2009). However, the accuracy of this statement is debatable. In addition, the presence of more passenger mutations than driver mutations has led to the development of many bioinformatic tools that prioritize mutations for functional testing. This will be discussed more in the bioinformatic tools section. The percentage of nonsense, missense, indel, and frameshift mutations will suggest whether the driver gene is an oncogene or tumour suppressor. A 20/20 rule suggests that driver genes are oncogenic where >20% of somatic mutations are missense and repeated in the same position. In contrast, driver genes with >20% somatic mutations that are inactive are likely to be tumour suppressors. Identifying the driver genes in different cancer types will ultimately result in improved personalised treatment (Vogelstein *et al.*, 2013).

There are two ways to identify driver genes and differentiate driver mutations from passenger mutations: (i) frequency-based approaches and (ii) functional-based approaches. Genes mutated more than the background mutation rate are considered drivers and more likely to be cancerous (Lawrence *et al.*, 2013). Once the driver gene is identified, the frequency-based method can be applied to determine which individual mutations are causing that gene to be a driver oncogene or tumour-suppressor. However, this method needs a large sample size. For example, the BRAF V600E mutation and infrequent driver genes and mutations may not be detected.

Estimating the background mutations (synonymous mutation) between different cancer types is almost impossible since the difference between the cancer samples is more than 1000-fold (Hodis *et al.*, 2012). In contrast, the variation between samples from the same cancer type is estimated to be 5-fold due to false-positive

increases underestimating variability. Furthermore, it has limited accuracy in the gene with low background mutations (Lawrence *et al.*, 2013). Therefore, an alternative way to estimate the background frequency is to investigate mutations in the untranslated and the intron regions since mutations in this region evolve high rate, similar to synonymous base substitution, with the assumption that mutations in these regions are neutral, which is not always true for example mutations in intron impact RNA stability and translation, omics-scale studied showed dysregulation of the expression of the noncoding region had been implicated in cancer (Shahrouki and Larsson, 2012). Another approach models the background mutations through GC content rates, genomic features, and gene density which account for <40% of the variance in cancer (Hodgkinson, Chen and Eyre-Walker, 2012).

Another way of identifying driver mutations from passenger mutations is through function-based methods. Unlike the frequency-based method, the sample size is not an issue, and mutations can be inferred from one sample. Function-based methods study the impact of mutations in conserved regions, functional domains, and biochemical mutations similarity (Hodgkinson, Chen and Eyre-Walker, 2012). However, not all mutations in the conserved regions are drivers, and not all mutations in non-functional regions are passengers (Fröhling *et al.*, 2007). Function-based mutation approaches have been used for personalized medicine by selecting chemotherapy based on the mutation present in the sample for tumour treatment. In addition, protein-protein interactions networks are also essential in finding new driver genes; for example, in a protein that interacts with many other cellular proteins, i.e. central nodes, mutations can alter these interactions (Szklarczyk *et al.*, 2021). Recently transcriptome data has been used to correlate mutations in the transcription factor to target gene expression (Shah *et al.*, 2012).

Cancer is very complex and heterogeneous, and the role of the microenvironment has brought much attention to study gene function in co-culture and different organisms. However, different organisms have other cell biology than humans. For instance, most mouse cells have active telomerase, reducing the number of mutations but leading to more potent mutations, which is not the case for human cells (Jacks *et al.*, 1994).

Infrequent driver mutations.

The progression of cancer is determined by many driver genes rather than one frequently mutated gene. For example, in ovarian carcinoma, cellular tumour antigen p53 (TP53) has been observed to be mutated in 97% of the samples, along with other eight genes where mutations are found in low prevalence (Bell *et al.*, 2011). On the other hand, some genes were overlooked and not considered drivers until large samples were applied (7,299 exomes). For example, the Homeobox protein cut-like 1 (CUX1) mutated only in a low prevalence 1-5% in different cancer samples (Wong *et al.*, 2014). Similarly, a driver mutation sometimes appears in the gene's infrequently mutated region and can be overlooked, for example, in receptor-type tyrosine-protein kinase (FLT3) in adult acute myeloid leukaemia (AML) (Fröhling *et al.*, 2007). The common mutations that contribute to leukemogenesis appear in the kinase domain and the juxtamembrane but not outside these regions. Functional testing of other mutations conferred gain-of-function that activates downstream signalling. This raised the question of why so many driver genes are rarely mutated in cancer? Functional redundancy in the same pathway is believed to be the answer to this question. Pathways are defined as a series of protein interactions that lead to a specific product. i.e. if there are multiple genes in the same pathway classified as a

driver, the first mutation of any gene provides a selective advantage for cancer to grow and proliferate, making the other gene mutations rare and not act as a driver (Hua *et al.*, 2013). This specification has been observed in oncogene genes but not in the tumour-suppressor genes. This stipulation shifted the attention to identifying the driver pathway in which genes act as a driver. Driver pathways can be defined in a very similar way to candidate driver genes. The candidate driver pathway is considered when entire genes in the pathway are mutated more than background mutations. In this way, we overcome the sample issue we encounter in the frequency-based method since we look at all genes collectively instead of individually. Suppose we have a set of driver genes that did not mutate significantly, then the background mutations cannot be identified statistically. However, identifying driver pathways by looking collectively at all genes and inferring the driver genes is the best way to identify infrequent driver genes.

Bioinformatic tools for putative driver genes and mutations

Over recent years, a wealth of mutation data has been deposited into TCGA. This has motivated a vast development of bioinformatics tools to characterize and discover new putative driver genes, mutations, and pathways (Figure 3). However, these tools tend not to agree on specific driver gene/mutation leading to false-positive findings.

Gene-Based	Sequence-Based	Structure-based	Omic/network	Drugability	Mutations-Data	3D-Clustering
20/20+	SIFT	SDM	DriverNET	PHIAL	COSMIC	COSMIC-3D
e-Driver	MutationAssessor	mCSM	OncoIMPACT	DEPO	ClinVar	3DHotspots
MutSig2CV	VEST	Fold-X			OncoKB	Cancer3D
ActiveDriver	Polyphen2	I-Mutant			ClinGen	CRAVAT
MuSIC		Maestro			CIViC	

Figure 3 Bioinformatic databases and tools commonly used to report mutations and predict the impacts of somatic mutations on protein stability. The developed tools are divided into six sections: gene-based, which detect putative driver genes based on background mutations; sequence-based methods which predict the effect of mutations based on the amino acid sequence; structure-based tools that require 3D structure to predict the impacts of mutations; omic/network tools that drive genes based on cell pathways; and finally, 3D clustering that maps mutations to the 3D structure.

The presence of so many mutations makes it impractical to validate all of the mutations experimentally. More than 20 bioinformatics tools have been developed to predict the impact of mutations. The most common and convenient are the frequency-based and function-based methods. Gene-based tools such as MutSig2CV (Lawrence *et al.*, 2014) detect gene background mutations using synonymous mutational rate and genomic features, reducing the extensive false-positive selected driver gene. Sequence-based tools such as Polyphene2 (Adzhubei *et al.*, 2010) do not require a 3-dimensional (3D) structure to predict the impact of mutations, but rather it uses BLAST+ to generate a multiple sequence alignment (MSA), which is then further refined by Leon software. The final score is based on physical and evolutionary comparatives generated from the MSA. Structure-based methods require access to 3D structure from the Protein Data Bank (PDB) (Berman *et al.*, 2000) or through comparative modelling. The stability prediction is based on the change between folded and unfolded states or free energy changes (ΔΔG). For example, in Site-Directed Mutator (SDM) (Pandurangan *et al.*, 2017), the stability score is derived from the statistical potential energy function using an environmental-specific amino acid substitution table. Machine learning tools, such as mCSM (Pires, Ascher and Blundell, 2014), evaluate multiple features of mutations using graph-based signature encoding distances between an atom in the structure. One of the

drawbacks of the machine-learning methods is that more mutational data are needed to get a good training set. Furthermore, overfitting data can occur when more features from different cancer types are included. Accuracy is limited to the data set on which the algorithm was trained; it will perform poorly for mutation types not represented in the training set.

Zhang et al. have compared eight bioinformatics tools; no method reached an accuracy higher than 79%, and most of the tools showed a lack of specificity. Unfortunately, there are no tools available that will give both high specificity and sensitivity (Gnad *et al.*, 2013). However, meta-prediction, also known as the rule-based approach, which integrates the results from several developed methods into one using a consensus score or weighted average, is one way to reduce false positives obtained from individual tools despite some tools disagree entirely (Tang and Thomas, 2016). Furthermore, combining the results from different tools based on statistical methods, machine learning, and evolutionary conservation outperformed individual tools that could lead to an unbiased analysis (Bailey *et al.*, 2018).

PopMuSiC (Dehouck *et al.*, 2011), and OncodriveFM (Mularoni *et al.*, 2016), are function-based methods that have been used to study mutations from 12 different cancer types in 3,200 samples. PopMuSiC predicted 232 driver mutations, whereas OncodriveFM predicted 259 driver mutations. Sixty-eight mutations were predicted as drivers by both methods (Ciriello *et al.*, 2013). In another study, the IntOGen-mutations pipeline used frequency and functional approaches to predict diver mutations from 4,623 exomes in 13 cancer sites. It ranked the most likely driver genes in 13 cancer types (Gonzalez-Perez, Jene-Sanz and Lopez-Bigas, 2013). Of

course, bioinformatic tools used to predict driver genes or driver mutations will never be absolute; but they are invaluable tools in prioritising mutations for functional tests. Comprehensive characterisation of 9,423 tumour exomes samples identified 299 driver genes using 26 computational tools. In addition, more than 3400 putative mutations were identified by sequence and structural-base analysis. Experimental validation showed that 60-80% of these putative mutations are likely to be a driver. The *in-silico* saturation mutagenesis machine-learning-based method (boostDM) developed to evaluate potential oncogenic drivers in 185-specific tissue in humans outperforms the experimental saturation mutagenesis in identifying drivers (Bailey *et al.*, 2018).

Methods

Modelling gene products in CGC.

Protomer Modelling

The target gene sequences were retrieved from the COSMIC cancer gene census (COSMIC-CGC) (Tate *et al.*, 2019) and blasted against the Protein Data Bank (PDB) to calculate the structural coverage. The DISOPRED3 tool (Jones and Cozzetto, 2015) was used to distinguish disordered unfolded regions from the globular domain of the target sequence. The gene sequence was screened against several databases to inform the modelling, including the Pfam database for an extensive collection of protein families annotated using hidden Markov models (Finn *et al.*, 2014). CATH: database provides evolutionary relationships between protein domains (Sillitoe *et al.*, 2021). SCOP: database classified protein based on 3D similarity (Andreeva *et al.*, 2014). SMART: database provided annotation of

signalling domain that usually multidomain and not detected by Pfam, functional class, and functionally important residues (Schultz *et al.*, 1998). UniProt: database provided comprehensive annotation on protein sequence, function and interaction (Bateman *et al.*, 2021).

After sequence and structure annotations were retrieved, protein sequences not associated with the experimentally solved structure were searched against PDB using FUGUE, a sequence-to-structure comparison program, which recognises distance homologues using environment-specific substitution table as well as structure-dependent gap penalties (Shi, Blundell and Mizuguchi, 2001). Position-Specific Iterative (PSI-BLAST), another sequence-to-structure comparison program, defines a profile-profile alignment or position-specific score matrix (PSSM) from multiple alignments (Ramsay *et al.*, 2000). HHsearch, another sensitive tool used to detect distant homologues, uses Hidden Markov Models (HMMs) (Söding, 2005). Selected templates with multiple experimental techniques such as X-ray, NMR, and cryo-EM were based on identity, coverage and resolution, i.e. the higher resolution structures were selected over lower solution and X-ray templates were favoured over NMR or cryo-EM templates.

All the templates were inspected visually to prevent long loops usually generated because of the gap in the PDB file of the selected template. Finally, the target sequence was aligned against selected templates using Clustal Omega (Sievers *et al.*, 2011) multiple sequence alignment software and the alignment file produced by Clustal Omega was processed to MODELLER version 9.23 to generate the final target model (Šali and Blundell, 1993) (Figure 5).

Homodimer and Heterodimer Modeling

All higher-order assemblies are modelled based on evidence from the literature and UniProt annotation. Building homo/heterodimer models are divided into four parts:

I. Gene product with no structural representation in the PDB.
II. Gene product with structural representation in the PDB, but each domain is solved separately.
III. Gene product with one domain solved as a dimer, and another domain solved as a monomer.
IV. Gene product with domain solved in a different conformational state

Including co-factor as well as nature ligands are essential for biological relevance. Therefore, DNA binding protein modelled with DNA, GTPase proteins modelled with GDP and GTP, metal ions were included where possible such as in the zinc-finger domain.

For example, Cyclic AMP-dependent transcription factor ATF-1 included the Basic leucine Zipper Domain binding to DNA as a dimer. Therefore, each domain was modelled separately using PDB ID: 5ZKO as a template to generate the final homodimer structure (Figure 4A).

An example of the second and the third points is RAC-beta serine/threonine-protein kinase (AKT2), the PH domain solved as monomer PDB ID:1P6S, covering the amino acid sequence from 1-111, and the kinase domain solved separately as homodimer PDB ID:3D0E covering the amino acid sequence from 146-480. The unsolved amino acid region is between amino acids 112-145, which are predicted to be disordered. The homodimer structure was obtained by modelling the three regions together to obtain the full protomer. Then, each protomer is superimposed on the solved kinase homodimer structure to generate the final model (Figure 4B).

Proteins with single-pass membrane helices, known as bitopic proteins, are retrieved from the Membranome database, which provides structural information on more than 6000 bitopic proteins. Furthermore, the orientation of proteins in membranes (OPM) server was used to position all transmembrane proteins included in the COSMIC CGC 3D (Lomize *et al.*, 2006). For example, in the Ephrin type-A receptor 3 (EPHA3), the membrane single-pass region was taken from the Membranome database and incorporated into the final model (Figure 4C).

Gene product structures that have been solved experimentally in two different conformations were taken into consideration. For instance, in activin receptor type-1 protein (ACVR1). The kinase domain was experimentally solved in two conformational states (PDB ID: 4DYM, 6GIN). Therefore, two independent models were generated. The first ACVR1 conformation was generated by modelling the full protomer, i.e. kinase domain linked to single-pass transmembrane and extracellular region. Then superimposed the full protomer to the kinase domain (PDB ID: 4DYM) and extracellular hetero 4-mer (PDB ID: 3EVS) to get the complex of hetero 4-mer. The second conformation was modelled similarly, except the PDB ID: 6GIN of the kinase was used as a template to superimpose the full modelled protomer (Figure 4 D/E).

Intrinsically disordered regions (IDRs) were detected using DISOPRED3 and incorporated within the single domain or in the multidomain structure. Thus, in the examples mentioned above, the disordered region in EPHA3 modelled structure was included between the kinase and the transmembrane, whereas in the AKT2 modelled structure was included between the kinase domain and the PH domain.

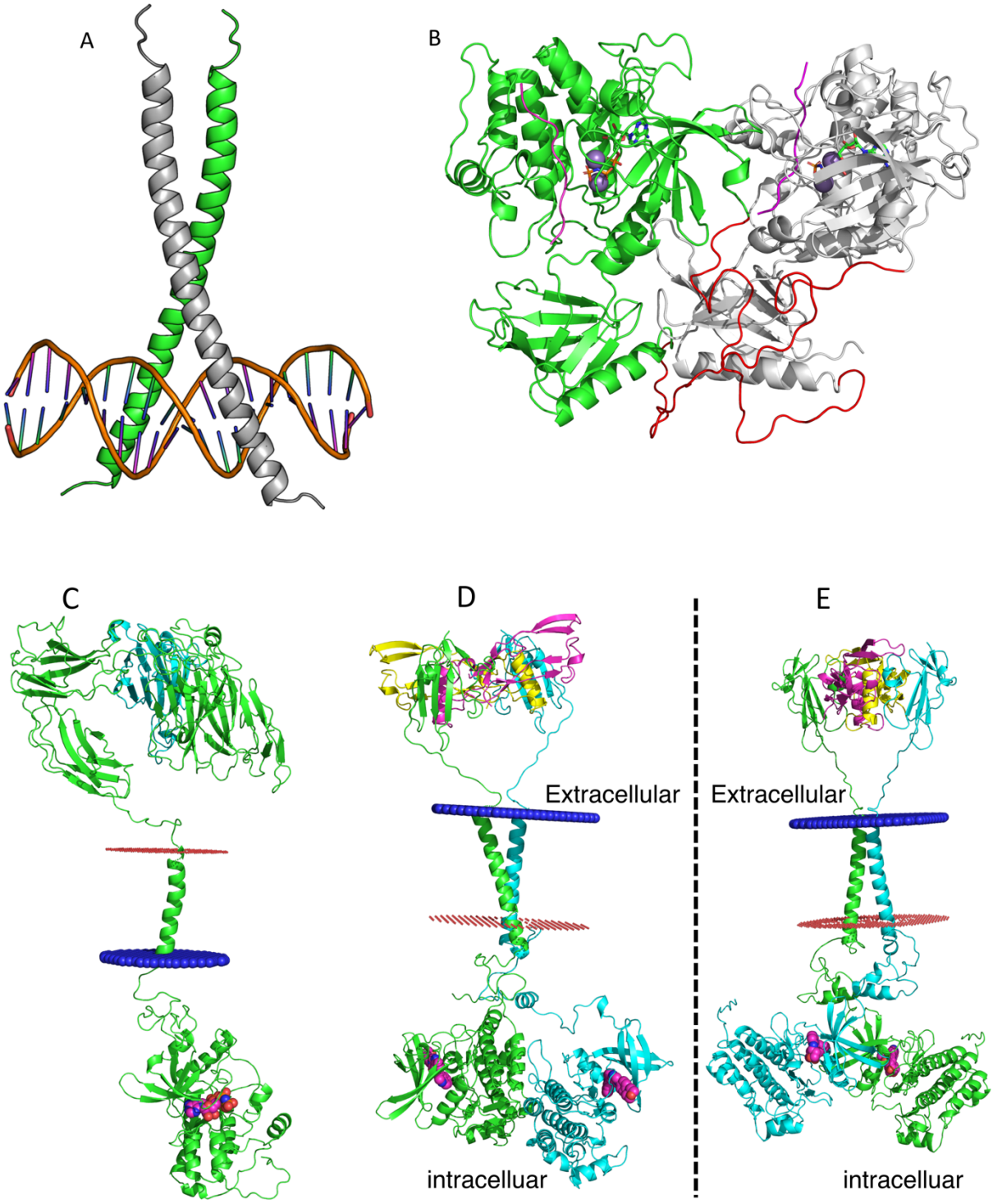


Figure 4 (A) Cyclic AMP-dependent transcription factor ATF-1, modelled as a homodimer, coloured in green and grey and DNA in light orange. (B) RAC-beta serine/threonine-protein kinase (AKT2) PH domain (PDB ID: 1P6S), covering the region between amino acids 1-111 coloured in wheat, the intrinsically disordered region between 112-145 coloured in red, and a protein kinase domain (PDB ID:3D0E) covering the region between 146-480 coloured in light purple-blue. ANP (phosphor-amino phosphonic acid-adenylate ester) ligands in stick, manganese ions in spheres dark purple-blue, and glycogen synthase kinase -3 beta in

magenta. (C) Ephrin type-A receptor 3 (EPHA3) single-pass transmembrane is in green, whereas ephrin -A5 is coloured in cyan. The ligand binds to the kinase domain represented in sphere magenta. (D/E) Activin receptor type-1 (ACVR1) heterotetramer comprising ACVR1 homodimer represented in green and cyan. The ligand in the cytoplasmic region is coloured in yellow and magenta, where all the transmembrane regions are highlighted between red/blue circular structure, which represents the protein-membrane region

All heterodimer complexes are built from selected templates. No protein-protein docking was carried out for any of the modelled oligomers in the COSMIC CGC 3D. From the example mentioned above, ACVR1, the solved hetero 4-mer, was constructed using selected templates. Since the goal is to bring biological relevance to the built models, all homo/heterodimer complex interface regions were evaluated using a well-developed tool PISA (Protein Interfaces, Surfaces and Assemblies) (Krissinel, 2015). PISA is a tool that calculates macromolecule surface and interfaces from the model coordinates file generated by MODELLER. The output generated represents a score for interface area, interface hydrophobicity, hydrogen bonds and salt interactions.

Protomer homo/heterodimer side-chain optimization was essential for improving the quality of the final models. A minimised side-chain implemented into Foldit (Kleffner *et al.*, 2017) was used to remove clashes between homo/heterodimer interface and protein-ligand clashing.

Quality assessment is one of the essential steps in comparative modelling. Selecting which method to evaluate the built model can be challenging but is important since it will determine model utility. Furthermore, the model's accuracy depends on the availability of high-resolution templates. Two quality assessments were implemented

to evaluate the comparative modelling. First, PROCHECK (Laskowski *et al.*, 1993) was downloaded from the EMBL-EBI website and installed locally. It produced a PostScript plot of stereochemical analysis of residue geometry. PROCHECK also highlighted regions that needed further attention. Since IDRs were included in our final models, PROCHECK proved very useful. The second quality assessment implemented is MolProbity (Chen *et al.*, 2010), which evaluates all-atom contact analysis, side-chain rotamers, Ramachandran criteria, steric clashes, C deviation, and dihedral angles. The MolProbity overall score represents the log-weighted value of all these features.

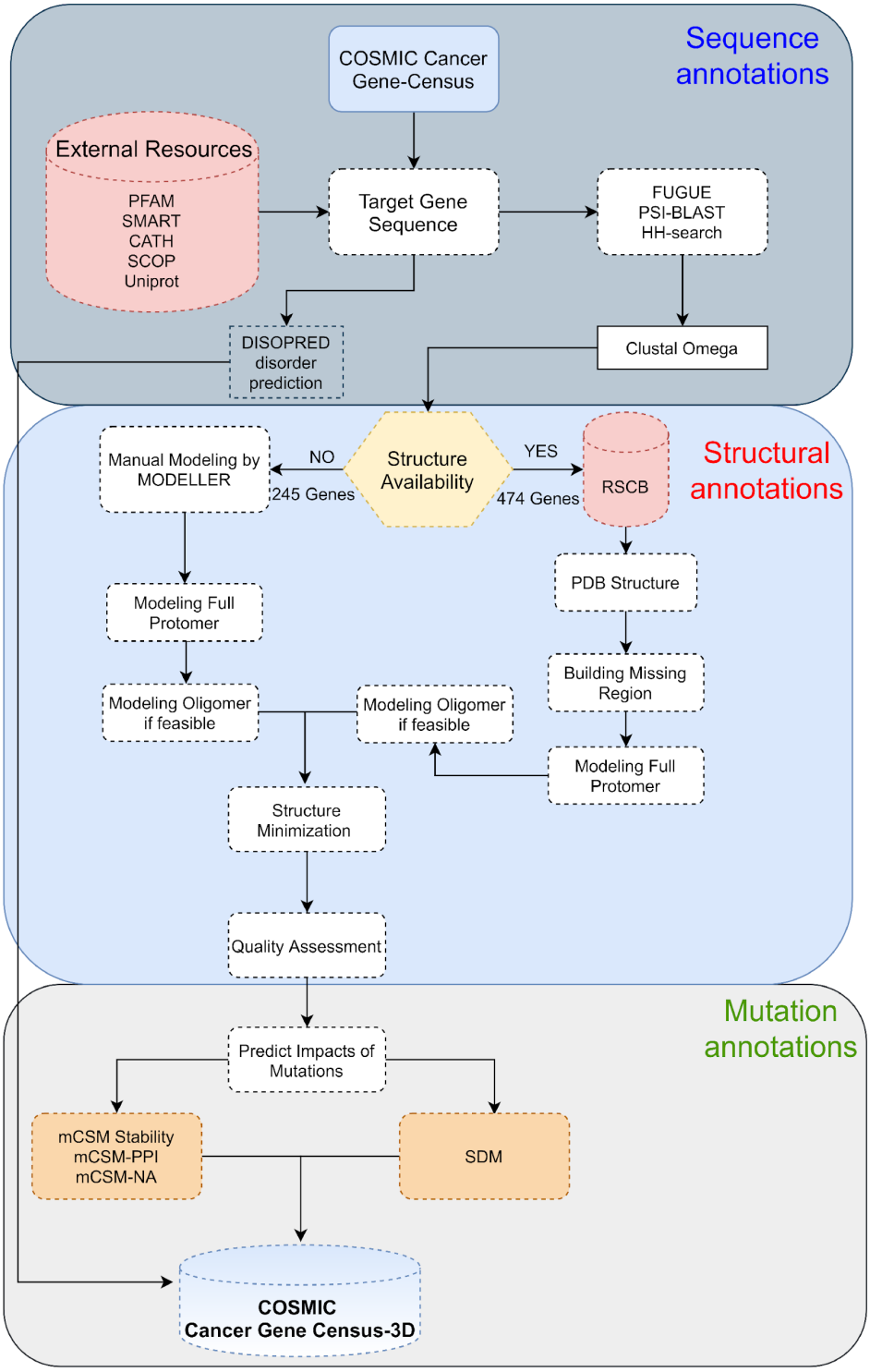


Figure 5. Simplified flowchart for manual modelling pipeline, starting from the target gene in COSMIC CGC, annotating gene domains, finding PDB hits and homologous structures, building model complexes and finally predicting the impact of the mutations on the modelled structures.

Mutation data

Over the past year, many mutation data have been generated. COSMIC CGC database is populated with curated driver and potential driver genes based on evidence from the literature. All missense mutations for 714 genes were downloaded from the COSMIC CGC (https://cancer.sanger.ac.uk/cosmic/download). Mutation data were pre-processed and filtered, and only missense mutations were processed. Mutation frequency was calculated based on recurrence. Two independent structure-based methods were used to predict the impact of mutations and hypothesise putative mutations that affect protein stability. The first of these, SDM, is a knowledge-based statistical method that can predict only protein stability. The second tool is mCSM, a machine learning method that predicts the impact of mutations on stability (mCSM-Stability) as well as on the protein-protein interface (mCSM-PPI) and protein-DNA interaction (mCSM-NA). These data are represented in the **mutation table** (Figure 6) in the COSMIC CGC 3D database (Figure). mCSM-PPI was used for homo/heterodimer models, whereas mCSM-NA was run if DNA or RNA were included in the final modelled structure.

Mutations	Mutations Frequency	mCSM Stability	mCSM PPI	mCSM NA	SDM Stability
K315N	29	-2.082	-0.19		-0.19
D177E	11	-0.561	-0.731		0.58
I458T	5	-2.791	-0.901		-2.16
R400C	4	-2.204	-1.247		-1.19
P125S	4	-0.28	-0.184		0.07
L310Q	4	-0.768	-0.479		0.15
R400H	3	-2.233	-0.472		-0.22
P118L	3	-0.711	-0.618		-0.14

Figure 6. Representation of mutations in the COSMIC CGC 3D database. The first letter amino acid is the wild type in the mutations column, followed by the amino acid position, and the second letter amino acid is the mutant. The mutation columns, by default, are sorted by frequency. However, the table can be sorted based on mCSM stability, mCSM PPI, and SDM stability predictions.

Website development

All data were stored in the PostgreSQL database (version 9.3). The front-end, also known as the client-side user interface, was developed using HyperText Markup Language (HTML), Cascading Style Sheets (CSS), Bootstrap version 4, and jQuery (https://jquery.com/). The back-end was developed using Express.JS, a web application framework using node.js. Express (version 4) is used in the back-end to fetch information from the stored table, and Embedded JavaScript (EJS) was used as a template engine to render the final HTML page. All the tables were produced by DataTable JavaScript (https://datatables.net/), which allows the user to search keywords and sort data. A high-performance graphic MolStar viewer was implemented to view all model coordinate files and the experimentally solved structures (Sehnal *et al.*, 2021). The protein feature viewer (Watkins *et al.*, 2017) was implemented to view all target protein sequence features (Watkins *et al.*, 2017). An open-source multiple sequence alignment visualization (MSAViewer) was

implemented to view target template sequence alignments (Yachdav *et al.*, 2016). In addition, the d3.js heatmap was implemented to produce colour scale and hover effect dynamic interactive mutations data. The website is designed for large (laptop), medium (iPad) and small (iPhone) devices and have been tested in multiple popular browsers such as Chrome, Safari, and Firefox. All source code of the COSMIC CGC 3D (Ali F Alsulami *et al.*, 2021) is stored in GitHub private repository. This author is responsible for continuous updates, backup, and error repair.

Results

Website analysis

Home page

The COSMIC CGC 3D home page (Figure 7) has a short description that provides overall structural coverage and the importance of developing this database. In addition, it has a search form where users can query the database with UniProt ID (Q04771), Gene ID (90), and Gene name (ACVR1). The button is under the search form will show the result page as a demo. The navbar at the top of the page designed using bootstrap 4 provides a link to five different pages.

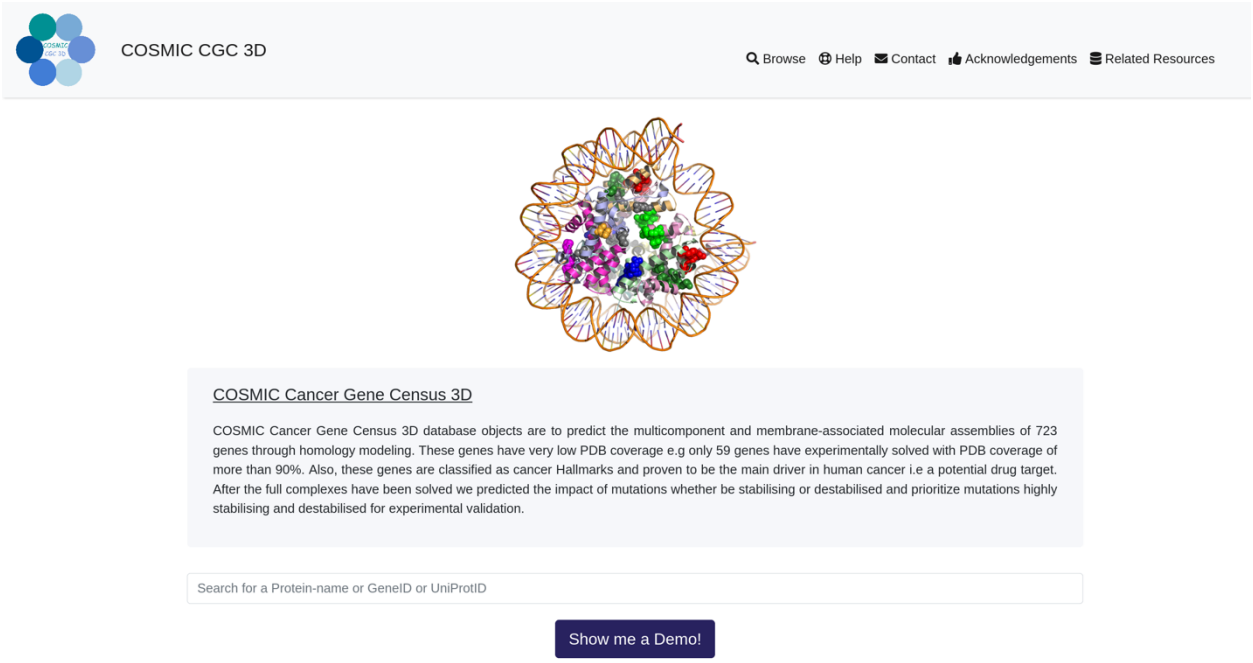


Figure 7. COSMIC CGC 3D home page. A short description, including some statistics of the modelled genes shown in the middle of the page. The nav bar at the top of the page is used to navigate through different pages.

Browse page

The COSMIC CGC 3D browse page (figure 8) provides all genes deposited in the database. The table contains gene ID, UniProt ID gene name, protein coverage, percentage intrinsically disordered and a short description of each gene. In addition, the table can be sorted and searched for a keyword.

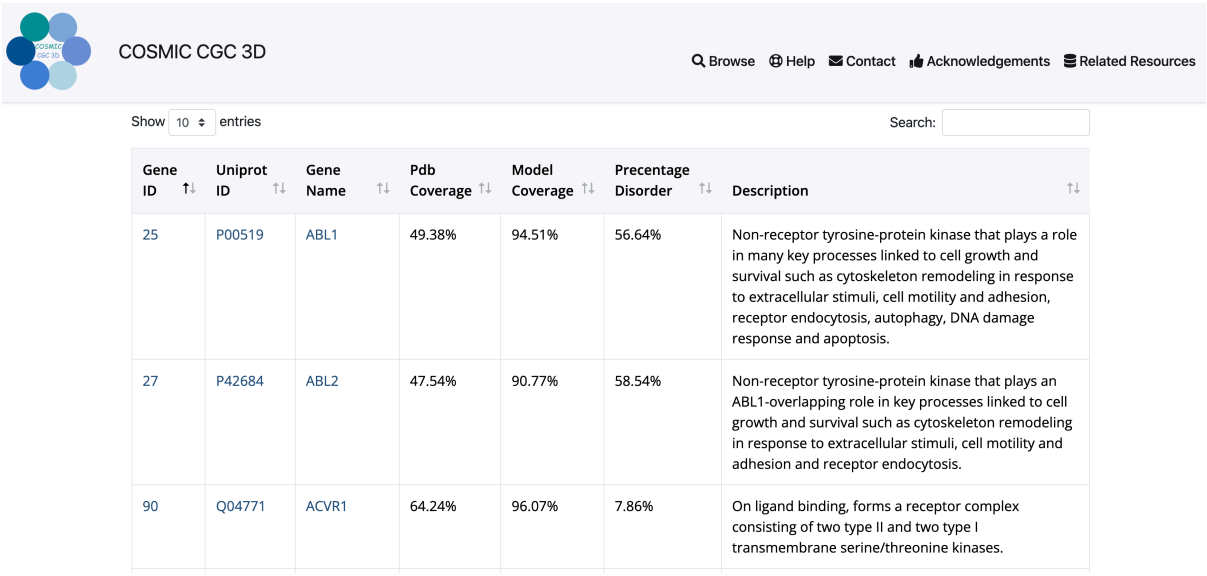


Figure 8. The browse page. All genes deposited in the COSMIC CGC 3D can be found on this page. In addition, essential information such as description, model coverage, and PDB coverage are included.

Contact and Acknowledgements pages

The contact page was developed to receive questions from the users, whereas the acknowledgements were developed to thank all the funding and support received. (Figure 9B)

Help page

The help page (figure 9A) in COSMIC CGC 3D provides a graphical and textual description of how the database can be queried. It offers the user detailed guidance

on better use of the tools implemented. In addition, it describes in detail how data can be fetched programmatically using Application Programming Interface (APIs). The currently implemented API includes queries by gene_id, gene_name, and uniprot_id represented in the queries section. A query takes a single identifier such as 90, or ACVR1, or Q04771 and returns the information about the gene in JSON files. Examples https://cancer-3d.com/api/models/Q04771 OR https://cancer-3d.com/api/models/90 OR https://cancer-3d.com/api/models/ACVR1

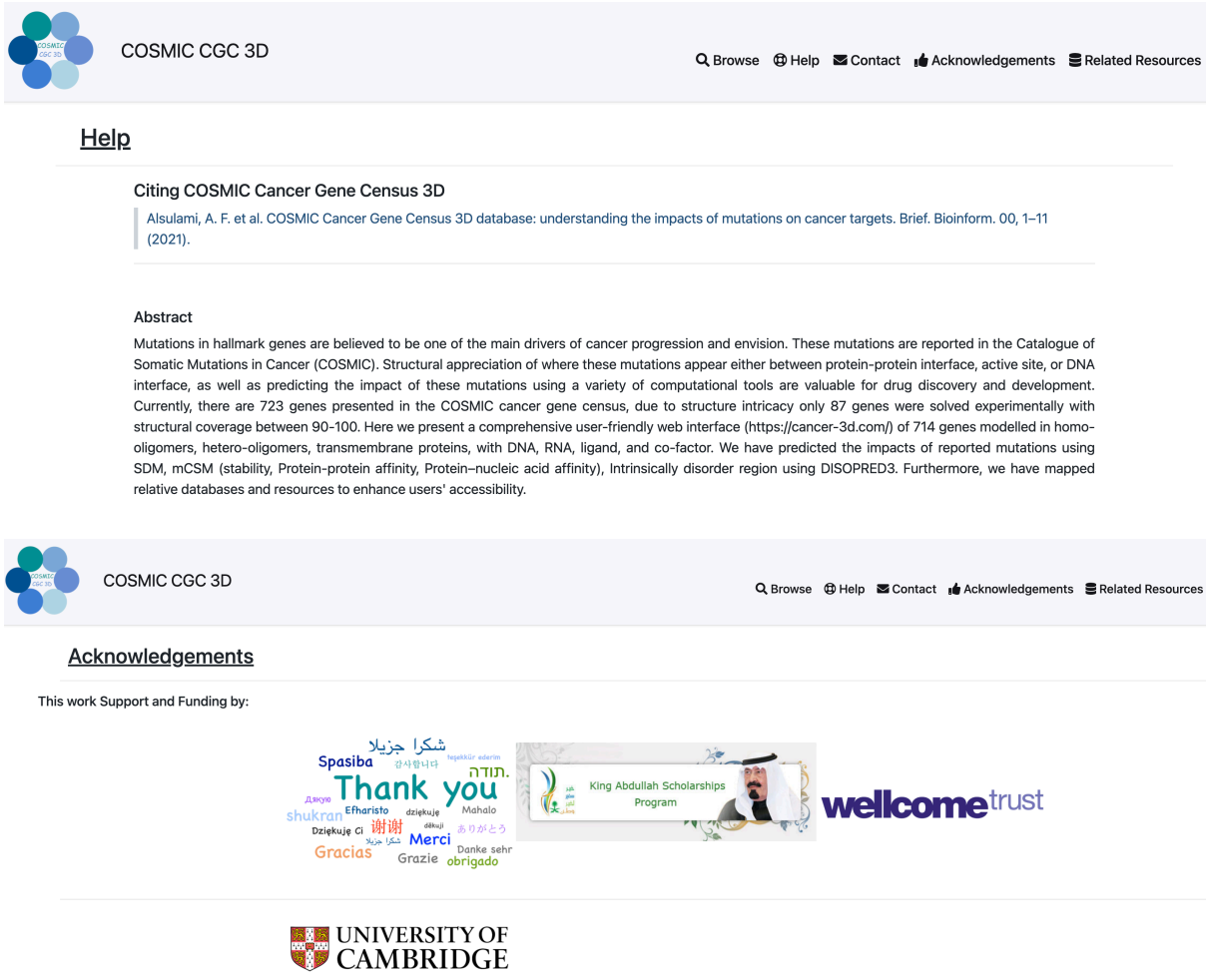


Figure 9. Help and Acknowledgment pages. It includes all tools implemented in COSMIC CGC 3D, how the database can be queried and API usage.

Related resource page

COSMIC CGC 3D integrates multiple tools used to build the website and generate the data. All the tools are freely available under academic license. URLs of these resource publications and websites are provided. (Figure 10)

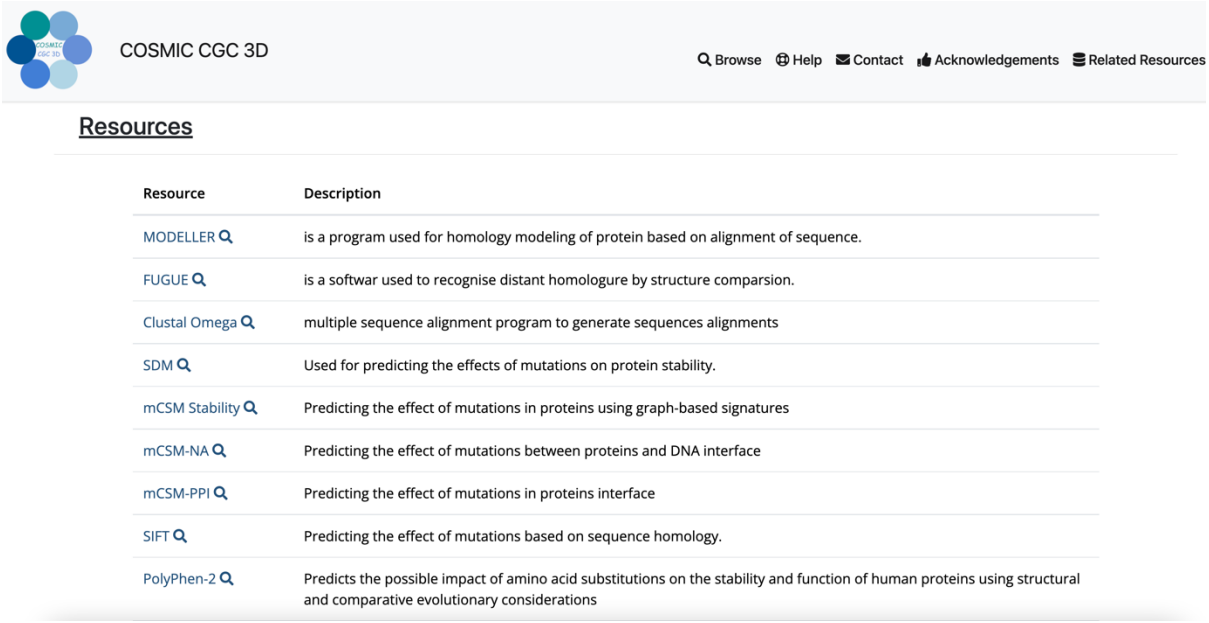


Figure 10. Resource page. It includes all external software used to generate the data deposited into the COSMIC CGC 3D.

Result page

The main page of the COSMIC CGC 3D is the results page (Figure 11). It has all the data about the queried target gene. It involves four components:

I. Gene description and external resource URLs.

II. Structural and sequence annotations

III. Models and PDB tables

IV. Mutations table and heatmap.

To reduce data redundancy, each gene has a short description and URLs to a different external database such as UniProt, MobiDB (Piovesan *et al.*, 2021), Pfam, and COSMIC. The Models/PDB table is divided into two parts. The first contains disordered sequence percentage calculated using DISOPRED3, models coverage and experimental coverage. The second part gives information about the built model, including model coordinates files, templates selected to build the model, oligomeric state of the model, oligomeric interface analysis, quality assessment analysis by PROCHECK, MolProbity and extra model information. The multiple sequence alignment of the target and templates can be viewed under the Model/PDB table using MSAViewer. In addition, the user can colour the alignment based on hydrophobicity and nucleotide etc. and visualize the identity score along with other features.

All model coordinate files in the Model/PDB table and experimentally solved structures in the PDB table can be visualized using MolStar. Each chain ID is coloured differently. Furthermore, interaction through stacking and hydrogen bonding can be viewed by clicking on the target residue or ligand. This helps visualize the gain or loss of interaction between the wild-type and the mutant. Furthermore, the IDRs of each gene target are directly plotted under the MolStar viewer so that the user can correlate the IDR sequences with the target model. Finally, the solved experimental structures were retrieved from the PDB where available.

The sequence and proteomics annotations can be visualized using ProtVista, which is a JavaScript component. The various annotations ProtVista included are domain and site, topology, i.e. transmembrane, variants, and post-translational modification. By default, each of these categories are collapsed. Visualizing all these features together facilitates the identification of patterns related to protein functions.

The mutations table includes mutation frequency and prediction from structure-based methods SDM and mCSM (Stability, PPI, and NA). In addition, the heatmap is implemented to make it easy for the user to visualize and spot the difference between stabilising and destabilising mutation prediction. The heatmap scale indicator represents a dark blue square for destabilising mutations and a red square for those that are stabilising. In addition, these heatmap indicators are included in the mutation table. Thus, all data represented in the COSMIC CGC 3D can be downloaded and visualized and analysed.

Figure 11. The results page. These include all data: gene name with links to external data sources, including UniProt, Pfam, COSMIC and COSMIC-3D; a brief description of queried gene. Models and PDB tables provide information about the modelled and experimental structures. MolStar can be used to display 3D structures of the target-gene models; UniProt viewer is used to visualise domains; DISOPRED3 predicts disorder of the target gene, and the heatmap and mutations table provides predictions of the impacts of mutations reported in COSMIC CGC.

Data statistics

There are 715 genes deposited in the COSMIC CGC 3D. The largest predicted model structures deposited in the COSMIC CGC 3D database are: low-density lipoprotein receptor-related protein 1B (LRP1B), which is a cell surface protein (4599 residues); Protocadherin Fat 1 (FAT1) cell membrane protein essential for cell-cell contact (4588 residues); and Protocadherin Fat 3 (FAT3) another cell membrane protein probably involved in a neurites derived interaction. On the other hand, the smallest protein we have modelled is cytochrome c oxidase subunit 6C (COX6C), one of the cytochrome c oxidase components in the mitochondrial electron transport chain.

There are eight genes presented in the COSMIC CGC without mutation annotations (table). Therefore, no models have been built for these genes. The largest reported gene in the COSMIC CGC is the Mucin-16 (MUC16), coding for 14,507 amino acids. In addition, there are 283 genes the COSMIC CGC classified as hallmarks. These genes are involved in multiple metabolic pathways giving cancer cells the advantage to growth and proliferation. Unfortunately, only 59 gene products have experimental structure coverage of more than 90%. In contrast, 410 gene products have the structures of regions/domains solved independently in different PDB files.

Table 1. Genes presented in COSMIC Cancer-Gene-Census without mutation annotation.

Gene Symbol	Gene ID	Gene name
TRA	6955	T cell receptor alpha locus
TRB	6957	T cell receptor beta locus
TRD	6964	T cell receptor delta locus
IGH	3492	immunoglobulin heavy locus
IGK	50802	immunoglobulin kappa locus
IGL	3535	immunoglobulin lambda locus
HMGN2P46		high mobility group nucleosomal binding domain 2 pseudogene 46
MALAT1		metastasis associated lung adenocarcinoma transcript 1 (lnc-RNA; non-protein coding)

In order to gain insight into protein function, the COSMIC CGC gene was searched against the Pfam database using the Hidden Markov Model (HMM) method. 476 genes have more than one domain hit, indicating that most of the COSMIC CGC genes are multi-domain high order assembly. The most frequent domains presented in the (Figure 12B) indicate that many genes are associated with DNA binding. We have modelled 402 genes with structural coverage above 80%, roughly 60% of the cancer gene census proteins. Furthermore, we have modelled 119 transmembrane proteins. Most genes have IDRs less than 50% of the amino acid sequence. However, 71 genes are predicted to be more than 80% highly intrinsically disordered. The average MolProbity score for all modelled structures is ~3.2, with a

standard deviation of 0.64. The lowest MolProbity score is 0.51, whereas the highest MolProbity score is 4.58. (Figure 12)

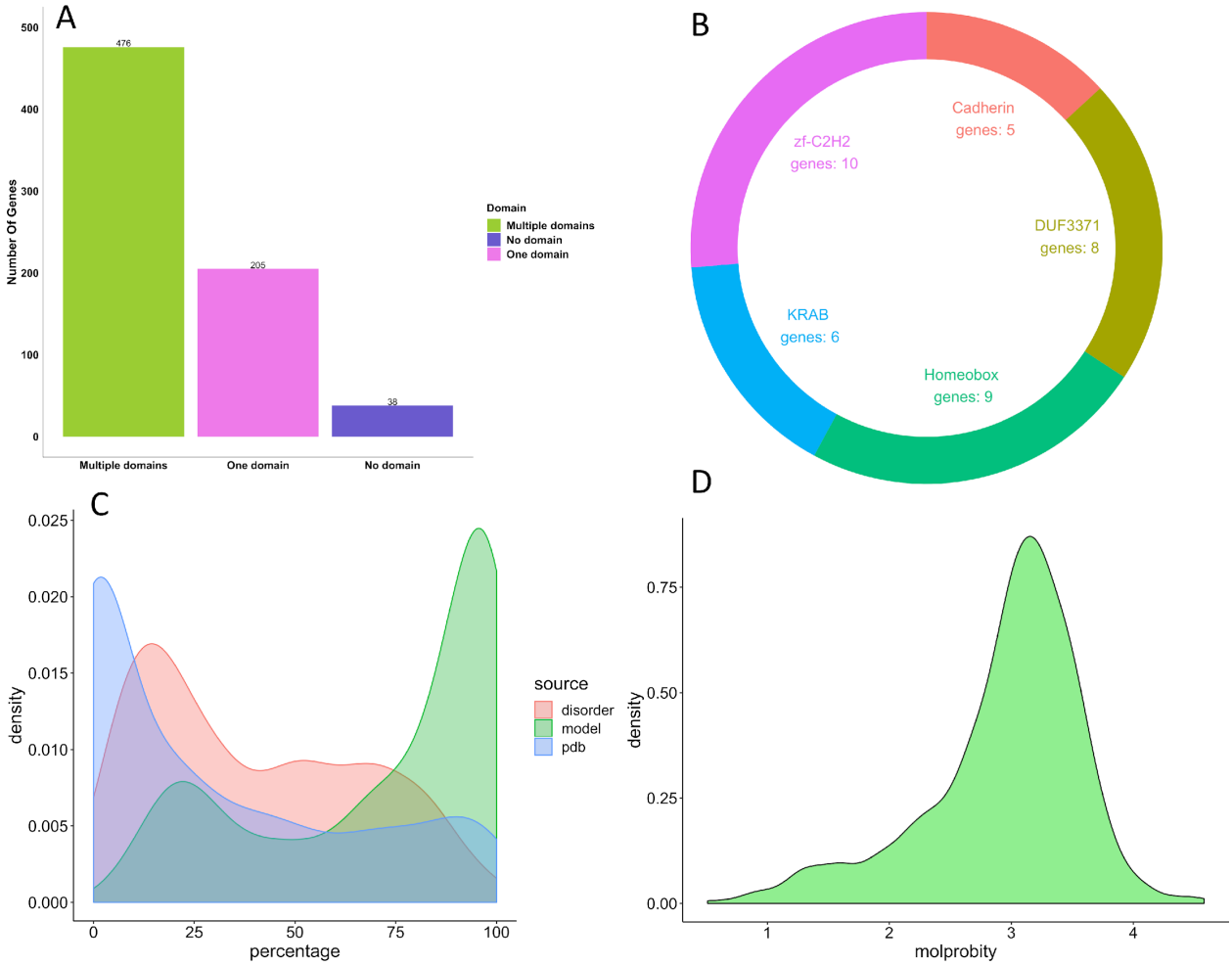
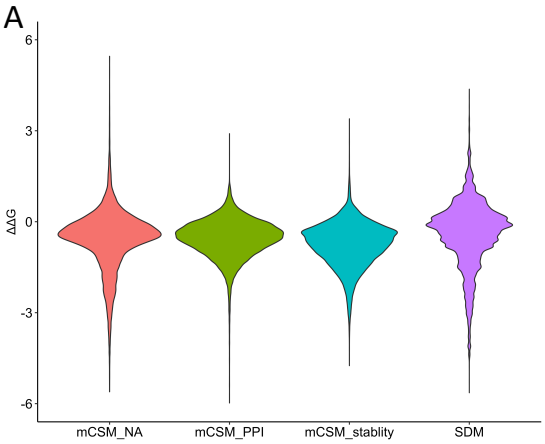
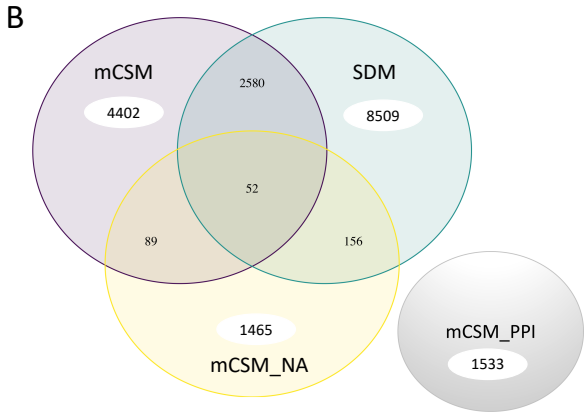


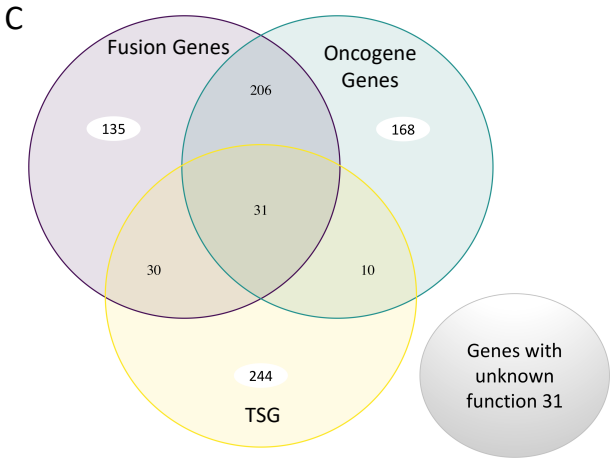
Figure 12. Structural analysis of genes presented in the COSMIC CGC 3D. (**A**) COSMIC CGC Pfam domain annotation; in purple, 38 genes show no domain annotations, whereas, in green and magenta, 681 genes show Pfam hits. (**B**) There are five domains annotated to be in multiple genes, of which three domains are associated with DNA binding. (**C**) Density plot for model coverage of the gene products: green colour represents modelled structures for gene products in COSMIC CGC that are close to 100% coverage of gene, whereas blue colour represents experimental structure coverage for genes in COSMIC CGC. The percentage of disordered regions predicted by DISOPRED is represented in red. (**D**) Density plot of the MolProbity score of all the models deposited in the CGC 3D.

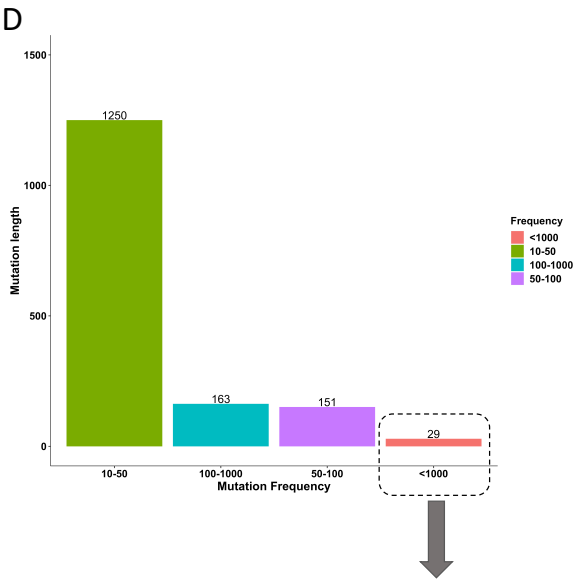
Understanding the impact of mutations on protein stability was the primary goal of modelling cancer gene census proteins. The most mutated gene with structural annotation is cellular tumour antigen p53 (TP53), whereas the least reported gene is chromosome 15 open reading frame 65 (C15orf65). Thus, 7123 mutations reported in the cancer gene census were predicted to be highly destabilising by the mCSM stability tool. In addition, there are 2632 mutations predicted to be highly destabilised by the SDM tool. Of these, only 2632 mutations were highly destabilised by SDM and mCSM tools. Furthermore, fewer mutations 1458 were predicted to destabilise the interface by mCSM-PPI, and 1710 mutations were predicted to destabilise the protein-DNA interface. However, the most frequent mutations are driver mutations, usually occurring in the active site region or allosterically. (Figure 13)

There are 715 modelled structures in high order assembly with ligand, DNA, Ribonucleic Acid (RNA), and metal ions. To illustrate, some of the gene products modelled as homo/heterodimers deposited into the COSMIC CGC 3D are further discussed below.


A
6
3
0
-3
-6
ΔΔG
mCSM_NA
mCSM_PPI
mCSM_stablity
SDM


B
mCSM
4402
2580
SDM
8509
52
89
156
1465
mCSM_NA
mCSM_PPI
1533


C
Fusion Genes
135
206
Oncogene Genes
168
31
30
10
244
TSG
Genes with unknown function 31


D
1500
1000
500
0
Mutation length
1250
163
151
29
10-50
100-1000
50-100
<1000
Mutation Frequency
Frequency
<1000
10-50
100-1000
50-100

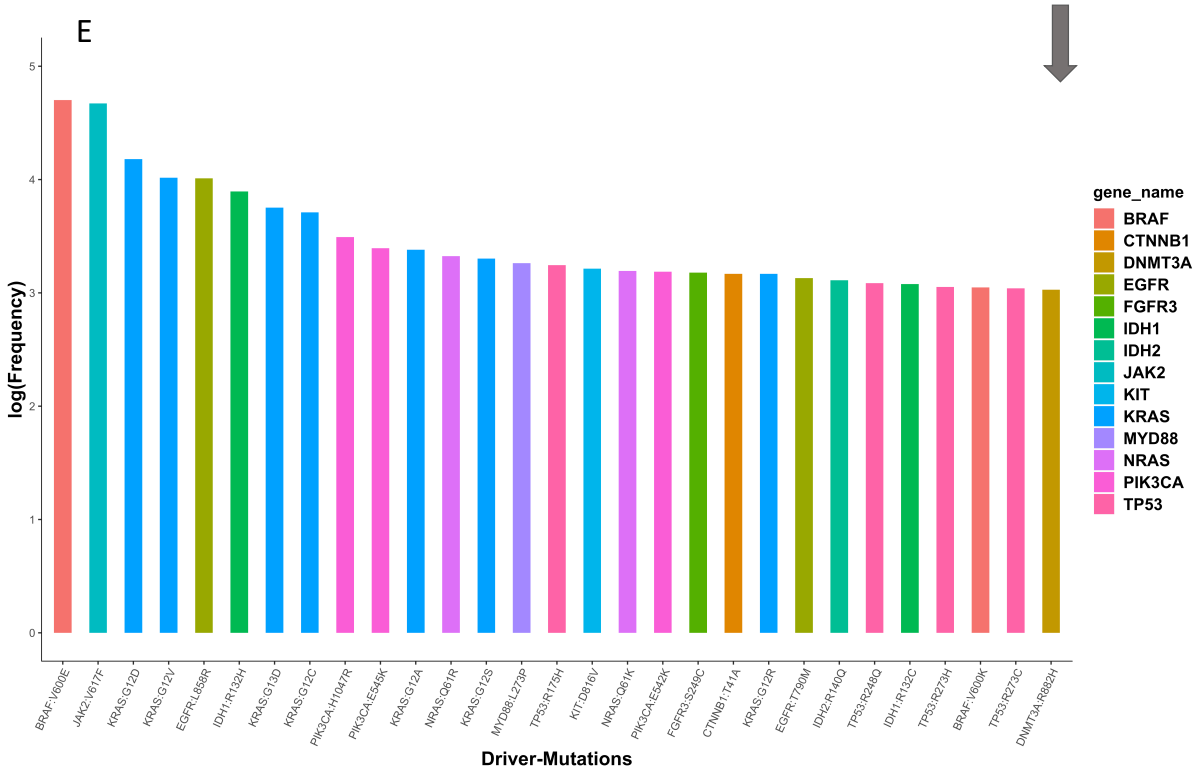

E
5
4
3
2
1
0
log(Frequency)
BRAF:V600E
JAK2:V617F
KRAS:G12D
KRAS:G12V
EGFR:L858R
IDH1:R132H
KRAS:G13D
KRAS:G12C
PIK3CA:H1047R
PIK3CA:E545K
KRAS:G12A
NRAS:Q61R
KRAS:G12S
MYD88:L273P
TP53:R175H
KIT:D816V
NRAS:Q61K
PIK3CA:E542K
FGFR3:S249C
CTNNB1:T41A
KRAS:G12R
EGFR:T790M
IDH2:R140Q
TP53:R248Q
IDH1:R132C
TP53:R273H
BRAF:V600K
TP53:R273C
DNMT3A:R882H
Driver-Mutations
gene_name
BRAF
CTNNB1
DNMT3A
EGFR
FGFR3
IDH1
IDH2
JAK2
KIT
KRAS
MYD88
NRAS
PIK3CA
TP53

Figure 13. Analysis of mutations with structural annotations deposited in the COSMIC CGC 3D database. (**A**) The distribution of impacts of mutations predicted by SDM and mCSM (nucleic acid, PPI and stability), the majority of the mutations are predicted to be natural, a small number of mutations are predicted to be highly destabilising ΔΔG value negative. Also, a small number of mutations are predicted to be stabilising ΔΔG positive. (**B**) The resemblance of highly destabilizing mutations between mCSM (stability, NA and PPI) and SDM tools, there was no resemblance between mCSM-PPI which predicted the impact of mutations and other tools. (**C**) The overlap annotations of fusion genes, oncogene genes or tumour-suppressor genes in the COSMIC CGC 3D, only 31 genes acted as Oncogene, fusion, and tumour suppressor genes (**D**). Frequencies of the most mutated residues reported in COSMIC CGC with a structural annotation in the COSMIC CGC 3D database 29 mutations reported more than 1000 times. (**E**) Frequencies (log) of the most frequently mutated residues have over 1000 occurrences (8 of these occur in KRAS, and four occur in TP53). Each gene is coloured differently.

COSMIC CGC 3D Modelling examples

Since all 714 genes were deposited into the COSMIC CGC 3D cancer gene census, genes are essential in cancer growing proliferation. All examples below were selected based on the issue faced during modelling the high order assembly of the cancer gene census. The androgen receptor homodimer modelled structure included an intrinsically disordered region linking two domains, DNA and natural ligand. The SDHC was selected as an example of heterodimer assembly. The KCNJ5 as homotrimer transmembrane transporter and PI3CB as heterodimer complex with inhibitor modelled into the active site.

Androgen receptor (AR)

AR is a nuclear hormone receptor that acts as a transcription factor. The binding of endogenous natural ligands stimulates the binding of AR to DNA through the DNA binding domain (DBD). AR is the primary drug target in prostate cancer. However,

after patients are exposed to abiraterone or enzalutamide, castration resistance develops. Castration-resistant prostate cancer (CRPC) is a form of advanced prostate cancer that no longer completely responds to treatments that lower testosterone. The AR included three main functional domains; the N-terminal domain, DBD, and ligand-binding domain, which is highly conserved. In addition, the DNA binding domain has zinc fingers that recognise and bind DNA so, facilitating AR regulation of gene expression (Tan *et al.*, 2015).

Including biological aspects into the final modelled structure is essential to see where mutations are located i.e in the DNA domain or the other region (Figure 14). The final homodimer AR modelled structure included all the three domains linked to each other. The DBD is built with DNA and zinc ions, and the natural occurring ligand 5-alpha-dihydrotestosterone (DHT) is incorporated into the ligand-binding domain. The templates used to build the final AR homodimer were PDB ID; 5JJM, for the ligand-binding domain, which covers the amino acid region between 668-920. In addition, PDB ID; 1R4I was selected for DBD, covering the region between 551-622. Finally, the hinge region linked the ligand-binding domain and DBD was built based on PDB ID; 5CJ6, 2AM9, covering the region between 623-667. The MolProbity score of the final modelled structure is 2.4, indicating that the modelled structure could be used in silico to assess the impact of mutations in the DNA-protein interface or around the ligand-binding site.

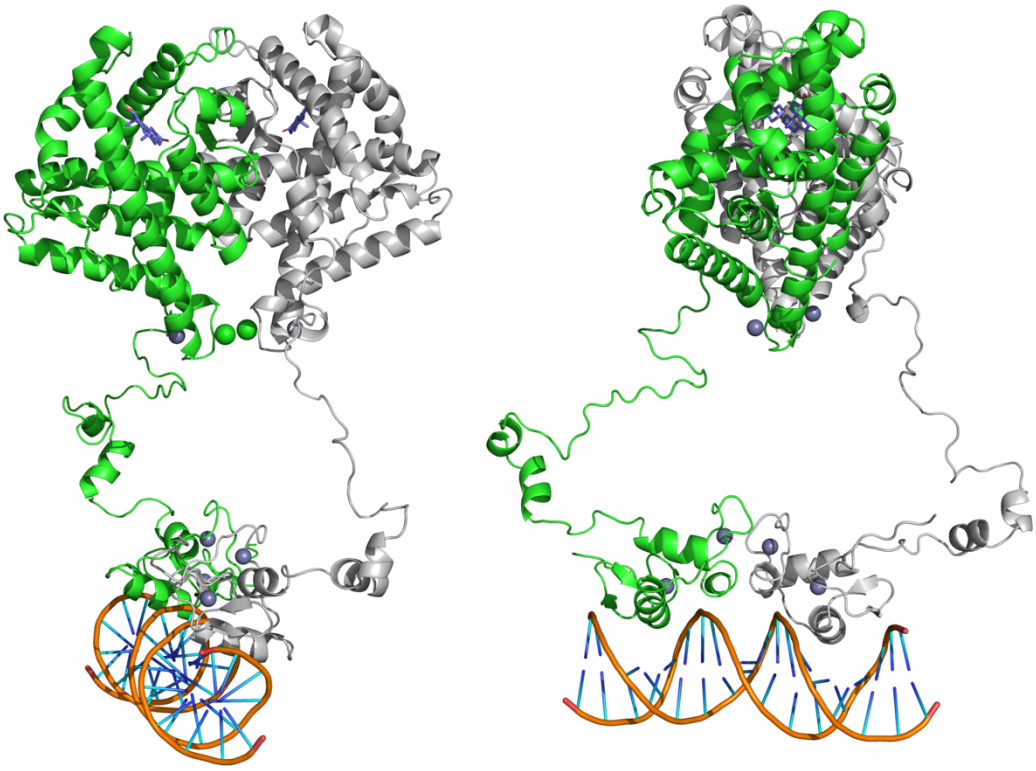


Figure 14. Modelled homodimer androgen receptor. Font and side view are represented with each protomer coloured differently, zinc ions represented in sphere dark-silver, whereas chloride ions represented in sphere green. The DNA is coloured in orange, and the DHT ligand is represented in a blue stick.

Succinate dehydrogenase cytochrome b560 subunit, mitochondrial (SDHC)

SDHC is a membrane anchoring subunit involved in the complex II of the mitochondrial electron transport chain. SDHC is involved in the tricarboxylic acid cycle part of carbohydrate metabolism. Complex II consists of four subunits: flavoprotein, iron-sulfur protein, SDHC and SDHD (Horsefield *et al.*, 2006). SDHC has 0% PDB coverage, and the final hetero 4-mer modelled structure included all complex II subunits with 81.6% PDB coverage. Biological insight into the final model is essential (Figure 15). Therefore, the final modelled structure incorporated the heme between the two transmembrane SDHC and SDHB subunits. One template was selected PDB ID; 1ZOY to build the final hetero 4-mer with the square root of

the average of squared errors (RMSD) 0.28, a MolProbity score of 3.13, and TM-align score=1). Tm-align ranges from 0-1 and measures the shape similarity between the proposed model and the native structure. A score of 1 indicates that the proposed model has folded very well onto the template, whereas a score of 0 or closer to 0 indicates that the template fold is unlikely to be the same.

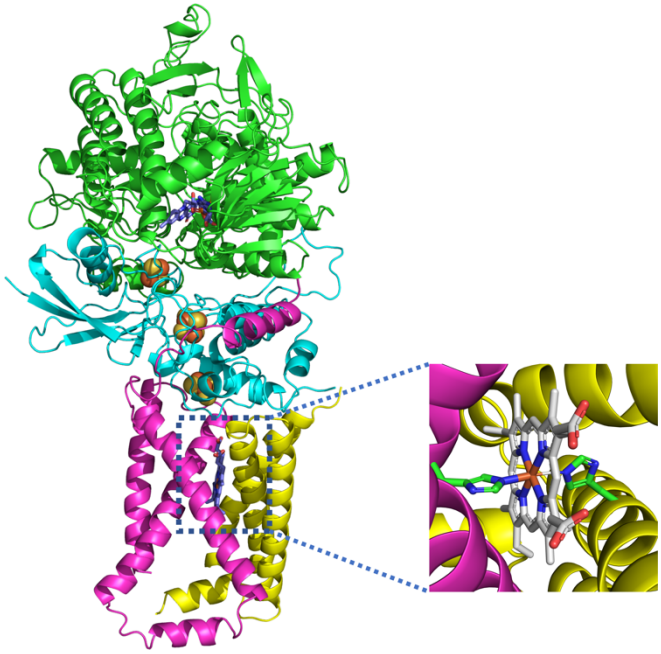
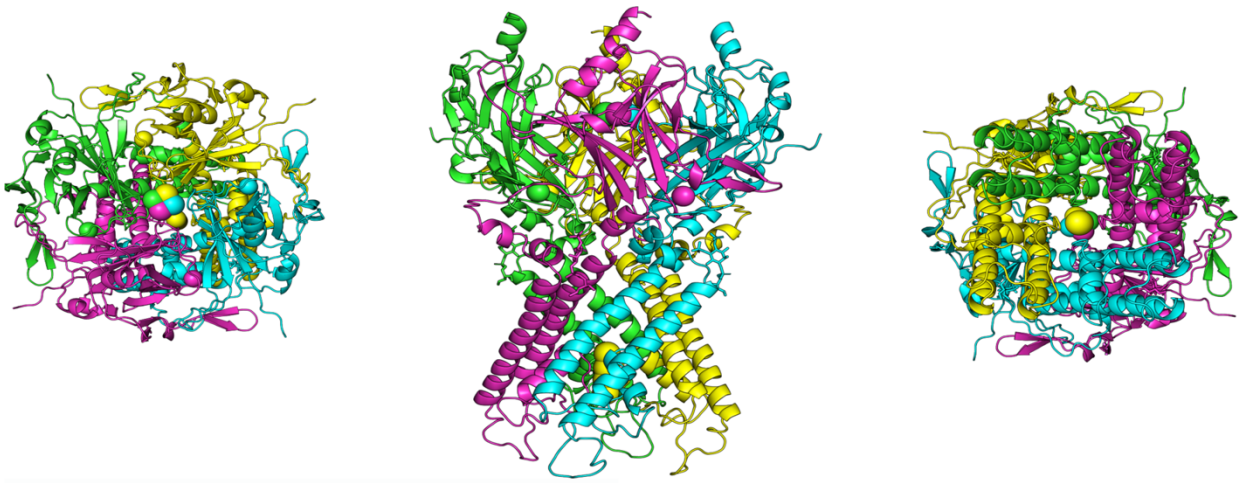


Figure 15. The heterotetrameric structure of SDHC. The heterotetramer consisted of FAD-binding protein (green), iron-sulphur protein (cyan), SDHC (magenta) and SDHD (yellow). The heme molecule, located between SDHC and SDHD, is shown in stick white. As illustrated in the insert, it is coordinated by histidines in green from SDHC and SDHD.

G protein-activated inward rectifier potassium channel 4 (KCNJ5)

KCNJ5 is an inward rectifier potassium channel controlled by G proteins. Inwardly rectified channel allows the flow of positive ions into the cell. In this case, potassium triggers the membrane potential back to resting potential. Mutations in the KCNJ5 result in the loss of the inward rectification and shift the selectivity toward sodium ions instead of potassium, increasing aldosterone production (Hattangady *et al.*, 2016). KCNJ5 has 0% PDB coverage. The final structure is modelled as a homo 4-

mer based on PDB ID; 3SYP, which can be helpful to locate the detrimental mutants i.e either on the pore of the channel or between the homodimer interface. The final RMSD and TM-score are 1.72, 0.9, respectively, and the MolProbity score is 2.7. the predicted interface between the homo 4-mer using PSIA shows multiple hydrogen bonding and salt bridge interactions where the large radius indicates the likelihood of finding an interface within the biological assembly (Figure 16). The final modelled structure included potassium ions, giving more biological insight into where potassium ions pass through the channel. The modelled structure could be used in silico for saturation mutational analysis as well as molecular docking.

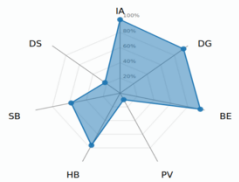


Interface A || C

Summary

	Monomer 1		Monomer 2	
Monomer ID	A		C	
Class	Protein		Protein	
Symmetry operation	X,Y,Z		X,Y,Z	
Symmetry ID	1_555		0_555	
Interface atoms	335	12.7%	329	12.5%
Surface atoms	1693	64.2%	1690	64.1%
Total atoms	2638	100.0%	2638	100.0%
Interface residues	87	26.8%	86	26.5%
Surface residues	318	97.8%	319	98.2%
Total residues	325	100.0%	325	100.0%
BSA, Å²	3139.6	15.1%	3278.8	15.8%
ASA, Å²	20761.9	100.0%	20737.0	100.0%
Solvation energy, kcal/mol		-288.8		-289.7
SE gain, kcal/mol		-16.4		-12.9

Interaction radar



Interface parameters

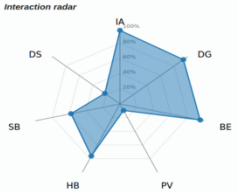
IA	Interface area, Å²	3209
DG	Solvation Energy, kcal/mol	-29.33
BE	Total Binding Energy, kcal/mol	-38.01
PV	Hydrophobic P-value	0.6538
HB	Number of Hydrogen Bonds	12
SB	Number of Salt Bridges	9
DS	Number of Disulphide Bonds	0

Interface A || D

Summary

	Monomer 1		Monomer 2	
Monomer ID	A		D	
Class	Protein		Protein	
Symmetry operation	X,Y,Z		X,Y,Z	
Symmetry ID	1_555		0_555	
Interface atoms	328	12.4%	335	12.7%
Surface atoms	1693	64.2%	1693	64.2%
Total atoms	2638	100.0%	2638	100.0%
Interface residues	85	26.2%	86	26.5%
Surface residues	318	97.8%	318	97.8%
Total residues	325	100.0%	325	100.0%
BSA, Å²	3282.6	15.8%	3133.8	15.1%
ASA, Å²	20761.9	100.0%	20753.8	100.0%
Solvation energy, kcal/mol		-288.8		-289.4
SE gain, kcal/mol		-13.0		-16.2

Interaction radar



Interface parameters

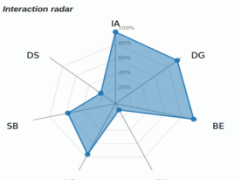
IA	Interface area, Å²	3208
DG	Solvation Energy, kcal/mol	-29.16
BE	Total Binding Energy, kcal/mol	-37.83
PV	Hydrophobic P-value	0.6593
HB	Number of Hydrogen Bonds	12
SB	Number of Salt Bridges	9
DS	Number of Disulphide Bonds	0

Interface B || C

Summary

	Monomer 1		Monomer 2	
Monomer ID	B		C	
Class	Protein		Protein	
Symmetry operation	X,Y,Z		X,Y,Z	
Symmetry ID	1_555		0_555	
Interface atoms	331	12.5%	331	12.5%
Surface atoms	1690	64.1%	1690	64.1%
Total atoms	2638	100.0%	2638	100.0%
Interface residues	85	26.2%	85	26.2%
Surface residues	318	97.8%	319	98.2%
Total residues	325	100.0%	325	100.0%
BSA, Å²	3283.3	15.8%	3128.9	15.1%
ASA, Å²	20774.0	100.0%	20737.0	100.0%
Solvation energy, kcal/mol		-289.1		-289.7
SE gain, kcal/mol		-13.1		-16.4

Interaction radar



Interface parameters

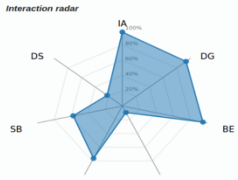
IA	Interface area, Å²	3206
DG	Solvation Energy, kcal/mol	-29.43
BE	Total Binding Energy, kcal/mol	-38.11
PV	Hydrophobic P-value	0.6466
HB	Number of Hydrogen Bonds	12
SB	Number of Salt Bridges	9
DS	Number of Disulphide Bonds	0

Interface B || D

Summary

	Monomer 1		Monomer 2	
Monomer ID	B		D	
Class	Protein		Protein	
Symmetry operation	X,Y,Z		X,Y,Z	
Symmetry ID	1_555		0_555	
Interface atoms	334	12.7%	332	12.6%
Surface atoms	1690	64.1%	1693	64.2%
Total atoms	2638	100.0%	2638	100.0%
Interface residues	87	26.8%	86	26.5%
Surface residues	318	97.8%	318	97.8%
Total residues	325	100.0%	325	100.0%
BSA, Å²	3142.3	15.1%	3283.3	15.8%
ASA, Å²	20774.0	100.0%	20753.8	100.0%
Solvation energy, kcal/mol		-289.1		-289.4
SE gain, kcal/mol		-16.5		-12.9

Interaction radar



Interface parameters

IA	Interface area, Å²	3213
DG	Solvation Energy, kcal/mol	-29.38
BE	Total Binding Energy, kcal/mol	-38.05
PV	Hydrophobic P-value	0.6553
HB	Number of Hydrogen Bonds	12
SB	Number of Salt Bridges	9
DS	Number of Disulphide Bonds	0

Figure 16. The homotetrameric structure of KCNJ5 with each protomer is coloured differently. The top and bottom views of the channel show the potassium ions inside the channel. (https://cancer-3d.com/model/KCNJ5). Plotting the homodimer interactions as radar plots, each edge represents different parameters such as interface area (IA) and hydrogen bonding (HB).

Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit beta isoform (PIK3CB)

PI3Ks play essential roles in several cellular processes, including cellular proliferation. PI3Ks generate a phosphatidylinositol 3,4,5-triphosphate (PIP3) from phosphatidylinositol 4,5-diphosphate using adenosine triphosphate (ATP) (Jia *et al.*, 2008). The final hetero 2-mer was built based on PDB ID; 2Y3A, with a MolProbity score of 3.23. Furthermore, the final modelled structure built with 2-(1H-indazol-4-yl)-6-{[4-(methylsulfonyl) piperazin-1-yl] methyl}-4-morpholin-4-yl-thieno[3,2-d] pyrimidine inhibitor as well as phosphatidylinositol 3-kinase regulatory subunit beta (Figure 17). The piperazine pharmacophore is very common for PIK3 isoforms (Miller, Thompson and Gabelli, 2019)(Zhang *et al.*, 2011), in which the indazole ring binds to Asp813, which derives the potency. However, selectivity is the main issue for the PIK3 isoform. Therefore, the finial heterodimer modelled structure will give more biological insight into where the new ligand should be designed to gain selectivity ligand.

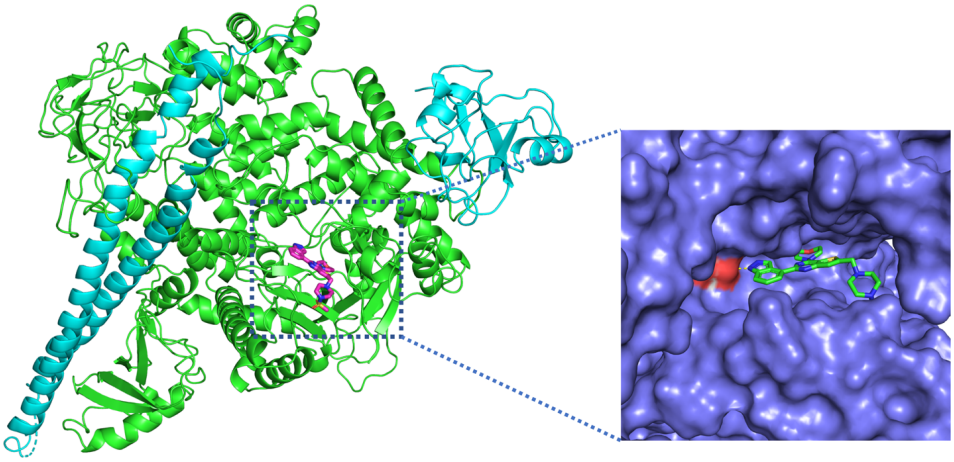


Figure 17. The heterodimeric structure of phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit beta-isoform is coloured in green with the 3-kinase regulatory subunit beta coloured in cyan. The 2-(1H-indazol-4-yl)-6-{[4-(methylsulfonyl) piperazin-1-yl] methyl}-4-morpholin-4-yl-thieno[3,2-d] pyrimidine represented in stick magentas. The surface representation in blue shows how the ligand fits into the active site with hydrogen bond interaction to Asp813 coloured in red.

3-ketodihydrosphingosine reductase (KDSR)

KDSR is an essential enzyme in synthesising sphingolipids, vital lipid components in membrane trafficking, apoptosis, cell proliferation and migration. In the de novo synthesis of sphingolipids, KDSR reduces 3-ketodihydrosphingosine to dihydrosphingosine. KDSR has a large cytoplasmic domain and two transmembrane helices. Two templates were selected PDB ID; 4NBU, 3P19, the MolProbity score of 3.32. The final model (Figure 18) was built with dihydro-nicotinamide-adenine-dinucleotide phosphate (NADPH) to give more biological insight into catalytic activity.

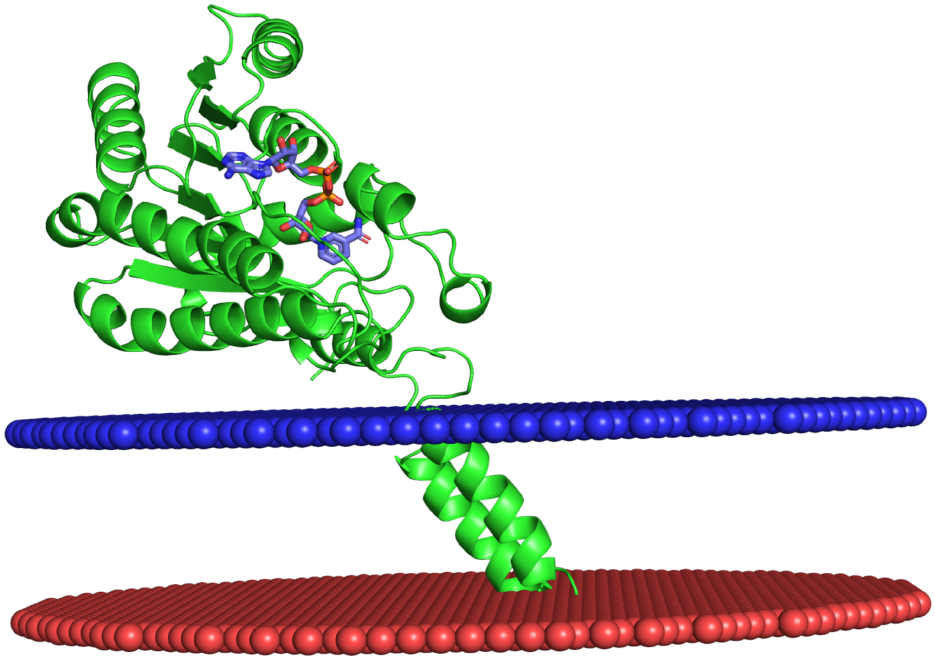


Figure 18. The protomer structure of 3-ketodihydrosphingosine reductase is coloured in green with NADPH ligand represented in a blue stick. The two transmembrane helices are highlighted between blue (extracellular) and red (intracellular) membrane surfaces, defined as circular structures.

Predicting the impact of Mutations

To study the impacts of mutations on stability, protein-protein interface and protein-DNA as predicted by SIFT (Ng and Henikoff, 2003), and PolyPhen2 (Adzhubei *et al.*, 2010) are coloured from light blue to dark blue. Interface sequence and structure-based methods were implemented. The sequence-based tools SIFT, and PolyPhen are already incorporated into the UniProt viewer in the section that shows variants from large scale studies, COSMIC, 1000-Genomes, and ClinVar (Pérez-Palma *et al.*, 2019). The mutations can be filtered based on those that are likely to cause disease coloured in red and those that are likely to be benign coloured in light green (supplementary-1). Missense mutations from COSMIC CGC are highlighted in the mutational table in the COSMIC CGC 3D database. The prediction results are based on SDM, mCSM-stability, mCSM-NA and mCSM-PPI. A value greater than +2

indicates mutations more likely to be stabilising and coloured in red, whereas mutations lower than -2 indicate mutations destabilising and coloured in dark blue. Most missense mutations reported in COSMIC have structural annotations in COSMIC CGC 3D. The majority of mutations shown are found between 10 and 50 times. However, 29 mutants were reported to be mutated more than 1000 times in 14 different genes. All these mutations are in hallmark genes except DNA (cytosine-5)-methyltransferase 3A (DNMT3A). These mutations are possible drivers and occur at different regions in the protein, for example, G12D, G12V in GTPase KRas (KRAS) located at the guanosine diphosphate (GDP) and guanosine triphosphate GTP binding site. In addition, DNMT3A (R882H) mutants occur at the protein-DNA interface. (Figure 13)

Interestingly, the most frequently reported mutation in the COSMIC CGC 3D has not been predicted to be highly stabilising or destabilising by structural-based tools mCSM and SDM. On the other hand, multiple infrequently mutations are predicted to be highly destabilised, affecting protein stability, protein-protein binding, and protein-DNA interaction. Detecting rare driver mutations tends to be very challenging but remains an essential task. However, it is not feasible to test all infrequent mutations to detect rare driver mutations. Therefore, newly putative rare driver mutations were hypothesised based on prediction from SDM and mCSM tools. However, this only prediction and experimental validations are needed (Figure 19).

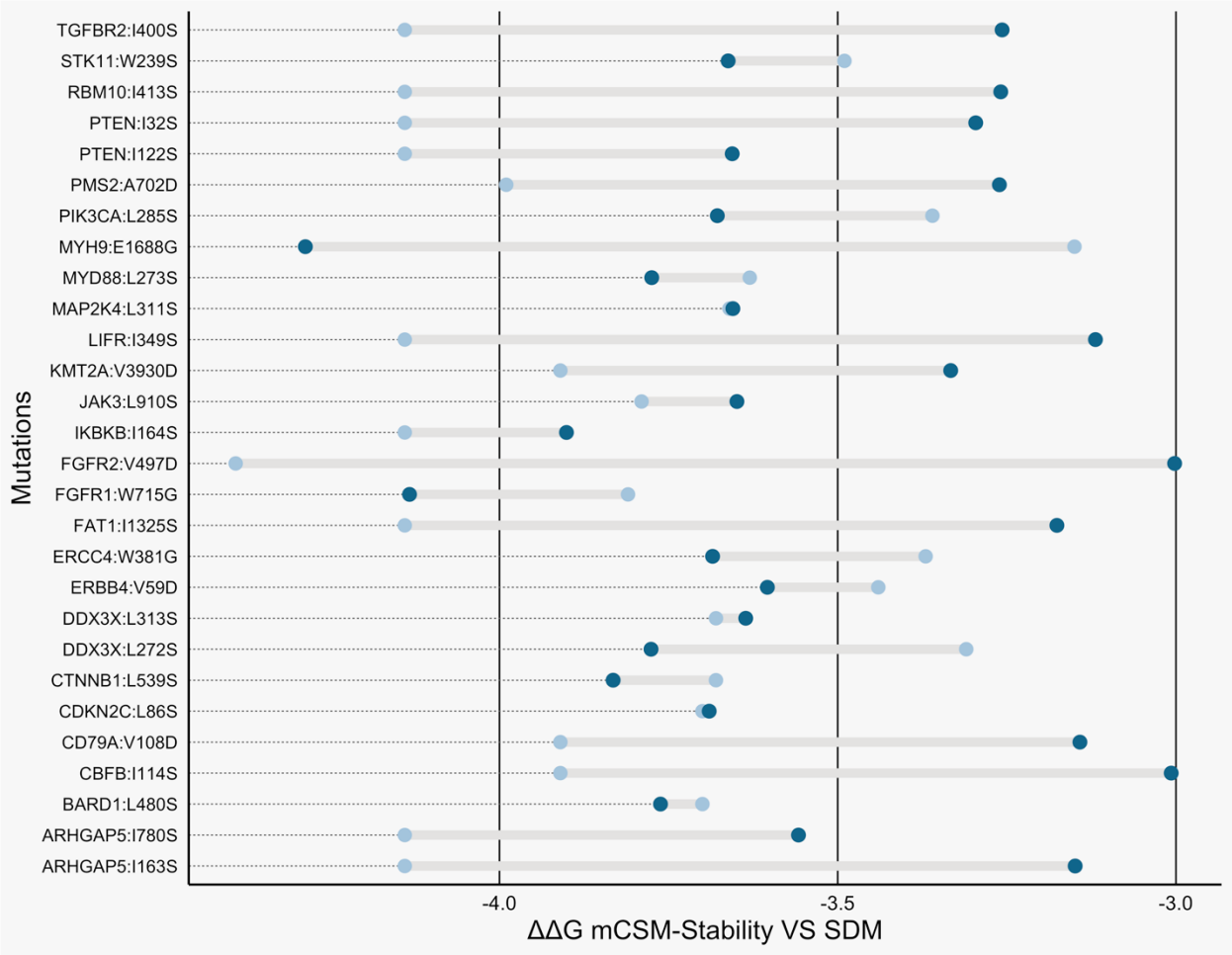


Figure 19. Prediction of rare mutations. On the y-axis is the gene name, followed by residue wild-mutant residue. Whereas in the x-axis predicted value from mCSM and SDM stabilities. The SDM value is coloured in circler light blue, and mCSM is coloured in dark blue.

Single infrequent driver mutation is rare to be driver since the key definition of the driver mutant is to be able to give a selective advantage to cancer cells. Protein is not a fixed structure, as visualized in the built models, but rather exists in conformational ensembles between the active and inactive states. Therefore, rare driver mutations combined with other mutations likely shift the protein conformation equilibrium toward an active state for oncogene and an inactive state for a tumour suppressor. To understand rare driver mutation on the stability, we need to consider allosteric mutations since it perturbs protein structure by breaking/forming or

weakening interactions, enhancing the relative stability toward inactive states and affecting ligands, ions lipid, and post-translational modification and more.

Examples of mutations in cancer

There is a massive number of reported mutations in the COSMIC CGC 3D. The selected mutation examples are based on the impact of mutations on protein stability and protein-protein interface and protein-DNA interface. (Figures 19, 22, 24)

Fibroblast growth factor receptor 1 (FGFR1)

FGFR1 is a tyrosine-protein kinase transmembrane protein essential in cell proliferation and migration and involves phosphorylation of multiple downstream signalling molecules, including RAS, AKT1, and MAPK1/ERK1 (Sarabipour and Hristova, 2016). The complete modelled structure has been solved (Figure 20A). The FGFR1 W715G mutant is reported only once in COSMIC CGC and is predicted to be highly destabilising by SDM and mCSM stability. It appears in the C-lobe region of the kinase between the active site and the homodimer interface. The helix is directly linked to the activation loop and interface region (Figure 20 B/C). Therefore, the hypothesis is that the destabilisation of the active state happens allosterically by affecting the activation loop. However, in most mutations cases, a single rare mutation on its own cannot shift the equilibrium toward active or inactive states. As seen in the Figure, mutations appear in many regions. The cooperation of spatially proximal mutations is the key to shifting the equilibration toward active and inactive states.

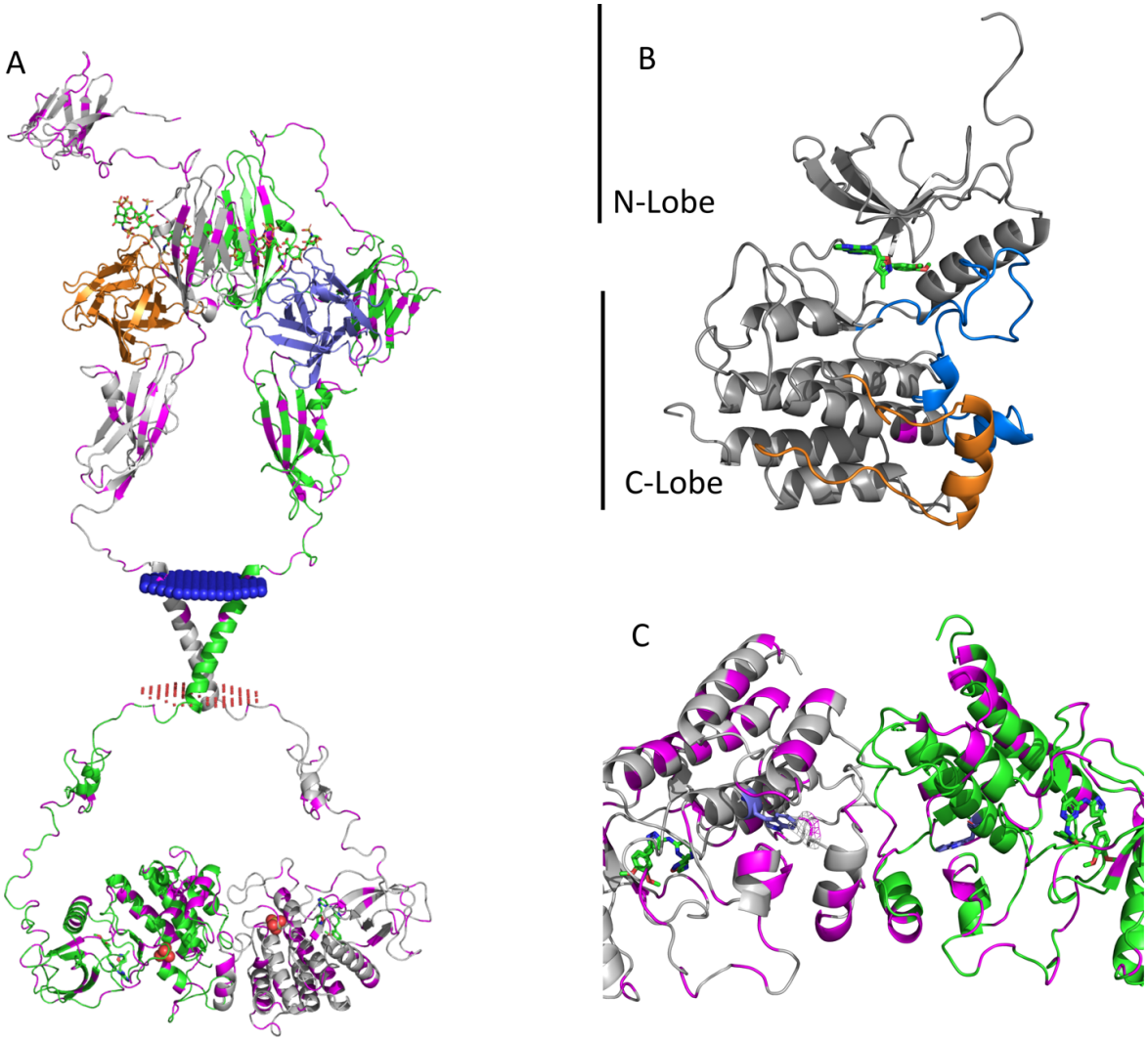


Figure 20. (A) modelled hetero 4-mer, fibroblast growth factor receptor 1 coloured in white and green, the ligands coloured in blue and orange. The oligosaccharides showed in stick green. The magenta colour represents reported mutations for FGFR1. (B) The blue loop in the kinase domain is the activation loop, whereas the orange is the interface region between the two dimers, and the pink spot region represents the W715G mutant. (C) the highlight of the W715G mutant as a blue stick, which occurs between the active site and the interface region, and other mutations in magenta.

Glutamate receptor ionotropic, NMDA 2A (GRIN2A)

The GRIN2A is a ligand-gated ion channel with high permeability to calcium ions. It functions as a heterodimer that mediates excitatory synaptic transmission essential for brain function (Pierson *et al.*, 2014). The GRIN2A has four domains, the

extracellular amino-terminal domain (ATD), the extracellular ligand-binding domain, the transmembrane region, and the intracellular carboxyl-terminal domain. The most frequent mutation is D114N, located in the ATD domain. The highly destabilising mutant predicted to affect the heterodimerization of GRIN2A is Q811L occurring in the peptide linker between LBD and the transmembrane. The destabilizing predicted value by mCSM-PPI is -4.6 (Figure 21).

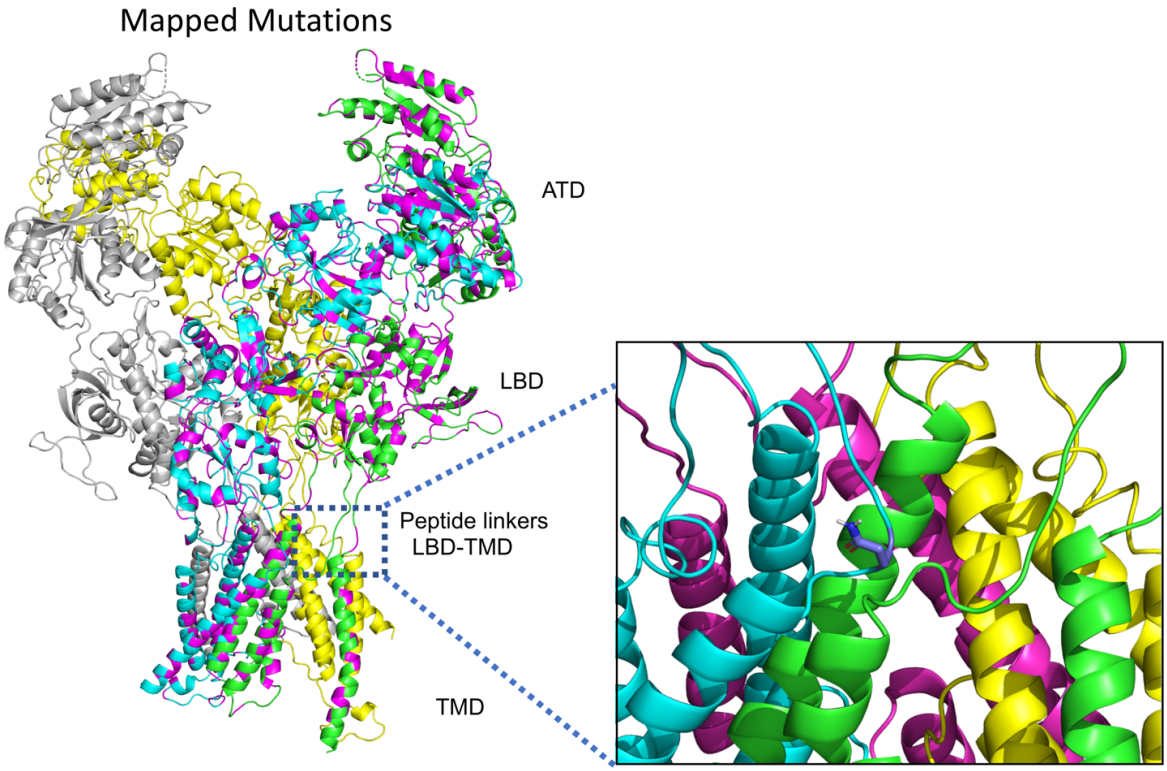


Figure 21. Mutations mapped to the modelled glutamate receptor ionotropic (NMDA 2A) hetero 4-mer. The glutamate receptor ionotropic is represented in green and cyan. Whereas Glutamate receptor ionotropic, NMDA 1, in yellow and white. The D114N mutant occurs in the peptide linker region represented in a blue stick.

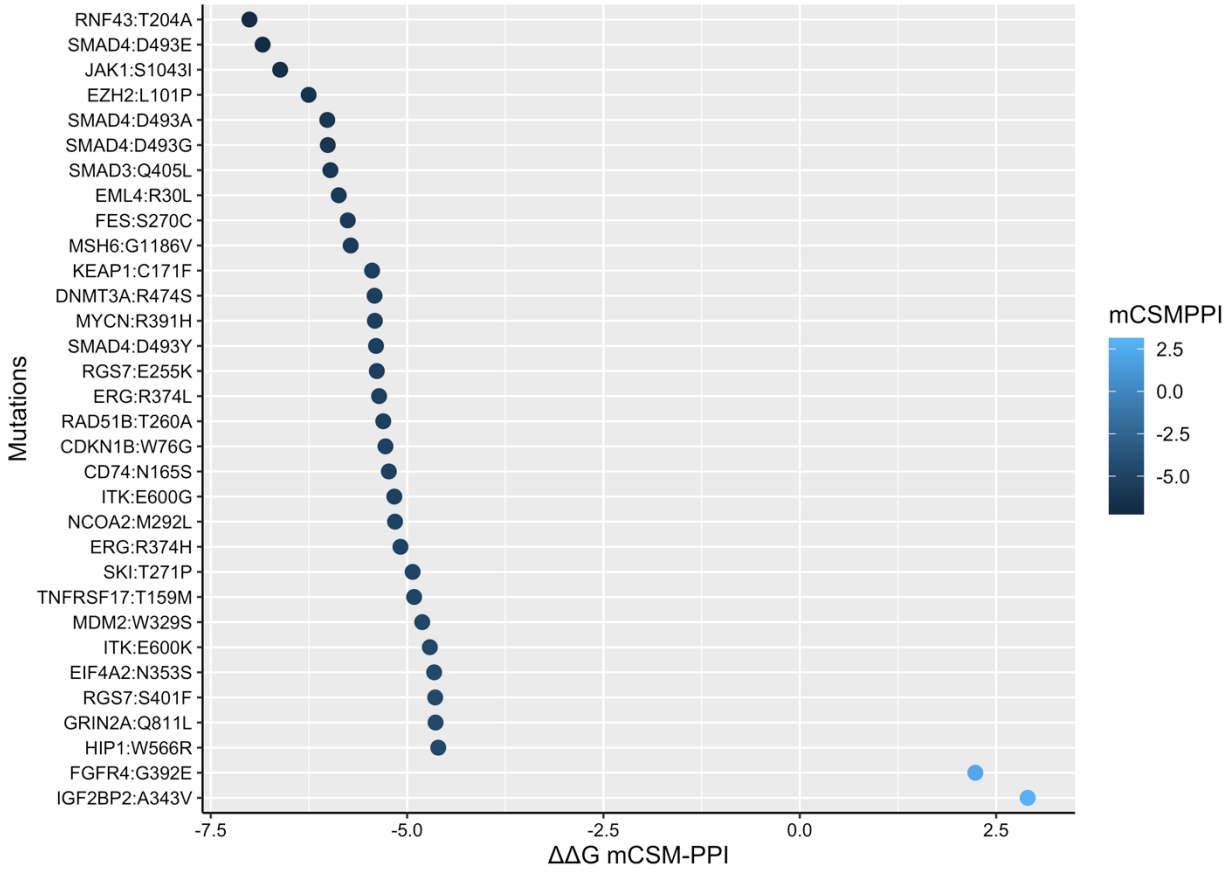


Figure 22. Prediction of rare mutations at the protein-protein interface predicted by mCSM-PPI. On the y-axis is the gene name, followed by residue number and type in wild and mutant. On the x-axis is the predicted ΔΔG value from mCSM-PPI. The highly destabilising mutations are coloured in dark blue, whereas highly stabilising mutations are coloured in light blue.

Forkhead box protein O1 (FOXO1)

Forkhead box protein O1 (FOXO1), a transcriptional factor that regulates metabolic homeostasis, plays an essential role in insulin signalling (Rajan *et al.*, 2016). It contains a distinct type of DNA-binding region (fork-head). The structure was solved

experimentally as a heterodimer with protein C-ets-1 (ETS1). There are 32 reported mutations in the fork-head DNA binding domain. The most frequent is R213C, whereas the mutant with the greatest impact on protein-DNA interaction is S218F (Figure 23)

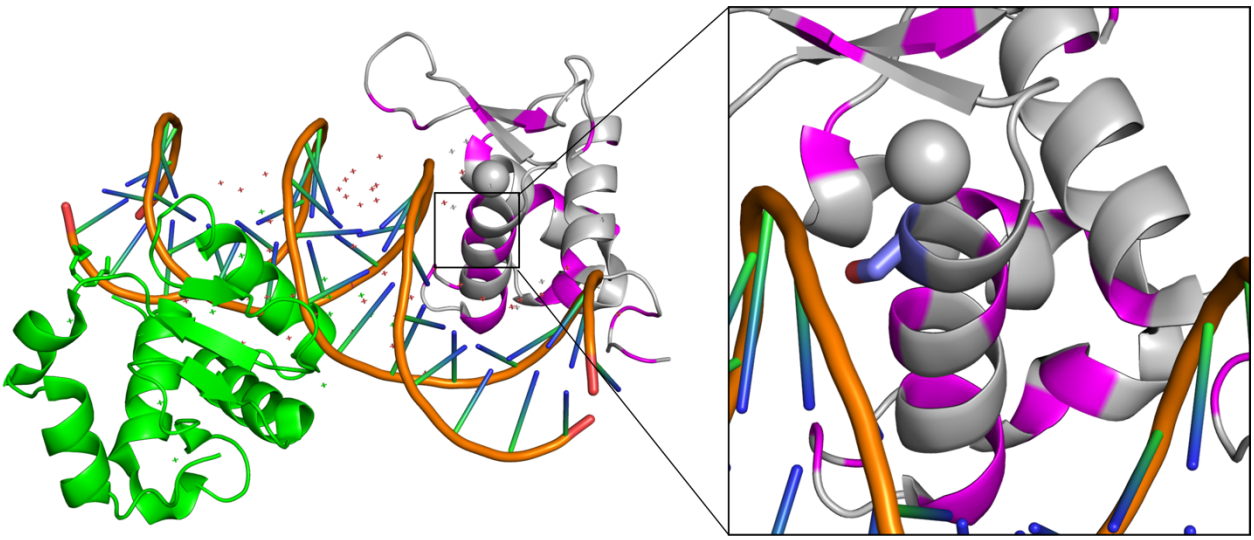


Figure 23. mapped mutations for experimentally solved forkhead box protein O1 structure (PDB ID; 4LG0) in white and protein C-ets-1 in green. The S218F is represented in a blue stick, and the DNA is coloured in orange.

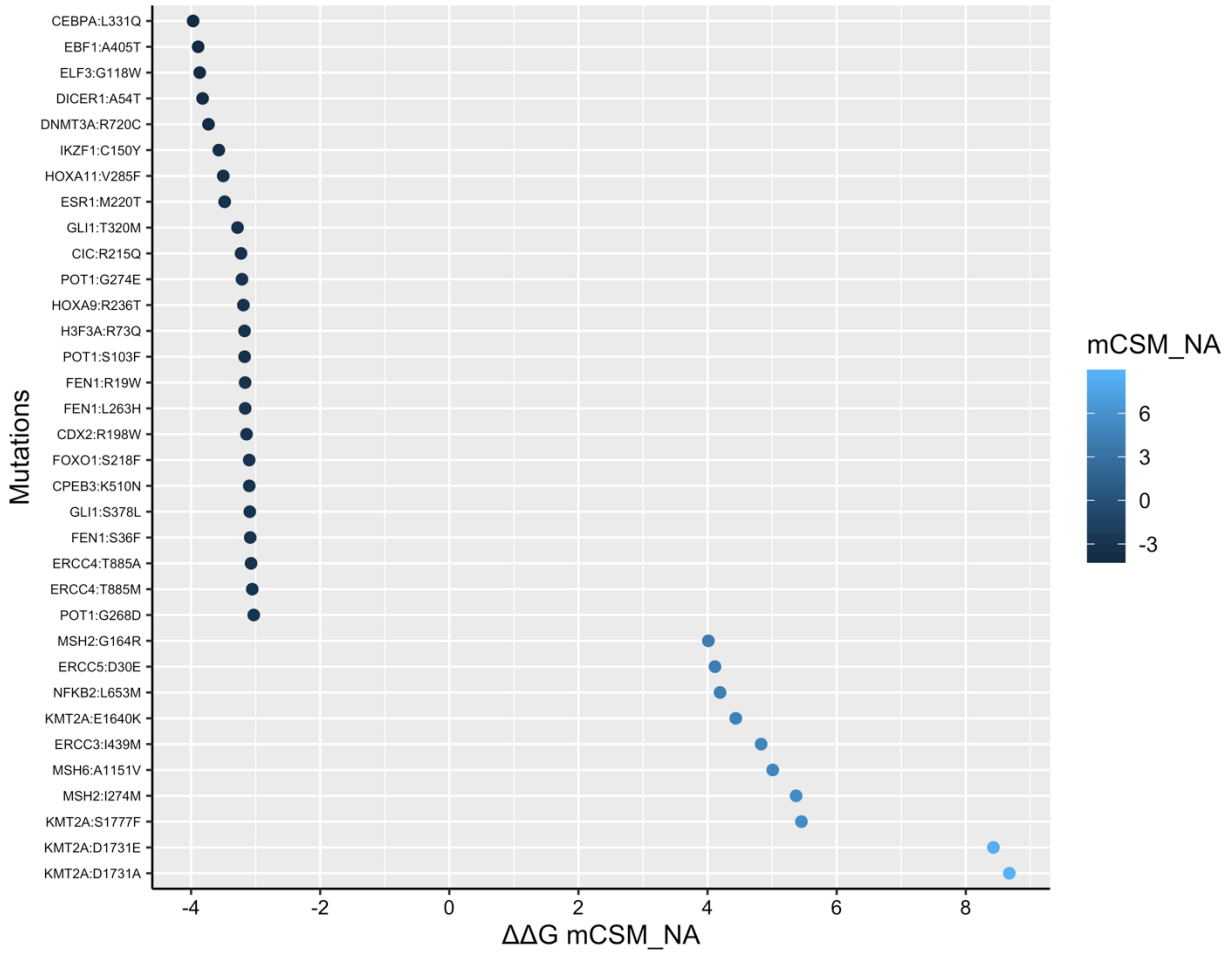


Figure 24. Prediction using mCSM-DNA of the impact of rare mutations at the protein-DNA interface. On the y-axis is the gene name, followed by wild type residue, number and mutant residue and the x-axis is the predicted value from mCSM-DNA. The highly destabilising mutations are coloured in dark blue, whereas highly stabilising mutations are coloured in light blue.

Comparing COSMIC CGC 3D models to Alphafold models.

The machine learning revolution has led to the development of the software AlphaFold, a new algorithm that predicts protein structure from sequence (Jumper *et al.*, 2021). The challenge of predicting protein fold has been in the field for a long time. AlphaFold works in a two-step process. The first step includes a convolutional neural network that takes an amino acid residues sequence as an input. Multiple features, such as multiple sequence alignment of evolutionarily related sequences, are included. The neural network's output is the confident distribution distance between the two amino acids in the 3D structure of the protein. After obtaining the final distance matrix, the second step is the gradient descent optimization, which folds the 3D structure to match the distance between the amino acid residues specified by the distance matrix.

We may then ask what is the difference between the COSMIC CGC 3D and AlphaFold models? There is some similarity but a huge difference between the modelled structures in the COSMIC CGC 3D and the recently developed AlphaFold database (https://alphafold.ebi.ac.uk/). The similarity tends to be in the predicted structure of the protomer, for example, LHFPL6 tetraspan subfamily member 6 protein (LHFPL6) (Figure 25), and sodium/potassium-transporting ATPase subunit alpha-1 (ATP1A1).

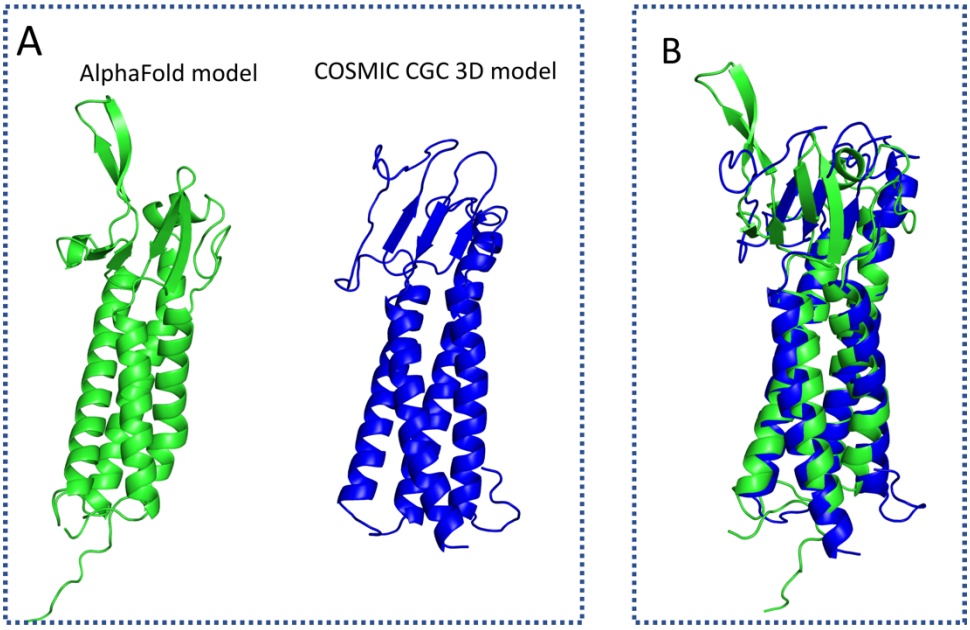


Figure 25. (A) Compression of models predicted by AlphaFold and COSMIC CGC 3D. LHFPL6 AlphaFold model in green and COSMIC CGC 3D model in blue. (B) A superimposition of the two models. They are very similar in the helical region, although slightly different in the beta-sheet region.

However, AlphaFold predicts protomer structures only, and several important biological features were missing. For examples:

I. There is no consideration of high order assembly homo/heterodimers. (Figure 26 C/D)
II. Intrinsically disordered regions are not defined, but models are built with colossal loops. (Figure 26 A/B)
III. Transmembrane regions are not well defined, i.e. there is no distinction between cytoplasmic and extracellular regions. (Figure 26 C/D)
IV. The model is missing ligands and metal ions. (Figure 26 G/H)
V. The modelled DNA/RNA protein/domains are missing the DNA/RNA component. (Figure 26 E/F)

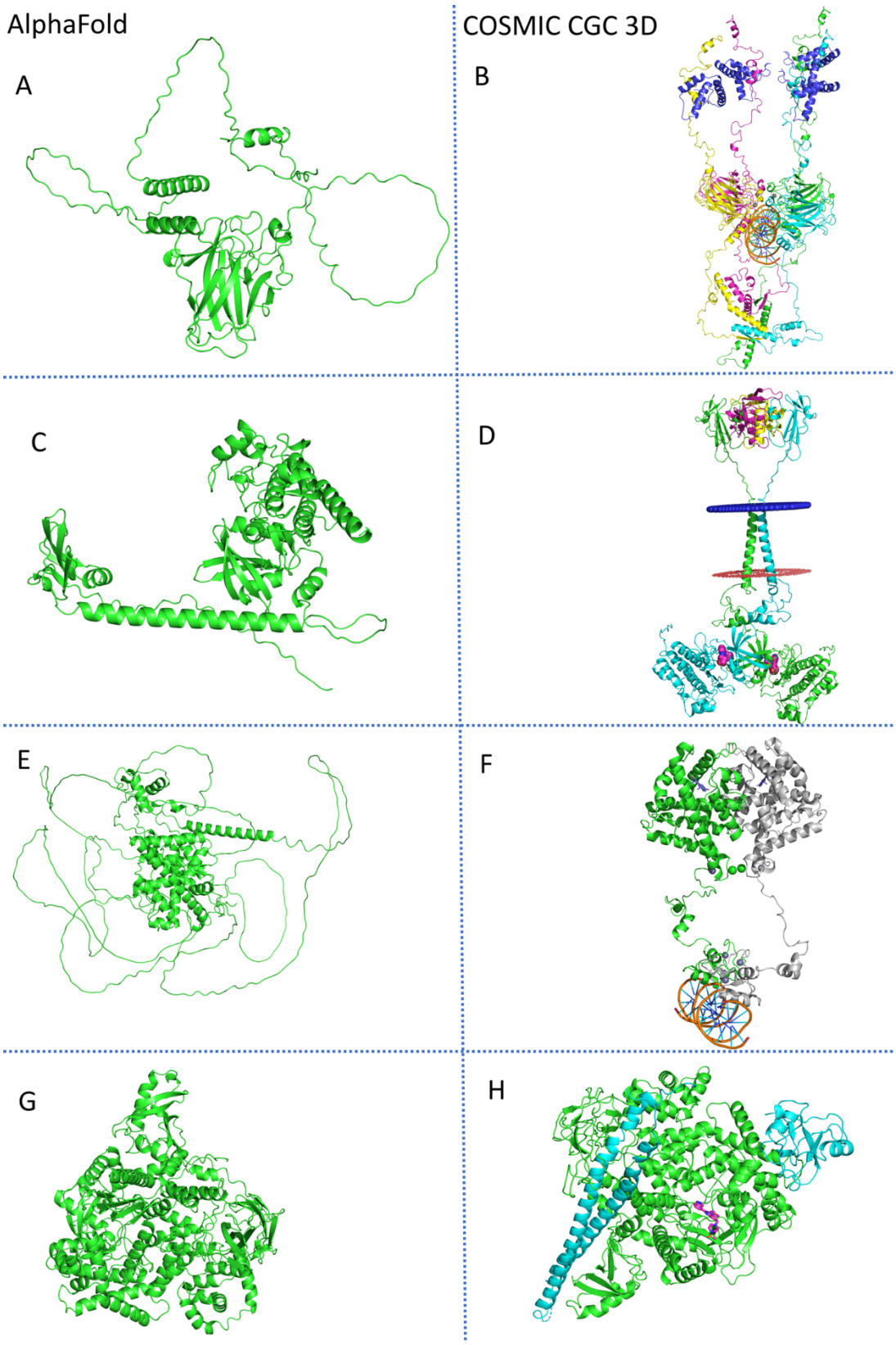


Figure 26. Comparison of AlphaFold models on the left and COSMIC CGC 3D models on the right. AlphaFold only modelled protomer structures and missing many biologically-relevant such as higher-order assembly DNA etc. (A/B) models of TP53. (C/D) models of Activin receptor type-2A. (E/F) models of Androgen receptor. (G/H) models of PIK3CB.

Discussion

The number of proteomic and genomic databases has increased dramatically in the past few years, annotating gene sequences and functions, domains and mutations. Examples are UniProt, Pfam, COSMIC and PDB. However, these databases are limited to experimentally solved structures. Genome3D (Sillitoe *et al.*, 2020) and ModBase (Pieper *et al.*, 2014) were developed to integrate models built by different groups. However, these databases are limited to small protomer structures. Furthermore, the AlphaFold database includes models for 20 organisms such as *E.coli, Human*, and *M.tuberculosis* etc. The COSMIC CGC 3D database includes models for the most challenging genes known as hallmarks genes. These gene products are modelled as higher-order assemblies, including homo and hetero-oligomeric complexes. The understanding of large structural assemblies has been increased recently by cryo-EM. However, the challenge remains in interpreting the disordered regions between domains and structures of small gene products that cryo-EM cannot resolve very well. Nevertheless, some models of these complex systems have been built with the approaches described here, such as cGMP-specific phosphodiesterase 6 (PDE6) (Maryam *et al.*, 2019) and SARS CoV2 3D proteome (Ali F. Alsulami *et al.*, 2021), which will be discussed in detail in chapter 3.

In order to predict the impact of mutations using a structural based method, 3D structures are required either by experimental approaches or by comparative modelling. Therefore, obtaining a 3D model with a reasonable quality assessment is important. However, this depends on the availability of solved templates. To overcome or reduce this issue, computational energy minimization of the final

modelled structure tends to be essential to remove side-chain clashing. Modelled structures also can be used in silico virtual screening to suggest new ligands.

Almost all the COSMIC CGC 3D gene products are modelled with intrinsically disordered regions between two domains or within the domain itself. However, the IDPRs lack a stable 3D structure and cannot be solved experimentally by X-ray or cryo-EM. Unfortunately, these regions tend to be considered non-functional regions and are omitted even in comparative modelling. The NMR conformational ensembling is one of the ways to represent IDPRs, whereas in COSMIC CGC 3D, we built them as a loop, and Foldit was used to remove clashes generated.

There remain many challenges in the modelling of wild-type proteins. These include:

I. The model's accuracy depends on the sequence similarity of the modelled protein to that of the target. Where the similarity is low, loops may differ, and residues may be inserted or deleted in the homologues to be modelled, leading to less accurate models.
II. The oligomeric state can vary between orthologues as well as paralogues. This makes automated modelling particularly challenging.
III. Conformations may differ according to a functional state, whether the proteins are enzymes (apoenzymes, holoenzymes or intermediate substrate-enzyme complexes), receptors or other regulatory proteins.
IV. The majority of protein structures of the human CGC have disordered regions, making the assembly of the domains very challenging.

The COSMIC CGC 3D database is the only database addressing all the points mentioned above. 127,443 mutations from COSMIC CGC have structural representation in the COSMIC CGC 3D. These mutations represent 71% of reported mutations in the COSMIC CGC. I will continue updating the modelled structures in the COSMIC CGC 3D by including a new curated gene from COSMIC CGC in the near future.

Conclusion

COSMIC CGC 3D is a well-annotated database, including links to external resources. It has been developed to increase the 3D structural coverage for the essential genes in cancer. The recent massive increase of reported mutational data makes it necessary to rely on comparative modelling and other computational tools to predict the impact of these mutations in high throughput fashion. Insertion of intrinsically disordered regions between or within globular domains is very experimentally challenging. Mapping curated mutations to modelled structures and exploiting novel methods such as SDM and mCSM will undoubtedly increase our knowledge of which mutations will impact the stability and interface. This will allow drug discovery to develop new ligands that interact with the new mutant form to combat emerging drug resistance. The COSMIC CGC 3D v2 will include other developed software that predicts the mutation's impact on ligand binding and stability, such as mCSM-Lig, MAESTRO, STRUM, and FOLD-X.

References

Adzhubei, I. A. *et al.* (2010) ‘A method and server for predicting damaging missense mutations’, *Nature Methods*, 7(4), pp. 248–249.

Alsulami, Ali F *et al.* (2021) ‘COSMIC Cancer Gene Census 3D database: understanding the impacts of mutations on cancer targets’, *Briefings in Bioinformatics*, 22(6), pp. 1–11.

Alsulami, Ali F. *et al.* (2021) ‘SARS-CoV-2 3D database: Understanding the coronavirus proteome and evaluating possible drug targets’, *Briefings in Bioinformatics*, 22(2), pp. 769–780.

Andreeva, A. *et al.* (2014) ‘SCOP2 prototype: A new approach to protein structure mining’, *Nucleic Acids Research*, 42(D1), pp. 310–314.

Bailey, M. H. *et al.* (2018) ‘Comprehensive Characterization of Cancer Driver Genes and Mutations’, *Cell*, 173(2), pp. 371-385.e18.

Bateman, A. *et al.* (2021) ‘UniProt: The universal protein knowledgebase in 2021’, *Nucleic Acids Research*, 49(D1), pp. D480–D489.

Bell, D. *et al.* (2011) ‘Integrated genomic analyses of ovarian carcinoma’, *Nature*, 474(7353), pp. 609–615.

Berman, H. M. *et al.* (2000) ‘The Protein Data Bank’, *Nucleic Acids Research*, 28(1), pp. 235–242.

Bray, F. *et al.* (2018) ‘Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries’, *CA: A Cancer Journal for Clinicians*, 68(6), pp. 394–424.

Bunting, S. F. and Nussenzweig, A. (2013) ‘End-joining, translocations and cancer’, *Nature Reviews Cancer*, 13(7), pp. 443–454.

Campbell, P. J. *et al.* (2020) ‘Pan-cancer analysis of whole genomes’, *Nature*, 578(7793), pp. 82–93.

Carracedo, A. and Pandolfi, P. P. (2008) ‘The PTEN-PI3K pathway: Of feedbacks and cross-talks’, *Oncogene*, 27(41), pp. 5527–5541.

Chen, V. B. *et al.* (2010) ‘MolProbity: All-atom structure validation for macromolecular crystallography’, *Biological Crystallography*, 66(1), pp. 12–21.

Ciriello, G. *et al.* (2013) ‘Emerging landscape of oncogenic signatures across human cancers’, *Nature Genetics*, 45(10), pp. 1127–1133.

Darwin, C. (1859) *ON THE ORIGIN OF SPECIES*.

Dehouck, Y. *et al.* (2011) ‘PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality’, *BMC Bioinformatics*, 12(151), pp. 1–12.

Finn, R. D. *et al.* (2014) ‘Pfam: The protein families database’, *Nucleic Acids Research*, 42(D1), pp. 222–230.

Franco, I. *et al.* (2019) ‘Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type’, *Genome Biology*, 20(1), pp. 1–22.

Fröhling, S. *et al.* (2007) ‘Identification of Driver and Passenger Mutations of FLT3 by High-Throughput DNA Sequence Analysis and Functional Assessment of Candidate Alleles’, *Cancer Cell*, 12(6), pp. 501–513.

Futreal, P. A. *et al.* (2004) ‘A census of human cancer genes’, *Nature Reviews Cancer*, 4(3), pp. 177–183.

Gnad, F. *et al.* (2013) ‘Assessment of computational methods for predicting the effects of missense mutations in human cancers’, *BMC genomics*, 14(Suppl 3), pp. 1-13.

Gonzalez-Perez, A., Jene-Sanz, A. and Lopez-Bigas, N. (2013) ‘The mutational landscape of chromatin regulatory factors across 4,623 tumor samples’, *Genome Biology*, 14(9), pp. 1–15.

Hanahan, D. and Weinberg, R. A. (2000) 'The Hallmarks of Cancer', *Cell*, 100(7), pp. 57–70.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: The next generation', *Cell*, 144(5), pp. 646–674.

Hattangady, N. G. *et al.* (2016) 'Mutated KCNJ5 activates the acute and chronic regulatory steps in aldosterone production', *Journal of Molecular Endocrinology*, 57(1), pp. 1–11.

Hodgkinson, A., Chen, Y. and Eyre-Walker, A. (2012) 'The large-scale distribution of somatic mutations in cancer genomes', *Human Mutation*, 33(1), pp. 136–143.

Hodis, E. *et al.* (2012) 'A landscape of driver mutations in melanoma', *Cell*, 150(2), pp. 251–263.

Horsefield, R. *et al.* (2006) 'Structural and computational analysis of the quinone-binding site of complex II (succinate-ubiquinone oxidoreductase): A mechanism of electron transfer and proton conduction during ubiquinone reduction', *Journal of Biological Chemistry*, 281(11), pp. 7309–7316.

Hua, X. *et al.* (2013) 'DrGaP: A powerful tool for identifying driver genes and pathways in cancer sequencing studies', *American Journal of Human Genetics*, 93(3), pp. 439–451.

Jacks, T. *et al.* (1994) 'Tumor spectrum analysis in p53-mutant mice', *Current Biology*, 4(1), pp. 1–7.

Jia, S. *et al.* (2008) 'Kinase-dependent and -independent functions of the p110β phosphoinositide-3-kinase in cell growth, metabolic regulation and oncogenic transformation', *Nature*, 454(7205), pp. 776–779.

Jones, D. T. and Cozzetto, D. (2015) 'DISOPRED3: Precise disordered region predictions with annotated protein-binding activity', *Bioinformatics*, 31(6), pp. 857–863.

Jumper, J. *et al.* (2021) ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, 596(7873), pp. 583–589.

Kern, S. E. and Winter, J. M. (2006) ‘Elegance, silence and nonsense in the mutations literature for solid tumors’, *Cancer Biology and Therapy*, 5(4), pp. 349–359.

Kleffner, R. *et al.* (2017) ‘Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta’, *Bioinformatics*, 33(17), pp. 2765–2767.

Kontomanolis, E. N. *et al.* (2020) ‘Role of oncogenes and tumor-suppressor genes in carcinogenesis: A review’, *Anticancer Research*, 40(11), pp. 6009–6015.

Krissinel, E. (2015) ‘Stock-based detection of protein oligomeric states in jsPISA’, *Nucleic Acids Research*, 43(W1), pp. W314–W319.

Laconi, E., Marongiu, F. and DeGregori, J. (2020) ‘Cancer as a disease of old age: changing mutational and microenvironmental landscapes’, *British Journal of Cancer*, 122(7), pp. 943–952.

Laskowski, R. A. *et al.* (1993) ‘PROCHECK: a program to check the stereochemical quality of protein structures’, *Journal of Applied Crystallography*, 26(2), pp. 283–291.

Lawrence, M. S. *et al.* (2013) ‘Mutational heterogeneity in cancer and the search for new cancer-associated genes’, *Nature*, 499(7457), pp. 214–218.

Lawrence, M. S. *et al.* (2014) ‘Discovery and saturation analysis of cancer genes across 21 tumour types’, *Nature*, 505(7484), pp. 495–501.

Lazebnik, Y. (2010) ‘What are the hallmarks of cancer?’, *Nature Reviews Cancer*, 10(4), pp. 232–233.

Li, Y. *et al.* (2019) ‘Gain-of-Function Mutations: An Emerging Advantage for Cancer Biology’, *Trends in Biochemical Sciences*, 44(8), pp. 659–674.

Lomize, M. A. *et al.* (2006) ‘OPM: Orientations of proteins in membranes database’, *Bioinformatics*, 22(5), pp. 623–625.

de Martel, C. *et al.* (2020) ‘Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis’, *The Lancet Global Health*, 8(9), pp. e180–e190.

Maryam, A. *et al.* (2019) ‘The Molecular Organization of Human cGMP Specific Phosphodiesterase 6 (PDE6): Structural Implications of Somatic Mutations in Cancer and Retinitis Pigmentosa’, *Computational and Structural Biotechnology Journal*, 17(2019), pp. 378–389.

Matias, C.-S. and Degregori, J. (2011) ‘How Cancer Shapes Evolution and How Evolution Shapes Cancer’, *Evolution*, 4(4), pp. 624–634.

Merlo, L. M. F. *et al.* (2006) ‘Cancer as an evolutionary and ecological process’, *Nature Reviews Cancer*, 6(12), pp. 924–935.

Miller, M. S., Thompson, P. E. and Gabelli, S. B. (2019) ‘Structural determinants of isoform selectivity in pi3k inhibitors’, *Biomolecules*, 9(3), pp. 1–35.

Mularoni, L. *et al.* (2016) ‘OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations’, *Genome Biology*, 17(1), pp. 1–13.

Ng, P. C. and Henikoff, S. (2003) ‘SIFT: Predicting amino acid changes that affect protein function’, *Nucleic Acids Research*, 31(13), pp. 3812–3814.

Pandurangan, A. P. *et al.* (2017) ‘SDM: A server for predicting effects of mutations on protein stability’, *Nucleic Acids Research*, 45(W1), pp. W229–W235.

Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011) ‘Sequencing technologies and genome sequencing’, *Journal of Applied Genetics*, 52(4), pp. 413–435.

Pérez-Palma, E. *et al.* (2019) ‘Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database’, *Nucleic Acids Research*, 47(W1), pp. W99–W105.

Pieper, U. *et al.* (2014) ‘ModBase, a database of annotated comparative protein structure models and associated resources’, *Nucleic Acids Research*, 42(D1), pp. 336–346.

Pierson, T. M. *et al.* (2014) ‘GRIN2A mutation and early-onset epileptic encephalopathy: personalized therapy with memantine’, *Annals of Clinical and Translational Neurology*, 1(3), pp. 190–198.

Piovesan, D. *et al.* (2021) ‘MobiDB: Intrinsically disordered proteins in 2021’, *Nucleic Acids Research*, 49(D1), pp. D361–D367.

Pires, D. E. V., Ascher, D. B. and Blundell, T. L. (2014) ‘MCSM: Predicting the effects of mutations in proteins using graph-based signatures’, *Bioinformatics*, 30(3), pp. 335–342.

Rajan, M. R. *et al.* (2016) ‘Systems-wide experimental and modeling analysis of insulin signaling through forkhead box protein O1 (FOXO1) in human adipocytes, normally and in type 2 diabetes’, *Journal of Biological Chemistry*, 291(30), pp. 15806–15819.

Ramsay, L. *et al.* (2000) ‘A simple sequence repeat-based linkage map of Barley’, *Genetics*, 156(4), pp. 1997–2005.

Šali, A. and Blundell, T. L. (1993) ‘Comparative Protein Modelling by Satisfaction of Spatial Restraints’, *Journal of Molecular Biology*, 234(3), pp. 779–815.

Sarabipour, S. and Hristova, K. (2016) ‘Mechanism of FGF receptor dimerization and activation’, *Nature Communications*, 7(10262), pp. 1–12.

Schultz, J. *et al.* (1998) ‘SMART, a simple modular architecture research tool: Identification of signaling domains’, *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp. 5857–5864.

Sehnal, D. *et al.* (2021) ‘Mol∗Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures’, *Nucleic Acids Research*, 49(W1), pp. W431–W437.

Sever, R. and Brugge, J. S. (2015) ‘Signal transduction in cancer’, *Cold Spring Harbor Perspectives in Medicine*, 5(4), pp. 1–21.

Shah, S. P. *et al.* (2012) ‘The clonal and mutational evolution spectrum of primary triple-negative breast cancers’, *Nature*, 486(7403), pp. 395–399.

Shahrouki, P. and Larsson, E. (2012) ‘The non-coding oncogene: A case of missing DNA evidence?’, *Frontiers in Genetics*, 3(170), pp. 1–8.

Shi, J., Blundell, T. L. and Mizuguchi, K. (2001) ‘FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties’, *Journal of Molecular Biology*, 310(1), pp. 243–257.

Sievers, F. *et al.* (2011) ‘Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega’, *Molecular Systems Biology*, 7(539), pp. 1–6.

Sillitoe, I. *et al.* (2020) ‘Genome3D: Integrating a collaborative data pipeline to expand the depth and breadth of consensus protein structure annotation’, *Nucleic Acids Research*, 48(D1), pp. D314–D319.

Sillitoe, I. *et al.* (2021) ‘CATH: Increased structural coverage of functional space’, *Nucleic Acids Research*, 49(D1), pp. D266–D273.

Söding, J. (2005) ‘Protein homology detection by HMM-HMM comparison’, *Bioinformatics*, 21(7), pp. 951–960.

Sondka, Z. *et al.* (2018) ‘The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers’, *Nature Reviews Cancer*, 18(11), pp. 696–705.

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009) ‘The cancer genome’, *Nature*, 458(7239), pp. 719–724.

Szklarczyk, D. *et al.* (2021) ‘The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets’, *Nucleic Acids Research*, 49(D1), pp. D605–D612.

Takeshima, H. and Ushijima, T. (2019) ‘Accumulation of genetic and epigenetic alterations in normal cells and cancer risk’, *npj Precision Oncology*, 3(1), pp. 1–8.

Tan, M. E. *et al.* (2015) ‘Androgen receptor: Structure, role in prostate cancer and drug discovery’, *Acta Pharmacologica Sinica*, 36(1), pp. 3–23.

Tang, H. and Thomas, P. D. (2016) ‘Tools for predicting the functional impact of nonsynonymous genetic variation’, *Genetics*, 203(2), pp. 635–647.

Tate, J. G. *et al.* (2019) ‘COSMIC: The Catalogue Of Somatic Mutations In Cancer’, *Nucleic Acids Research*, 47(D1), pp. D941–D947.

Vogelstein, B. *et al.* (2013) ‘Cancer genome landscapes’, *Science*, 340(6127), pp. 1546–1558.

Watkins, X. *et al.* (2017) ‘ProtVista: Visualization of protein sequence annotations’, *Bioinformatics*, 33(13), pp. 2040–2041.

Weinstein, J. N. *et al.* (2013) ‘The cancer genome atlas pan-cancer analysis project’, *Nature Genetics*, 45(10), pp. 1113–1120.

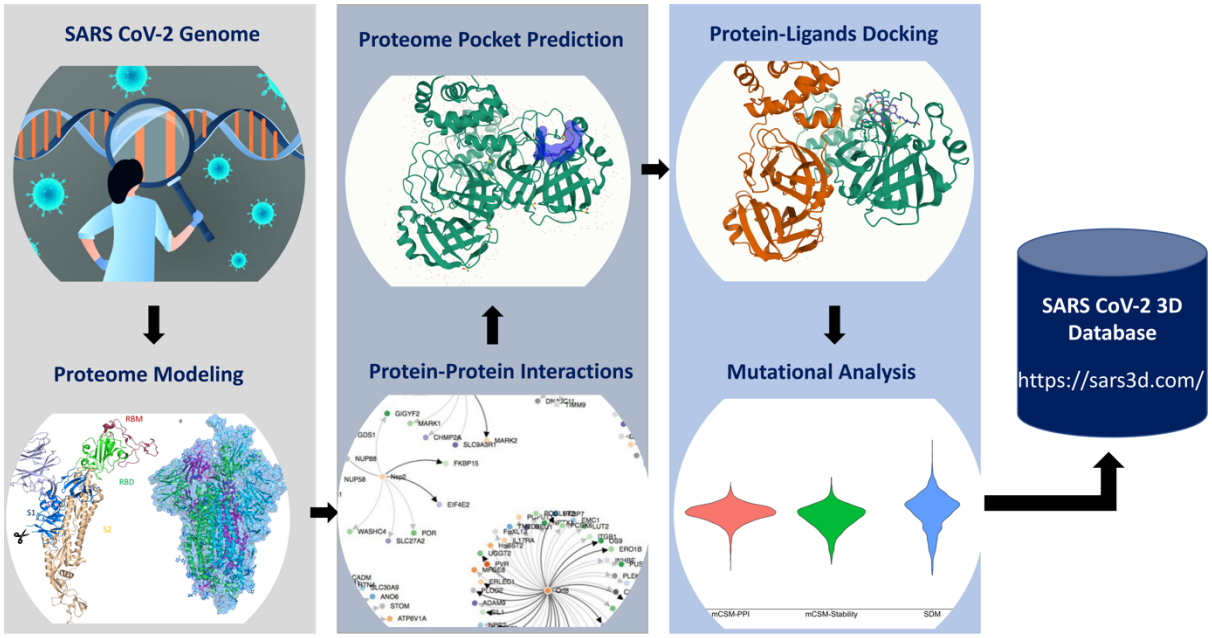
Wong, C. C. *et al.* (2014) ‘Inactivating CUX1 mutations promote tumorigenesis’, *Nature Genetics*, 46(1), pp. 33–38.

Wu, S. *et al.* (2018) ‘Evaluating intrinsic and non-intrinsic cancer risk factors’, *Nature Communications*, 9(1), pp. 1–12.

Yachdav, G. *et al.* (2016) ‘MSAViewer: Interactive JavaScript visualization of multiple sequence alignments’, *Bioinformatics*, 32(22), pp. 3501–3503.

Zhang, X. *et al.* (2011) ‘Structure of Lipid Kinase p110β/p85β Elucidates an Unusual SH2-Domain-Mediated Inhibitory Mechanism’, *Molecular Cell*, 41(5), pp. 567–578.

Chapter II: SARS-CoV-2 3D Proteome Database



Introduction

SARS CoV-2 background.

The severe acute respiratory syndrome coronavirus 2 (SARS CoV-2) was first detected in Wuhan, China late December 2019. Since then, it has spread dramatically, causing a lot of health and economic challenges worldwide (World Health Organization, 2020). The ongoing outbreak has triggered a significant threat to our global public health. According to the health world organisation, SARS CoV-2 led to 300 million reported cases and 5.2 million deaths worldwide (WHO, 2021). The origin of SARS CoV-2 is probably from bats and possibly another animal such as the pangolin, which acts as an intermediate host (Andersen *et al.*, 2020). Human to human transmission was confirmed when reported cases increased dramatically in February 2020 (WHO, 2020). As a result, the world health organisation WHO announced the COVID-19 outbreak as a pandemic that needs immediate health attention. The SARS CoV-2 was named by the International Committee on Taxonomy of Virus, whereas the WHO named the COVID-19. In February, the outbreak peaked in China, travel and outdoor activities were blocked because of those measures, the number of cases gradually decreased. However, due to less restrictive measures in the rest of the world, cases increased dramatically until restrictive measures were implemented. A high mortality rate was reported when the health system was overwhelmed, especially in the USA (Stokes *et al.*, 2021). The genetic evidence strongly suggests that the SARS CoV-2 is a naturally occurring virus, i.e. not synthetically made or modified. However, when and where first the virus gets into a human, this is still missing (Boni *et al.*, 2020). One case reported in France in 2019 with coronavirus was earlier than the outbreak (Carrat *et al.*, 2021).

Large samples from animals, environments and patients worldwide need to be taken and validated with a well-established assay to overcome these speculations.

The virus SARS CoV-2 belongs to seven human coronavirus families that involve coronaviruses 1 (SARS-1) and middle east respiratory syndrome (MERS), which cause more respiratory severe illnesses. Whereas HCoV-NL63, HCoV-229E, HCoV-OC43, and HKU1 usually cause mild respiratory disease (Gorbalenya *et al.*, 2020). The SARS CoV-2 is closely related to other coronaviruses, and it has roughly 79% sequence identity to SARS CoV-1 and 50% sequence identity to MERS (Cui, Li and Shi, 2019). However, the transmission rate is different even though the coronavirus family has a high sequence similarity. For example, SARS-1 tends to be more lethal but has more minor infections, whereas SARS CoV-2 is the opposite (Abdelrahman, Li and Wang, 2020). In addition, although SARS CoV-2 is mainly transmitted through contact from respiratory droplets from the mouth and nose, the WHO announced that the droplet could stay longer in the air than expected (WHO, 2018).

For this reason, social distance roles were implemented to reduce more spread of the virus. SARS CoV-2 and MERS infect bronchial epithelial cells causing viral pneumonia, including cough, dyspnea, and chest pain. Although all ages are susceptible for infection in general, the elderly is more likely to develop severe disease and require hospitalisation. In contrast, others are expected to develop mild symptoms. Loss of taste and olfactory are prevalent symptoms. Usually, people with these symptoms are carriers of the virus and require self-isolation (Lopez-Leon *et al.*, 2021). Therefore, early diagnosis of SARS CoV-2 is vital in controlling the spread of the virus.

The first genome sequence of SARS CoV-2 was released in January 2020 (GenBank: MN908947.3). Lately, other genome sequences have been released and deposited into the GISAID database (https://www.gisaid.org/). An analysis of the SARS CoV-2 genome shows sequence similarity to coronavirus in bats (RaTG13) and pangolins. The SARS CoV-2 is roughly 96% identical to RaTG13 and 92.2% to Pangolins (Wacharapluesadee *et al.*, 2021). Furthermore, all ORFs shared approximately 90% sequence Identity (Michel *et al.*, 2020). This high similarity supports that either bats or pangolins are the main reservoir of SARS CoV-2 (Wacharapluesadee *et al.*, 2021). However, the study shows that RaTG13 spike protein binds with low affinity to the human ACE2 receptor, hypothesising pangolins are more likely the intermediate host.

Since the spike protein in RaTG13, pangolins, and SARS CoV-2 are very similar, it has been tested against multiple AEC2 receptors. Pangolin and SARS CoV-2 spike proteins showed strong binding to human ACE2 and weak affinity to bat ACE2 protein. This similar binding affinity of Pangolin and SARS CoV-2 is due to the high structural and sequence similarity of the spike receptor-binding domain (RBDs) compared to the RBD of RaTG13.(Wrobel *et al.*, 2021). However, infected Pangolins showed clinical signs of cell inflammation and interstitial pneumonia, indicating that the Pangolins are unlikely to be the host for SARS CoV-2 (Li *et al.*, 2021). The host needs to be international for the virus to be an outbreak, for example, camels in MERS-CoV and palm civets in SARS CoV. In addition, the virus strains found in these hosts are almost identical to the virus found in humans, with more than 99% sequence identity. Therefore, although the virus in Pangolins has a spike receptor-

binding domain 92% identical to humans, the available data are insufficient to indicate that Pangolins are the intermediate host for SARS CoV-2.

SARS CoV-2 genome

The SARS CoV-2 is characterised as positive sense-single strand RNA about 120 nm in diameter. SARS CoV-2 genome consists of three regions from the 5' to 3' in order: firstly, the polyproteins pp1a and pp1ab (ORF1a/ORF1b). Secondly, the structure proteins include Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N). Thirdly, an accessory protein involved seven to eleven putative accessory proteins (ORF3a, ORF3b, ORF3c, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c, ORF10) (Wu *et al.*, 2020)(Lu *et al.*, 2020). The replicase ORF1a/ORF1b covers two-thirds of the genome encode for sixteen non-proteins (nsp1 to nsp16) that are proteolytically cleaved to be involved in transcription of the virus genome. The SARS CoV-2 spike protein size is 1273 amino acids, slightly larger than SARS-bat (1,269 amino acids) and SARS CoV-1 (1255 amino acids). The distinction between SARS CoV-2 and another type of coronaviruses family are the accessory proteins. For example, ORF8 showed only 40% sequence identity and had different roles inside the host. The SARS CoV-2 ORF8 does not include a motif that triggers the intercellular stress pathway. ORF8 deletion has been observed in patients indicating human adaptation from intermediate species (Su *et al.*, 2020).

The high spreading of the SARS CoV-2 results in genetic variation of different SARS CoV-2 strains. The National Genomics Data Center (NGDC) (https://ngdc.cncb.ac.cn/ncov/) in China aligned 4094258 genome sequences detected globally and identified 39079 mutations, including 29142 single-nucleotide

polymorphisms, 6577 deletions, 2779 insertions, and 581 indels (Xue *et al.*, 2021). There are 5681 unique mutations reported for the spike protein, which occur in all domains. Mutations such as V483A, and L455I, are proposed to be the most detrimental. However, more experimental evidence is needed to confirm the effect of these mutations on binding to the host receptor (Harvey *et al.*, 2021). The alteration of D614G in the S1 subunit became the domain and circulated internationally. The D614 variants (N, Y, G, E, A) postulated increasing viral loads. However, no clinical data suggested any link between these mutations and SARS CoV-2 severity (Harvey *et al.*, 2021). (Figure supplementary-2). This needs further experimental validation by comparing the wild type and the mutant using an animal model.

SARS CoV-2 mechanism.

The initial step of SARS CoV-2 involved binding the SI subunit of the spike protein to the human epithelial cell cellular surface receptor angiotensin-converting enzyme 2 (ACE2). Besides human SARS CoV-2 also recognised the ACE2 in multiple animals such as pangolin, cats etc. (Perlman and Netland, 2009). The level of ACE2 expression on the surface will influence viral pathogenicity. Theoretically, all organs that express ACE2 receptors can be infected by the SARS CoV-2 virus. Liver and heart issue, as well as intestinal inflammation, has been reported in COVID-19 patients.

Upon binding, spike protein undergoes proteolytic cleavage catalysed by two proteases (TMPRSS2, furin). This process is required for the SARS-S/ACE2 fusion. The insertion of unique four basic amino acid residues at the boundary site of S1/S2 domains cleaved by furin detaching the two domains S1/S2 and removing the structure constraint. Subsequently, the TMPRSS2 cleaves the fusion peptide located

downstream from the S2 domain after the spike protein undergoes conformational change exposing the internal fusion peptide (Zhang *et al.*, 2021). The four basic amino acid residues in SARS CoV-2 make it different from bat coronavirus RaTG13 enhancing the cleaving rate by furin protease. For this reason, SARS CoV-2 is believed to be more infectious than another coronavirus (Perlman and Netland, 2009)(Yang *et al.*, 2020).

The SARS-S/ACE2 interaction has been studied at the atomic level (Yang *et al.*, 2020). Other structural proteins such as Nucleoprotein (N) encapsulate the positive-sense single-stranded RNA. Membrane (M) and Envelop (E) ensure fusion in the viral particle during the assembly process. Upon the SARS CoV-2 entry, the virus releases the positive single RNA strand into the cytoplasm to replicate. The replication of the RNA involved two steps; firstly, the RNA was transcribed via replication-transcription complexes (RTC), producing negative RNA that in turn replicated into positive RNA, which repackaged into the viral offspring. In addition, the negative RNA can undergo discontinuous transcription firstly proposed by Sawicki that produces different mRNA in different lengths, therefore, encodes an additional protein known as subgenomic mRNA, which repackages again into the virus virion. Thus, the secretory pathway makes the virion progeny; translation of the mRNA in the rough endoplasmic reticulum then passes to the Golgi apparatus to be released in vesical and finally exocytosis to infect another cell. Spike protein open and close conformation of SARS CoV-2 and SARS CoV-1 shows different affinity to the ACE2 receptor. The receptor-binding domain of SARS CoV-2 shows a greater binding affinity to the ACE2 receptor conferring its higher infectivity (Figure 1) (Perlman and Netland, 2009) (Zhang *et al.*, 2021).

After the virus enters the cell, the replication process of the virus damage essential factor, releasing inflammatory mediators that trigger the secretion of multiple cytokines such as interleukin-1 and 2 Tumor necrosis factor-alpha (TNF-∝), cytokines are associated with SARS CoV-2 severity by increasing the endothelial cell permeability, damaging the alveoli and leading to Hypoxemia. Therefore, finding key cytokines induced by SARS CoV-2 and inhibiting its signalling is essential to stop immune damage to the lung. Furthermore, the spike protein was reported to down-regulate the ACE2 physiological function on the cell surface by dysfunctioning the renin-angiotensin system leading to an inflammatory response (Huang *et al.*, 2020)(Chen *et al.*, 2020).

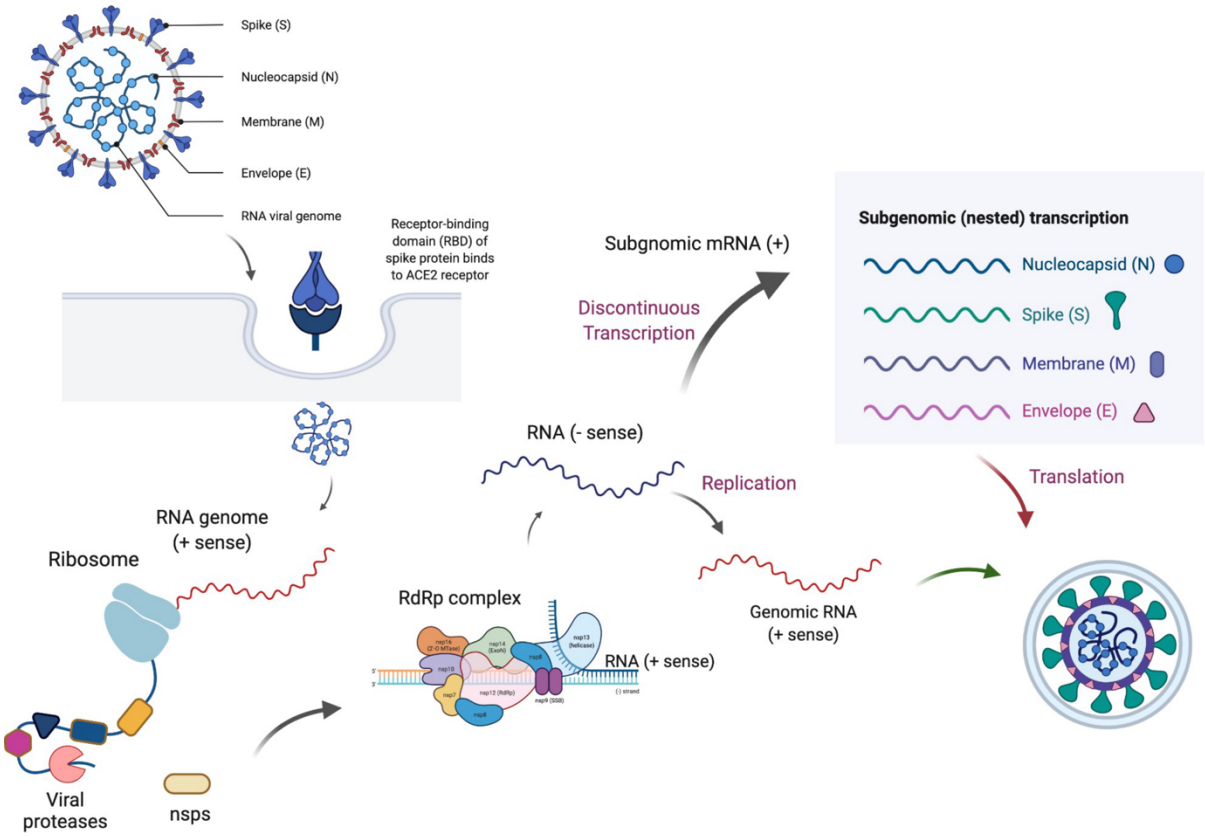


Figure 1. (made by biorender App) Cartoon representation of SARS CoV-2 enter mechanism. Starting with spike protein binds to ACE2 receptors releasing the virus genome into the host cell. Next, the virus uses host cell machinery to translate pp1a and pp1ab

proteins, which are then cleaved by the protease. Finally, the cleaved protein forms the RdRp complex, which replicates the (-sense) into full (+sense) and subgenomic (nested) transcription, which all repackage into the virion progeny to be released to infect another cell.

Genome translation

The release of the virus into the cytoplasm is just the beginning of SARS CoV-2 gene expression that is highly regulated in space and time. The SARS CoV-2 has two open reading frames, ORF1a and ORF1b produce what are known as pp1a and pp1ab polyproteins, respectively. The overlap in the ORF1a and ORF1b regions results in a ribosomal frameshift with efficiency between 40-70%. As a result, pp1a expresses 1-2 times more than pp1ab. The translation of pp1a expresses (nsp1-11) whereas pp1ab (nsp1-10, and nsp12-16). These sixteen proteins are auto proteolytically by two cysteine proteases, papain-like protease (nsp3) and chymotrypsin-like protease (nsp5), with nsp5 responsible for the cleavage of most of the polyprotein (Liu *et al.*, 2021). The rapid release of the host translation inhibitor (nsp1) enables it to target the host cell translation machinery (Schubert *et al.*, 2020). The non-structural protein (nsp2-16) supports the RTC complex by host immune evasion and modulating membranes. The nsp7 and nsp8 cofactors are critical to the SARS CoV-2 RNA synthesis machinery (Peng *et al.*, 2020). The non-structural protein (nsp12-16) contains the essential enzyme for including RNA synthesis, proofreading, and modification (Yan *et al.*, 2020). In general, the structural proteins (Spike, Envelope, Membrane, and Nucleocapsid) are not associated with SARS CoV-2 assembly and budding of the new virions. However, one study shows that SARS CoV-2 infected cells through lysosomal trafficking pathways via interfering with lysosomal acidification (Ghosh *et al.*, 2020). Although not all accessory proteins have been verified yet, it has been postulated that they are not essential for SARS

CoV-2 replication but have a role in pathogenicity. At least five accessory proteins were expressed due to leaky scanning of subgenomic RNA of the nucleocapsid protein (ORF3a, ORF6, ORF7a, ORF7b, and ORF8). ORF10 is located downstream of the nucleocapsid gene (Alexandersen, Chamings and Bhatta, 2020).

Spike protein and pathogenesis

The spike protein is the most studied receptor in SARS CoV-2. It is essential for virus attacks and invasion. As mentioned above, spike protein binds to an ACE2 receptor in humans and other animals with different potency. The spike protein contains S1 and S2 domains. The S1 was further divided into N-terminal and C-terminal, including the receptor-binding (RBD) residue 319-529 and the receptor-binding motif (RBM) residues 437-507, which are involved in interaction with the ACE2 receptor. Since the receptor-binding motif is the main target for neutralising antibodies, it is essential to study the virus-binding hotspot (Shang *et al.*, 2020). (Figure 2)

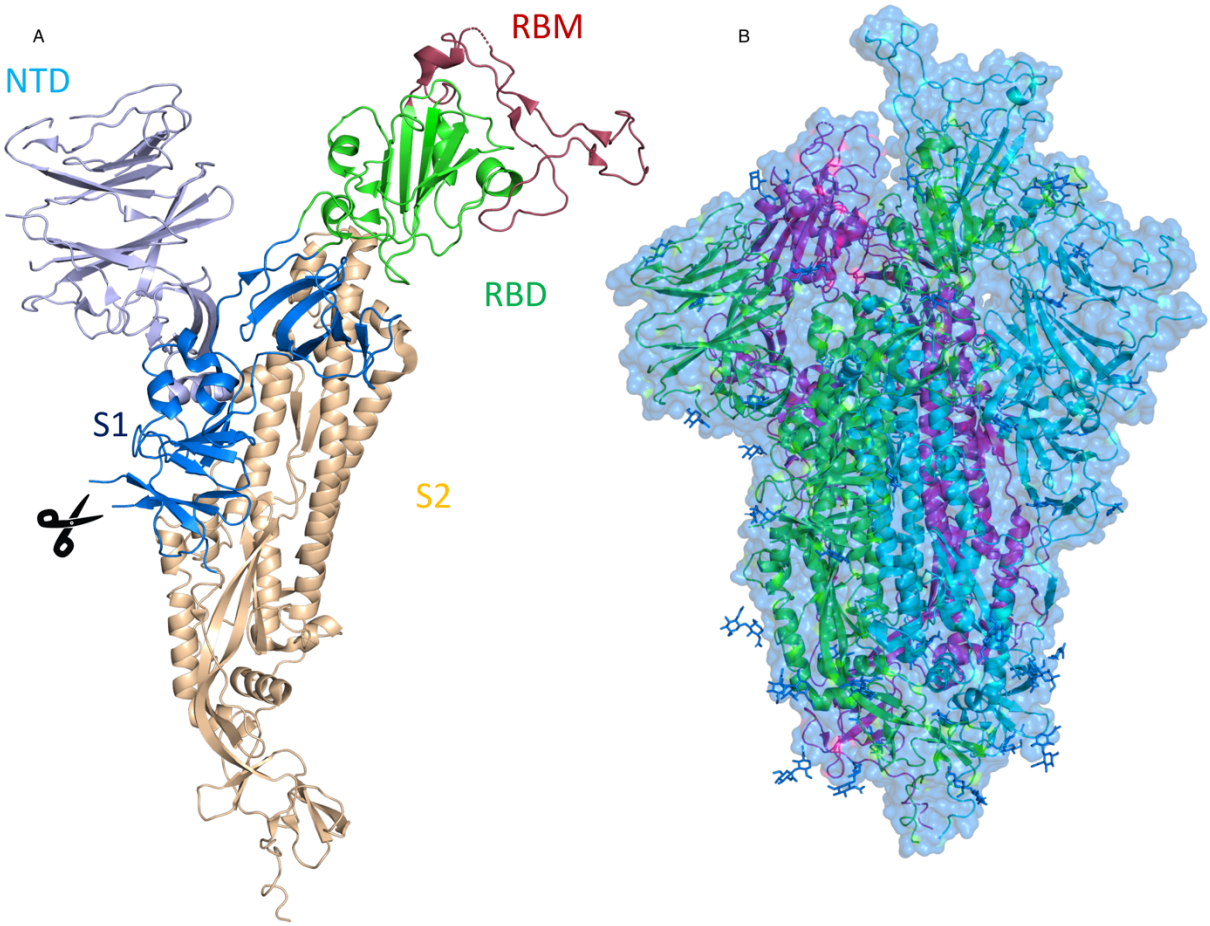


Figure 2. Spike monomer and homo 3-mer structures. (A) Upright is the receptor-binding domain which includes the receptor-binding motif in red. The scissors represent the boundary between the S1 domain coloured in blue and the S2 domains coloured in wheat. The amino-terminal domain (NTD) is coloured in light blue. (B) surface representation of the homo 3-mer structures in the open conformation (PDB ID: 6CRZ) each monomer coloured differently, the glycan represented in stick blue.

The SARS CoV-1 and SARS CoV-2 interaction with AEC-2 are formed through the receptor-binding domain. Comparing SARS CoV-2 receptor-binding domain to SARS CoV-1, they are different in five residues (Y455L, L486F, N493Q, D494S and T501N). Due to this difference in the virus-binding hotspot, more hydrogen bonding formed for SARS CoV-2 results in more compact conformation than SARS CoV-1. Structurally the difference between the spike in SARS CoV-1 and SARS CoV-2 is in the conformation loop in the AEC-2 binding ridge. The loop in SARS CoV-1 has a

tandem proline which allows it to take a sharp turn, whereas in SARS CoV-2 does not have a proline enabling the loop to adapt to different conformation. As a result of this difference, additional hydrogen bonding (Asn487-SARS-2 to Ala475-SARS-2) and (Ala475-SARS-2 to Ser19-AEC-2) are formed. The structure change at the AEC-2 binding ridge is caused mainly by four amino acid residues, namely G482, V483, E484, and G485, which allow the ridge to be more compact and enhance interaction with the N-terminal helix of the ACE2 receptor. (Figure 3).

Furthermore, Phe486 in SARS CoV-2 is larger than leucine in SARS CoV-1, its points into a hydrophobic pocket, whereas the leucine forms weaker Interaction with AEC-2. Finally, two lysines at the hotspots on ACE2 need to be accommodated by the hydrophobic environments. Neurulation of these lysine residues is essential for the receptor-binding domain to bind to the AEC-2 receptor. Biochemical data confirmed that the binding of the SARS CoV-2 receptor binding domain is much stronger than SARS CoV-1 due to the above structural features. The SARS CoV-2 recognised the AEC-2 receptor better than SARS CoV-1. Cryo-electron microscope structure of the spike protein revealed that in SARS CoV-2, the receptor-binding domain could exist in open and close conformations. The close confirmations of the SARS CoV-2 are hypothesised to be favoured for receptor binding (Shang *et al.*, 2020).

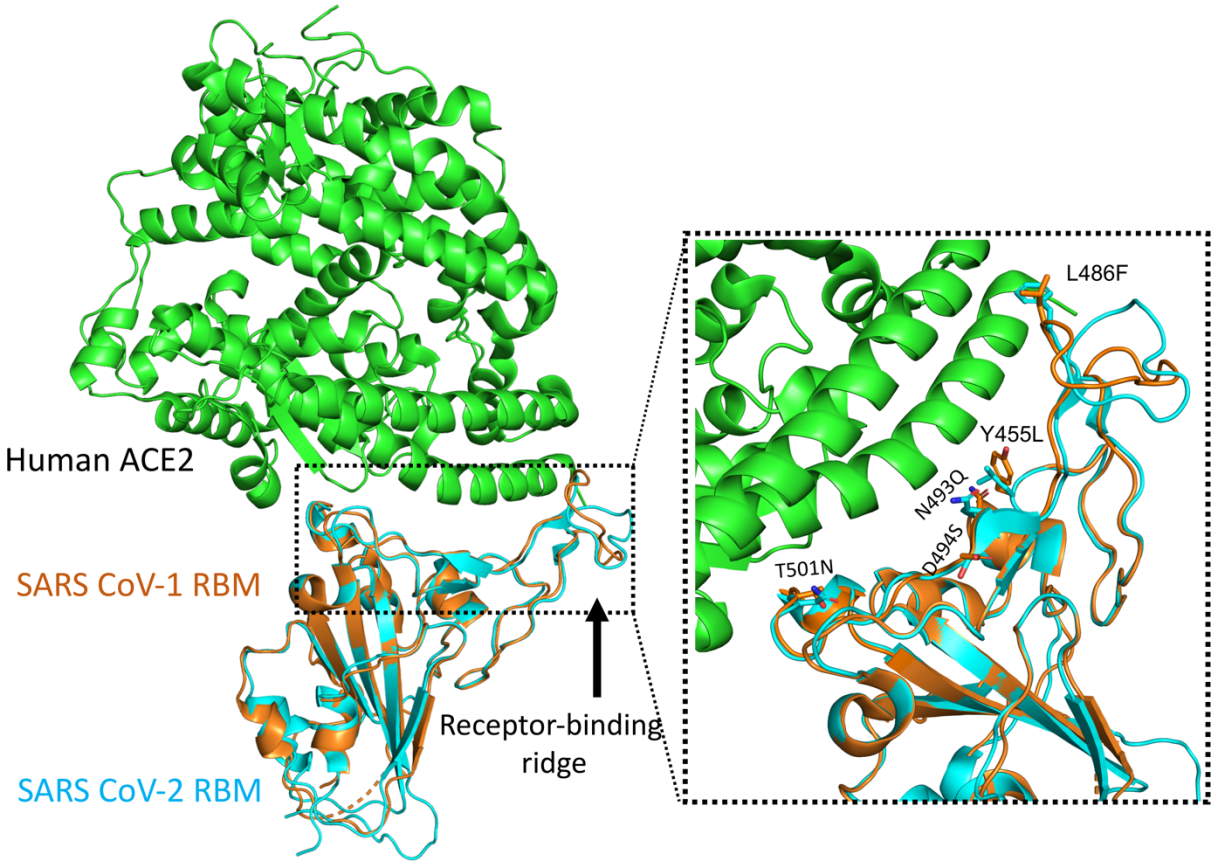


Figure 3. Crystal structure of AEC2 receptor in complex with spike receptor-binding domain SARS CoV-2 and SARS CoV-1. The human ACE2 is coloured in green, and the SARS CoV-2 chimeric RBD is coloured in cyan, whereas the chimeric RBD for SARS CoV-1 is coloured in orange. The difference in loop conformation is shown at the receptor binding ridge, whereas the residues are different at the interface hotspot shown in the sticks.

SARS CoV-2 drug development strategy and technology.

Multiple vaccines have been approved by the food and drug administration (FDA), including Pfizer, AstraZeneca, and Johnson and Johnson. Since the SARS CoV-2 remains longer in the human body, the best and fastest strategy is to develop a vaccine to combat the outbreak. This will allow people to build up immunity against the virus. The Pfizer vaccine brand name is Comirnaty and has been widely distributed worldwide. It works by introducing the cell with mRNA that makes the SARS CoV-2 protein. The immune system detects these proteins and triggers the

response to develop antibodies. However, no orally available drug has been proven. Therefore, screening available antiviral drugs with known toxicity and efficacy was a primary stage in drug development. Unfortunately, all the antivirals gave unsatisfactory results, and most patients showed no improvement (Ledford, 2021). This is potentially due to the upregulation of the ACE2 receptor and the target effect of the antiviral drug. Therefore, the second strategy is finding a drug based on conservation with another homologue since multiple SARS CoV-2 targets, including the catalytic site of SARS CoV-2, are highly conserved. Therefore, main proteases (M^{pro}, $3CL^{pro}$) drugs such as Lopinavir and Ritonavir were hypothesised (Zumla *et al.*, 2016).

Unfortunately, again no treatment benefit was observed. Another hypothesised drug is Remdesivir, which was used to treat Ebola. It interferes with viral RNA synthesis nsp12, nsp7, and nsp8. The scientific community has controversial this drug, and it is not yet accepted as a treatment for SARS CoV-2 (Gao *et al.*, 2020). Finally, chloroquine and hydroxy-chloroquine were hypothesised for SARS CoV-2 treatment at the beginning of the pandemic. Although both drugs are known to treat malaria, hydroxychloroquine is reported as more potent and effective (Li and De Clercq, 2020). However, a study in Brazil showed that hydroxy-chloroquine alone or combined with other drugs has no improvement in patient symptoms (Cavalcanti *et al.*, 2020). High throughput virtual screening, molecular docking, and artificial intelligence of the existing library of drugs or small molecules are powerful techniques. The main protease became the apparent target (Jin *et al.*, 2020) (Dai *et al.*, 2020). Other compounds were identified through molecular docking, structural fragment base, and cell-based assays, including Michael acceptor as a covalent bonding, Ebselen pharmacophore etc.

The emergence of new technology, artificial intelligence, and the fast development of detection methods of diagnosis has increased. One of these technologies, PROTACs, a new way of developing novel drug development, has gained much attention. PROTACs have three chemical elements: ligand binding to the target protein, ligand binding to E3 ubiquitin ligase, and a linker of the two ligands. The advantage of the 26S proteasome pathway is that it will degrade the target protein (Lecker, Goldberg and Mitch, 2006). The second technology is CRISPR-based nucleic acid detection for SARS CoV-2, which reduces the test time to less than one hour (Wang *et al.*, 2020). Many rapid nucleic acid kits have been developed to detect Spike, Envelope, and Nucleoprotein within an hour. However, a false-positive result is possible from these kits since it is affected by multiple factors such as sample method, sputum and throat swabs (Gootenberg *et al.*, 2017).

Valuable resource developed for SARS CoV-2.

Coronavirus3D is a resource for tracking the SARS CoV-2 lineages by identifying mutation variants from different geographical regions (Sedova *et al.*, 2020). Experimental determination of 3D structure is essential for understanding macromolecular interactions and identifying drug targets. 3D experimental structures help develop an effective and selective molecule against the target viral protein.

All SARS CoV-2 structures are deposited in the Protein Data Bank (PDB), a unique archive. Comparative models of SARS CoV-2 proteins from different groups have been developed and presented in Swiss-Model (Waterhouse *et al.*, 2018), I-TASSER (Yang and Zhang, 2015), AlphaFold, and SARS CoV-2 3D. The China National

Center for Bioinformation (CNCB) resource has integrated genomic variation on a global scale, with functional annotation and annotations of frequencies of mutations. Well established resources, such as Pfam for protein family annotation and InterPro for classifying proteins in families, have proved useful for predicting domains and other important features. Chemical Checker (Duran-Frigola *et al.*, 2020) has collated all small bioactive molecules that elicit activity against SARS CoV-2. ChEMBL database (Gaulton *et al.*, 2017) has manually curated bioactive drug-like molecules useful for drug-repurposing for SARS CoV-2. Large-scale genetic analysis has identified 1263 druggable gene products implicated in SARS CoV-2 activity. Some of these targets have approved drugs. The study shows two proteins, interferon alpha/beta receptor 2 (IFNAR2) and ACE2 protein, as useful drug targets for the early management of the virus and to reduce hospitalisation. COVID-19 Docking server has been developed to dock small molecules against various SARS CoV-2 proteins (Han *et al.*, 2021).

Methods

Proteome modelling

The target gene sequences were retrieved from the GeneBank: MN908947.3. The SARS CoV-2 sequences were blasted against PDB to calculate structural coverage. Other databases, such as Pfam, CATH, SCOP, SMART and UniProt were queried to map domains and annotate transmembrane regions. Sequences not associated with SARS CoV-2 structural annotations were blasted against PDB using FUGUE, PSI-BLAST, and HHsearch. The selected templates were based on sequence identity, coverage, and resolution. The target SARS CoV-2 sequences were re-aligned to the selected templates using Clustal Omega. The generated sequence alignment file was processed using MODELLER to produce the final model structure, then minimised using Foldit to remove steric clashes. Cofactors, such as metal ions and ligands, were obtained from the selected templates.

There are nine genes with a transmembrane region (nsp3, nsp4, nsp6, ORF3a, ORF7a, Spike, Envelope, Membrane, and Nucleoprotein), which are clearly annotated. The ProtCHOIR (Torres, Rossi and Blundell, 2021) automated pipeline software was used to build homodimeric structures for proteins such as the Envelope protein. ProtCHOIR uses PSI-BLAST to query a local homodimer database generated from PDB. MODELLER was used to create the models, GESAMT was used to align and superimpose protein structures in 3D, PISA was implemented to assess the dimer interface, and MolProbity to assess the finial homodimer structure. (Figure 4)

All the homo and heterodimer complexes were built in the same way, using COSMIC CGC 3D. For example, SARS CoV-2 nsp14-nsp10 heterodimeric complex was based on the crystal structure of the SARS CoV-1 coronavirus nsp14-nsp10 complex with functional ligands guanosine-p3-adenosine-5',5'-triphosphate. Open and closed conformations were considered based on solved structures, for example, for spike protein. At the time SARS CoV-2 3D was developed, a few proteins with full structural coverage had been solved experimentally by X-ray and Cryo-EM, such as nsp5 and nap16. Therefore, no modelled structures were created for SARS CoV-2 proteins with good experimental structural coverage. The biological assembly of the experimental structure was downloaded from the PDB and presented on the website.

Prediction of impacts of mutations

The emergence of SARS CoV-2 variants makes it necessary to understand the impacts of mutations on protein stability and protein-protein interfaces. Seven essential SARS CoV-2 drug targets were evaluated (nsp3, nsp5, nsp9, nsp12, nsp13, nsp15, and spike protein). Multiple computational tools with different computational approaches were used to characterise mutations:

I. mCSM-stability and mCSM-PPI are graph-based signatures that rely on the residues' distance patterns and environments.
II. DeepDDG, another tool, was used to predict the SARS CoV-2 mutations. It is a neural network trained based on 5700 experimental data points of protein thermodynamic stability curated from Protherm database and literature searched (Cao *et al.*, 2019).

III. Protein Variation Effect Analyser (PROVEAN) is a software that predicts the impact of mutations using an alignment-based score approach (Choi and Chan, 2015).

IV. MAESTRO structure-based tool uses a multi-agent machine learning system (Laimer *et al.*, 2015)

V. I-Mutant is another structural tool used. It is based on a support vector machine (SVM) (Capriotti, Fariselli and Casadio, 2005)

Most of these tools also quantify the change in the free energy $\Delta\Delta G$ between the wild type and the mutant forms. (Figure 4)

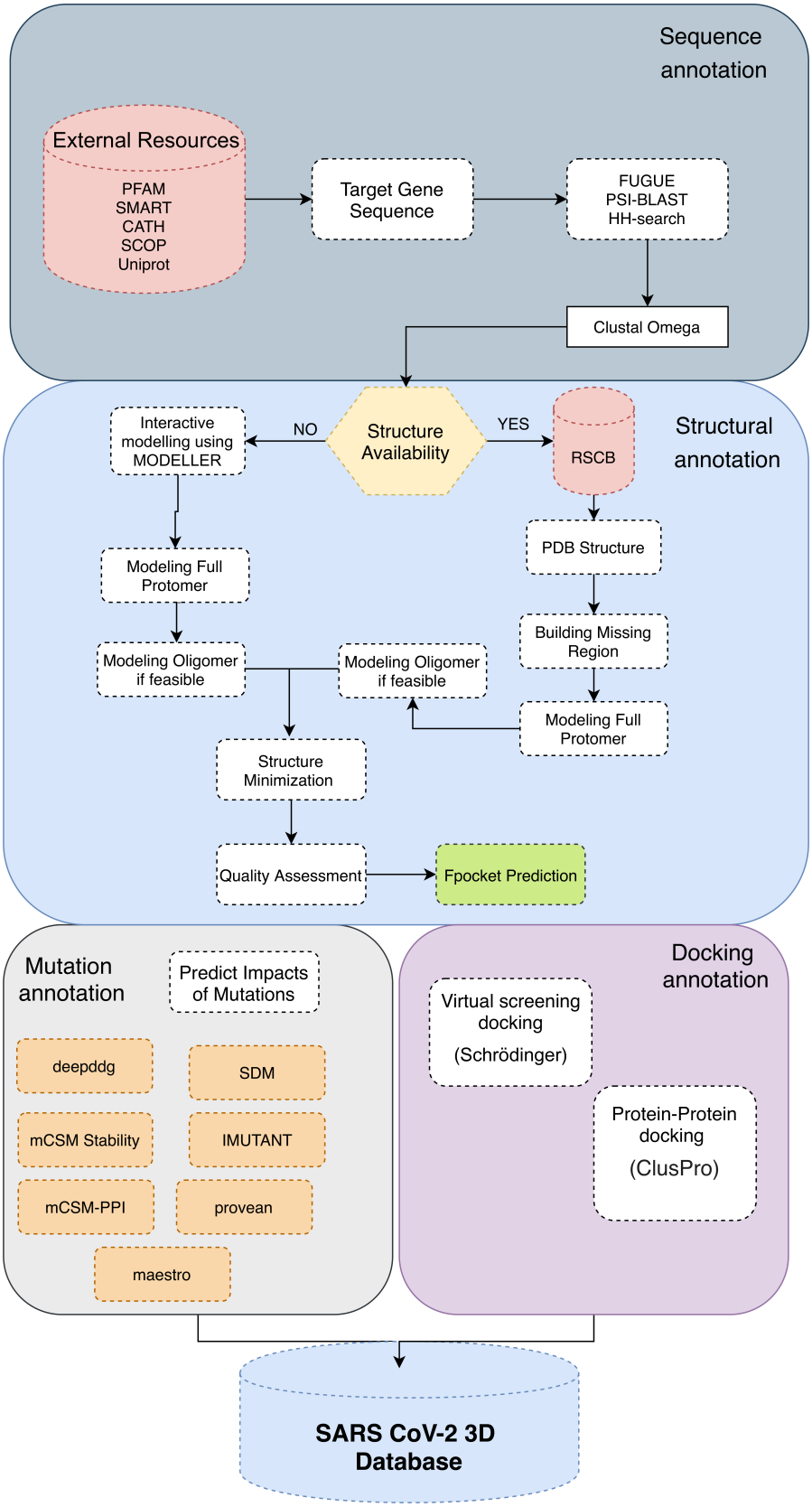


Figure 4. schematic flow of data deposited into the SARS CoV-2 database, including oligomeric modelling using MODELLER, mutations analysis using various algorithm tools, molecular docking using Schrödinger and Protein-protein docking using ClusPro.

Virtual screening docking

Molecular docking was performed using Schrödinger Suite 2020-2. The virtual screening docking calculation can predict different ligand-binding conformations known as poses and explore the behaviour of small molecules in the active site (Brooks *et al.*, 2008). The virtual screening docking has been performed using X-ray, cryo-EM experimental structure, and homology modelling on seven SARS CoV-2 targets. The target selections were based on tractability and ligandability:

I. nsp3 (PDB ID; 6XAA)
II. nsp5 (PDB ID; 6XMK)
III. nsp12 (PDB ID; 7C2K)
IV. nsp14, when the SARS CoV-2 developed, the nsp14 experimental structure was not solved. Modelled structure with high accuracy used for virtual screening docking.
V. nsp15 (PDB ID; 6XDH)
VI. nsp16 (PDB ID; 6WKQ)
VII. Spike protein (PDB ID; 6VSB)

1930 FDA approved ligands were downloaded from the eDrug 3D web resource (Douguet, 2018). However, the structure file from the PDB is not suitable for immediate use in molecular docking calculations. PDB files may have missing atoms, usually in an intrinsically disordered region with high B-factor and connectivity information between residues. These must be assigned along with charges and bond orders. The Protein Preparation Wizard in Schrödinger includes a set of tools that suitably prepare proteins for calculations, for example, refining imported protein structures and optimising oriented-hydrogen-bonding groups. The receptor grid-

generation panel in Schrödinger allow the grid file to be prepared from a prepared protein structure (Madhavi Sastry *et al.*, 2013). The square grid represents the physical property of a volume of the receptor, i.e. the active site that is searched for a potential ligand binding. Grid generation is a crucial step since the docking cannot be performed without it. The grid box was generated from the reference ligand centroid (Ban, Ohue and Akiyama, 2018). The default parameters were used, and no ligand or protein constraints were applied. The LigPrep panel in Schrödinger was used to prepare 1930 FDA approved ligands. The concept behind ligand preparation is to produce low energy 3D structures by generating different tautomers, stereochemistry, ring conformation, and adding hydrogen atoms. A total of 24992 conformers were generated from input 1930 ligands (Press, 2015).

The ligand-docking panel in Schrödinger was used to perform the molecular docking with prepared receptors and ligands. The ligand docking was performed in two steps. Firstly, the ligands were docked using high-throughput virtual screening (HTVS), intended for docking an extensive library of ligands. This method has very limited sampling with a docking rate of two seconds per compound. Next, the top 10% hits of the docked ligand by HTVS were redocked using standard precision (SP). SP is essentially the same algorithm as HTVS but without constrictions of the intermediate ligand conformations or torsion refinement and sampling. Despite the differences between HTVS and SP, both use a series of hierarchical filters to search the active site shape and properties to find the best possible position of the docked ligand.

Protein-protein-docking

A list of experimentally validated protein-protein interactions between SARS CoV-2 proteins and accessible human proteins has revealed multiple drug targets. The human target protein is mapped to PDB to fetch experimentally solved structures with more than 90% coverage. The selected PDB file is processed to ClusPro software, which is widely used for protein-protein docking (Weinstein *et al.*, 2013). ClusPro performs the protein-protein docking in three steps:

I. Rigid-body docking using PIPER, a docking program based on Fast Fourier Transform (FFT).
II. Selecting the lowest RMSD structure generated will likely represent the complex.
III. Refinement of the selected structure using energy minimisation.

The top 4 poses from 200 poses are selected and presented to the SARS CoV-2 3D database. However, all other poses can be downloaded and viewed locally from the help page on the SARS CoV-2 3D website

Pocket predictions

Binding-site identification is essential for drug discovery development to design small molecules that can block target protein function. The program Fpocket was selected among other tools because it is easy to use and implement on the website. Furthermore, the Fpocket detection can guide the virtual screening implemented when the target protein is unsolved with a small molecule, for example, nsp12. Finally, the Fpocket tool was applied to our final modelled structure represented in the Models/PDB table to predict potential ligand and fragment binding sites. The

pocket racking scores are based on possible small-molecule ligand-binding hydrophobicity and polarity scores (Le Guilloux, Schmidtke and Tuffery, 2009).

Website development

The website backbone is very similar to COSMIC CGC 3D but with different designs and new tool implementations. The front-end was developed using HTML5, CSS, Bootstrap (version 4.5), and jQuery, whereas the back-end was developed using Expres.Js, a web application framework using Node.js. The database tables are stored in the PostgreSQL server and queried using the pg-promise library. The Embedded JavaScript (EJS) is a language that generates dynamic HTML. The entire proteome of SARS CoV-2 was represented in a dynamic sunburst chart created using the Plotly.js library, which is open source. This user-friendly representation allows users to query specific genes. In addition, the SARS CoV-2 human protein-protein interaction was represented using a network graph interaction viewer created by the Data-Driven Documents (D3.js) package.

The SARS CoV-2 database data, including protein-protein interaction data, the SARS CoV-2 modelled proteome, and available experimental structures, can be queried through programmatic access using RESTFUL APIs. In addition, the data are presented in a JSON object that can be parsed by other software.

Results

Website analysis

Home page

The SARS CoV-2 home page (Figure 5) is divided into two parts. The first includes a short description of the SARS CoV-2 database and the computational idea behind developing the database. The SARS CoV-2 computational approaches include proteome oligomer modelling, binding site prediction, mutational analyses, and docking hypotheses. The second part of the home page has SARS CoV-2 proteome represented as sunburst and a table with geneID and gene names. More information about each gene can be displayed by hovering over any SARS CoV-2 gene presented on the table. The database home page also includes the ability to query by clicking on any genes in the sunburst or the table. In addition, the navbar at the top of the page provides links to navigate to different pages, and the database logo returns users to the home page.

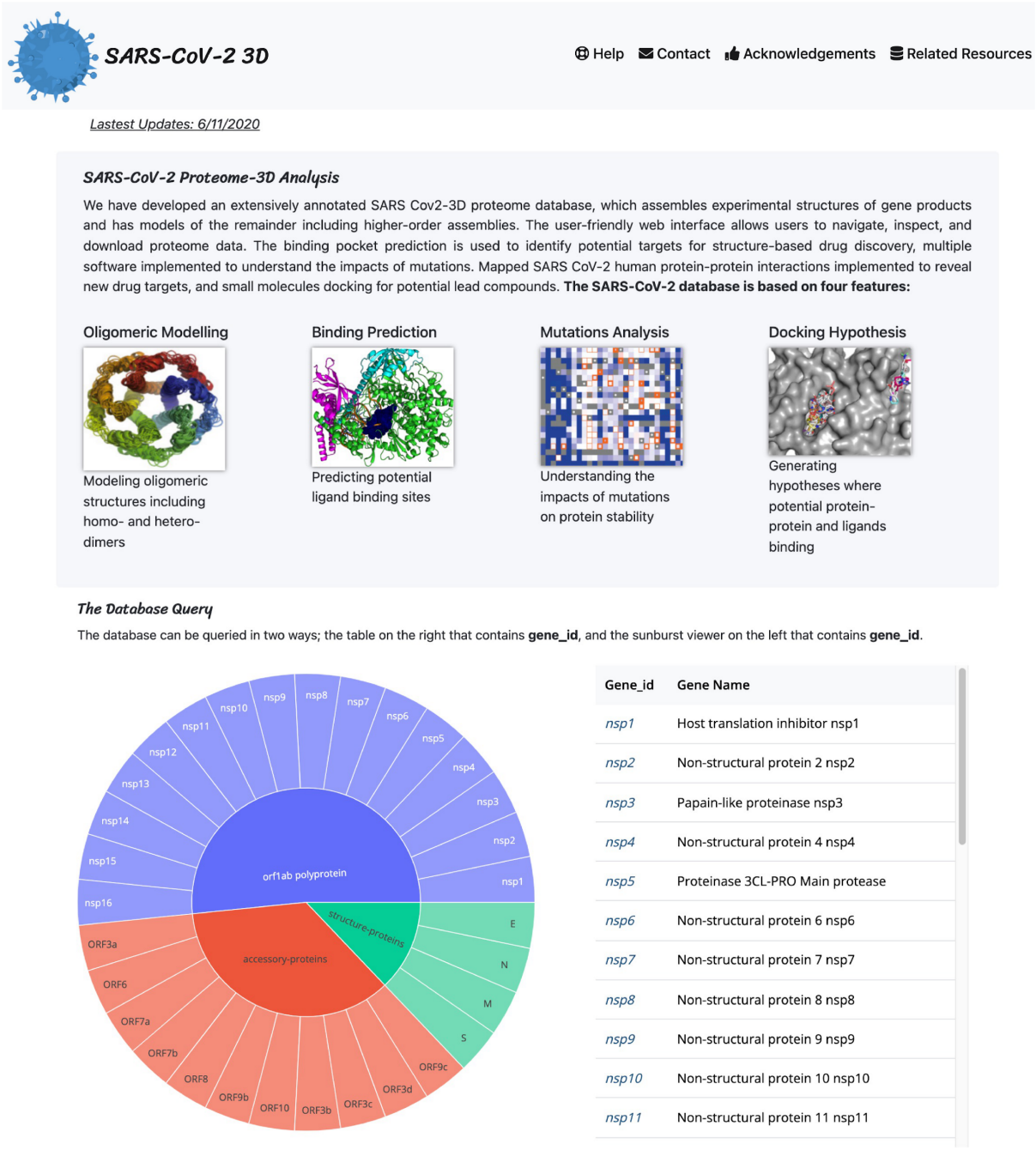


Figure 5. The SARS CoV-2 3D home page includes short descriptions of the database main features, including oligomeric modelling, binding prediction, mutation analysis and protein-ligand docking. In addition, the lower part of the home page includes the SARS CoV-2 proteome and a table listing each gene.

Result page

The SARS CoV-2 3D database results page (Figure 6) has all data about queried genes. It is organised in a very similar way as the COSMIC CGC 3D database. The results page includes a bar to search for other SARS CoV-2 proteins. Each gene has a short description of its function and is linked to other helpful modelling resources such as ITASSER, SWISS model, AlphaFold, and UniProt to compare other modelling structures. The Model/PDB table, represented under the gene name, includes important information about the model, such as selected templates, quality assessment of the model using MolProbity, calculated models and PDB coverage, oligomeric state of the built model, i.e. protomer or oligomers, oligomer interface assessment using Proteins, Interfaces, Structures and Assemblies (PSIA), and model information which includes the most reliable regions of the model.

Similarly to the COSMIC CGC 3D database, the MolStar viewer is implemented to visualise the SARS CoV-2 built models, experimentally solved structures, protein-protein docking and protein-ligand docking. MolStar has better features than other visualiser such as the NGL viewer, a web application for visualising macromolecular structures, map interaction between selected atoms such as hydrogen bonding, and clear visualisation of alpha and beta sheets. Representation on the top of the viewer of sequences of selected chains. Under the MolStar viewer is the pocket prediction table. The table includes pocket score, drug score, hydrophobicity and polarity scores. The table can be sorted based on user preference. ProtVista was implemented in a similar way to the COSMIC CGC 3D database. It is a powerful viewer, allowing the user to view gene annotation, such as domains.

All experimentally solved structures are represented in the PDB Hits table. This collapsible table will not present in query protein that has not been solved

experimentally. The mutations table includes mutation predictions from different tools described in the methods section. The mutation table can be sorted based on destabilising and stabilising for each tool presented. Mutations are represented as a wild-type residue, residue number in the reported PDB file, and mutant residue. The virtual screening ligands docking table describes the top five docked approved FDA drugs. The ligand names are responsive and linked to the DrugBank Database, which contains comprehensive information about drugs and drug-targets (Wishart *et al.*, 2018). The table can be ranked based on the Glide docking score, an empirical scoring function designed to separate ligands bond strongly from ligands bound weakly to the selected target active site. All the docked ligands can be viewed in the MolStar viewer on the website or downloaded locally. The virtual screening ligands docking table is designed so that when there is no virtual screen docking for a queried target, the table will not be shown, for example, nsp1. The last table on the result page defines the interactions of SAR CoV-2 and human proteins. The table includes docked human protein structure to the SARS CoV-2 target. The human PDB structure was mapped to the gene name and linked to the UniProt database. The lowest energy minimised structures can be viewed on the website. However, the entire 200 poses can be downloaded from the help page. The protein-protein interactions viewer shows the SARS CoV-2 gene as a node and each human protein target as an edge. Human genes with FDA drug targets are represented with black arrow indicators, whereas those without FDA drug annotation are represented in lighter indicators (Figure 6).

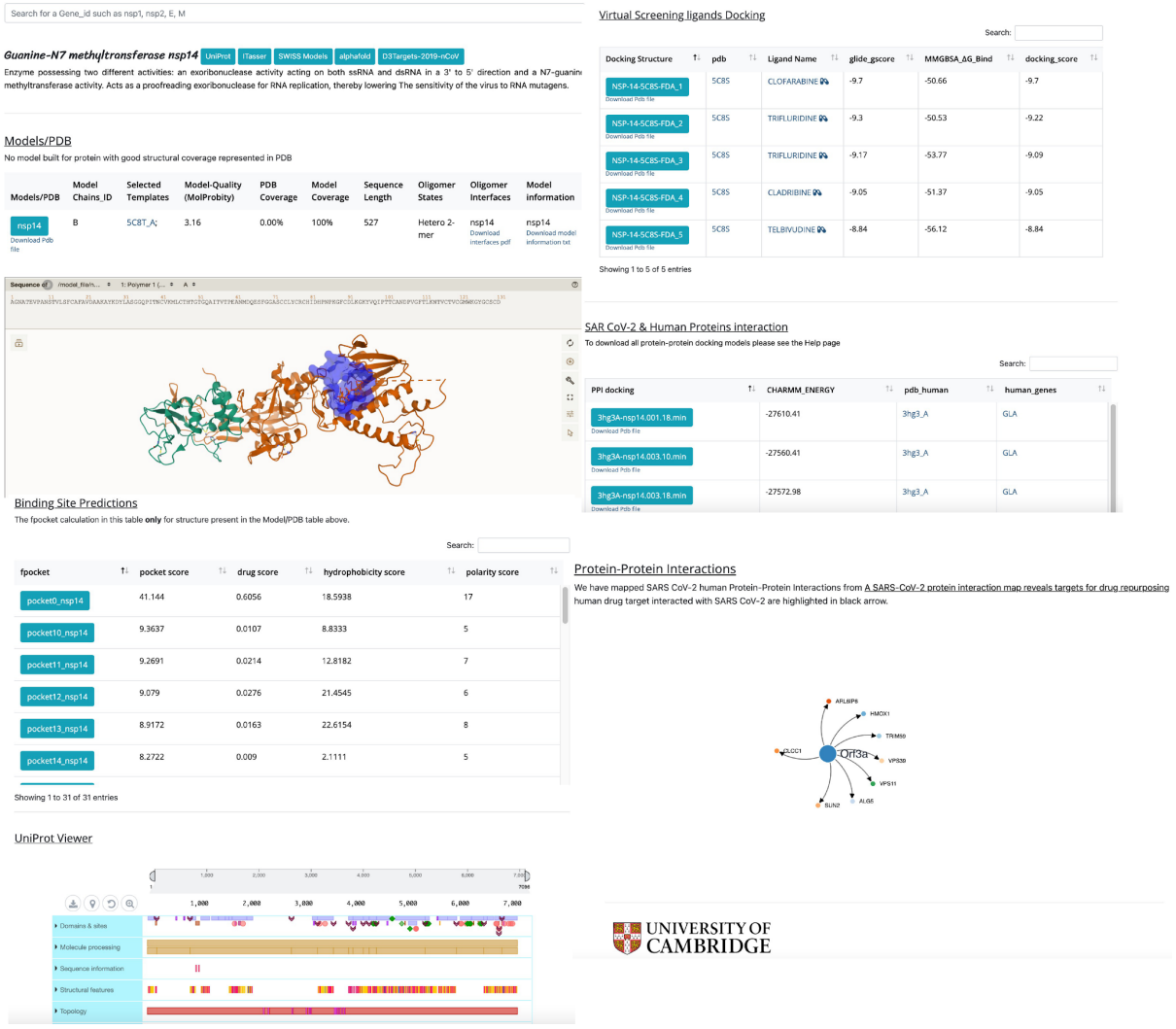


Figure 6. The result pages can be viewed from (https://sars3d.com/model/nsp14?showPocket=pocket0_nsp14). The data include: gene name with links to external data sources, including UniProt, ITasser, SWISS Models, AlphaFold, and D3Targets-2019-nCoV; a brief description about the gene; and a Models/PDB table, which provides information about the modelled and experimental structures. Mostar displays 3D structures of the target-gene models, and the UniProt viewer is used to visualise domains and other annotations. The virtual screening ligands docking table includes docked FDA antiviral drugs. The SARS CoV-2 $ human proteins interaction table includes docked SARS CoV-2 structure to human protein shown in the protein-protein interaction viewer.

Data statistics

The entire SARS CoV-2 proteome was built with close to 100% structural coverage. The modelled structures include not only protomers but also homo/heterodimer complexes where appropriate. The database was updated recently with four accessory proteins (ORF3d, ORF3c, ORF9c, and ORF3b); it is a continuing effort with regular updates. The papain-like protease (nsp3) was the largest built oligomeric model with 1945 amino acid residues, whereas the nsp11 is the smallest monomeric model with 13 amino acid residues. When the SARS CoV-2 3D database was developed, there were 15 proteins with complete/partial experimental structures with mean sequence coverage of 81.41%, whereas the modelled SARS CoV-2 structures have 97.5% structural coverage (Figure 7). The science for SARS CoV-2 is growing enormously fast. There are around 20,000 papers deposited in bioRxiv, which are only about Covid-19 work. This does not include published work that has not been deposited into bioRxiv. Therefore, the statistical evaluation reported here changes from time to time. For example, it now includes a new experimental structure, such as nsp14 in complex with (nsp14-nsp8-nsp7-nsp13-nsp9-nsp10) PDB ID;7EGQ, and nsp2 (PDB ID;7MSW) solved after the SARS CoV-2 3D database was released. In the section below, all newly solved structures are discussed and compared to built models.

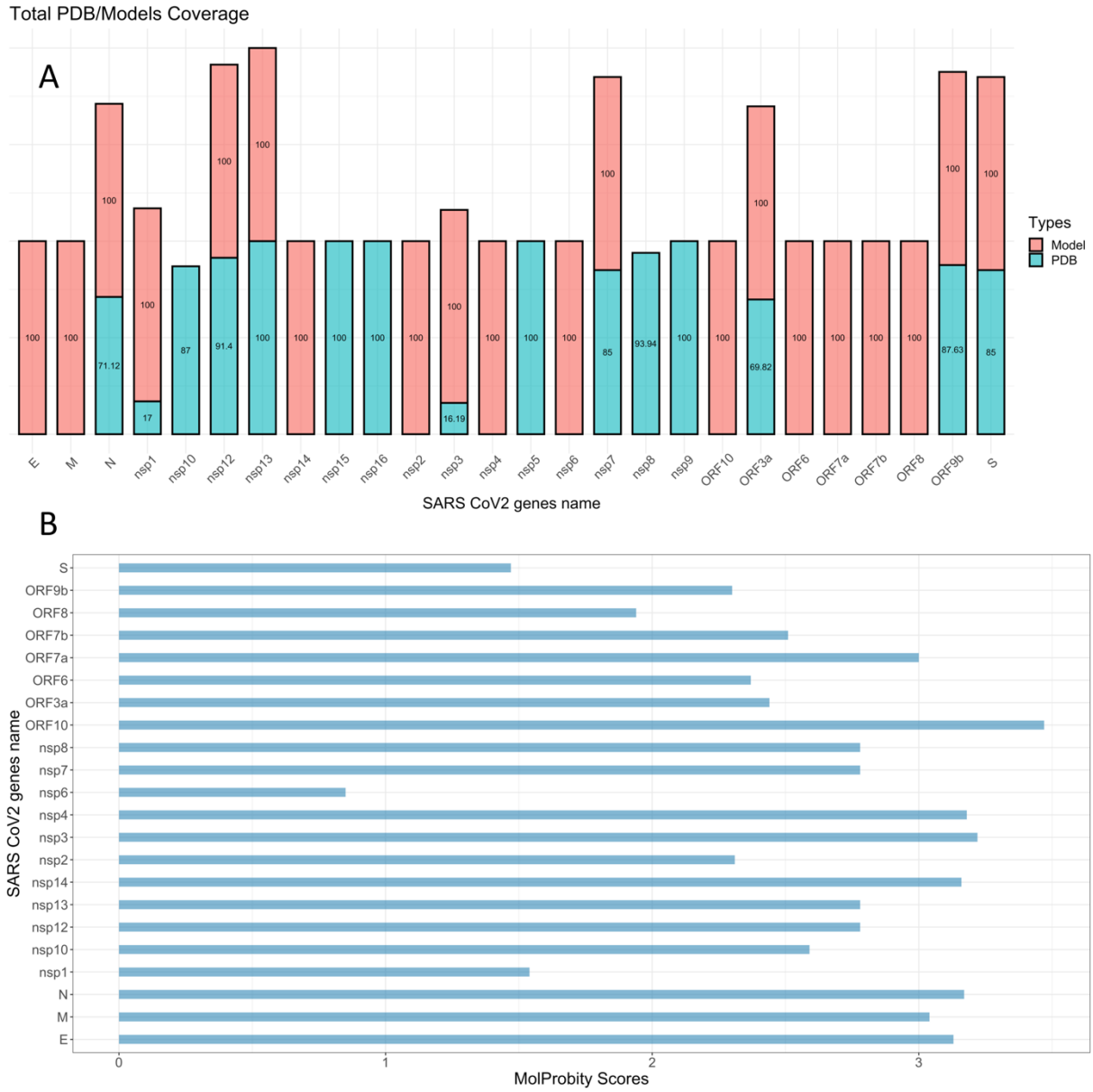


Figure 7. adapted from Alsulami et al. (2021). Statistical analysis of SARS CoV-2 modelled proteome. (A) Coverage of modelled structure is coloured in red and experimental structure is coloured in cyan. (B) MolProbity score of all modelled structures.

SARS CoV-2 3D Modelling protome

In this section, the discussion is mainly about SARS CoV-2 Orf1ab polyprotein and structural proteins. However, since accessory proteins classify as non-essential, these are described briefly.

Table 1 Modelled and experimentally solved structures of SARS CoV-2 proteins in the SARS CoV-2 3D database, including quality assessment score, model coverage, and experimental coverage.

Name	Model Acronym	Model Coverage.	MolProbity Score.	RCSB Structure available	RCSB Coverage.
Structural proteins					
Spike protein	S	100%	1.47	Yes	85%
Membrane protein	M	87%	3.04	No	-
Nucleocapsid protein	N	87%	3.17	Yes	71.12%
Envelope protein	E	76%	3.13	No	-
Non-structural proteins (Nsp)					
Nsp1	nsp1	100%	3.29	Yes	17.77%
Nsp2	nsp2	100%	2.31	No	-
Nsp3	nsp3	100%	3.22	Yes	16.19%
Nsp4	nsp4	100%	0.85	No	-

Nsp5	No Model	-	-	Yes	100%
Nsp6	nsp6	100%	0.85	No	-
Nsp7	nsp7	100%	2.78	Yes	85%
Nsp8	nsp8	100%	2.78	Yes	93.94%
Nsp9	No Model	-	-	Yes	100%
Nsp10	No Model	94%	2.59	Yes	81.29%
Nsp12	nsp12	100%	2.78	Yes	91.40%
Nsp13	nsp13	100%	2.78	Yes	100%
Nsp14	nsp14	100%	3.16	No	-
Nsp15	No Model	-	-	Yes	100%
Nsp16	No Model	-	-	Yes	100%
Accessory protein					
ORF3a	ORF3a	100%	2.44	Yes	69.82%
ORF6	ORF6	100%	2.37	No	-
ORF7a	ORF7a	100%	3.00	No	-
ORF7b	ORF7b	100%	2.3	No	-
ORF8	ORF8	100%	1.94	No	-
ORF9b	ORF9b	100%	2.3	Yes	87.63%
ORF10	ORF10	100%	3.47	No	-

Host translation inhibitor nsp1

The nsp1 of SARS CoV-2 has 84% sequence identity to that of SARS CoV-2, indicating similarity in function. The nsp1 protein suppresses the host innate immune system by inhibiting the antiviral defence mechanism (Thoms *et al.*, 2020). In addition, it mediates translation inhibition by interacting with mRNA channels. It binds to 43s pre-initiation complex and non-translating 80s ribosome through a small C-terminal domain covering the region between 148-180 (Figure 8). Both N-terminal and C-terminal domains have been solved experimentally (Thoms *et al.*, 2020)(Semper, Watanabe and Savchenko, 2021).

The nsp1 structure was modelled based on three templates PDB ID (2GDT_A, 5C5S_A, 6ZLW_i). The modelled nsp1 N-terminal perfectly aligns with the experimentally solved structure (PDB ID; 7K7P) with a TM-align score of 0.8. The template modelling score or TM-score measures similarity between two optimally superimposed protein structures. A TM-score of 1 indicates two structures are perfectly aligned, whereas a score of 0 indicates the two structures do not match. The perfectly aligned 3D structure between the generated model and experimentally solved structures indicates that carried modelling approach is reliable. Therefore, the modelled structures in the SARS CoV-2 3D database can be used to generate other hypotheses, such as those from mutational analysis and docking.

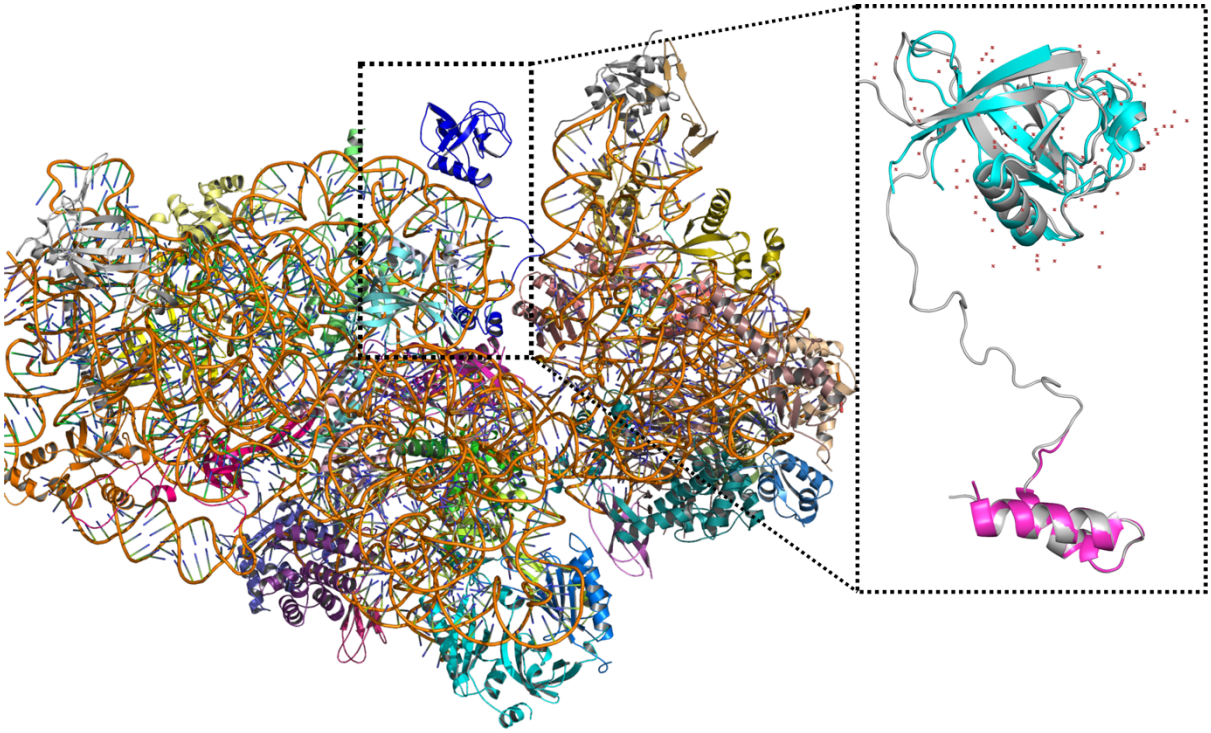


Figure 8. Modelled nsp1 structure coloured in blue with 40S ribosome. Two experimental structures were solved for nsp1 coloured in magentas, and cyan is overlaid to the modelled nsp1 coloured in white.

Non-structural protein 2 nsp2

The exact function and roles of nsp2 inside the host are still not understood. However, the newly solved crystal structure reveals that nsp2 may regulate intracellular signalling pathways by interacting with nucleic acid since it contains zinc fingers (C2CH2, C4, and C2HC types). Furthermore, nsp2 is associated with RTC and may play an essential role in translation. The entire structure was modelled using five template PDB IDs;(5F22_B, 3LD1_A, 5Y81_H, 1R7G_A, and 1G03_A). Interestingly, the nsp2 four domains are perfectly modelled (Figure 9). However, the orientations of domains relative to each other are not accurately predicted. The retention of the overall fold can be assessed by comparing the modelled nsp2 structure side by side with the experimentally solved structure. (Figure 9)

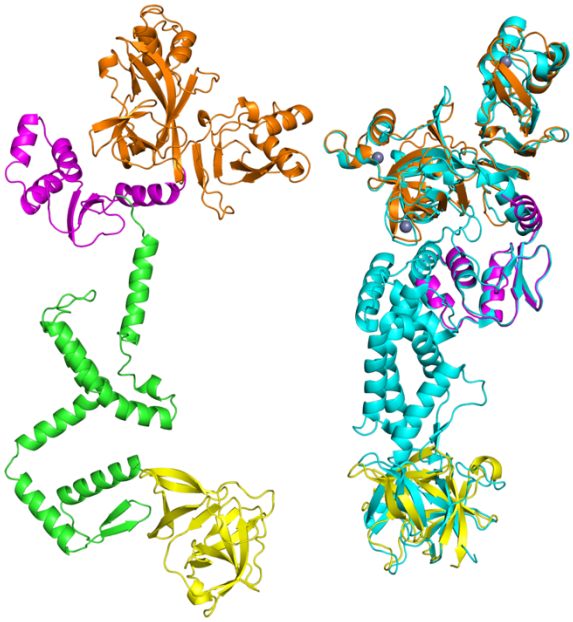


Figure 9. Modelled nsp2 structure on the left and experimentally solved structure on the right (PDB ID:7MSW). Each modelled domain is coloured differently. The domains overlay perfectly to the experimental structures except in the region coloured in green.

Papain-like protease nsp3

The extensively studied essential SARS CoV-2 protein, nsp3, enables viral spreading by cleaving the polyprotein to generate the functional replicase complex. Although the nsp3 protein shares 83% sequence identity with SARS CoV-1, it exhibits a different function in regulating the host interferon through the interferon-stimulated gene 15 protein (ISG15). The papain-like protease is the largest SARS CoV-2 protein with 1944 amino acid residues. There is no solved structure for the complete protomer. However, the SARS CoV-2 3D database has a complete homo 3-mer structure. The model structure was built based on 12 template PDB IDs (2GRI_A, 6VXS_A, 2W2G_A, 6W9C_A, 2K87_A, 3GA8_A, 1YX1_A, 6ORH_B, 1QWG_A, 3C8F_A, 1HA8_A, 1HUP;). Only two regions have been solved experimentally (PDB ID 6VXS); these are the region between 206-374 as a monomer and the region between 746-1060 (PDB ID 6W9C) as a homo 3-mer. One

of the challenging aspects of building complexes is predicting the oligomeric state of the protein. The nsp3 is a perfect example, one of the domains is solved as homo 3-mer, and the other is solved as a monomer. The final modelled nsp3 oligomer is a homo 3-mer based on nsp3 experimentally solved structure PDB ID 6W9C. (Figure 10)

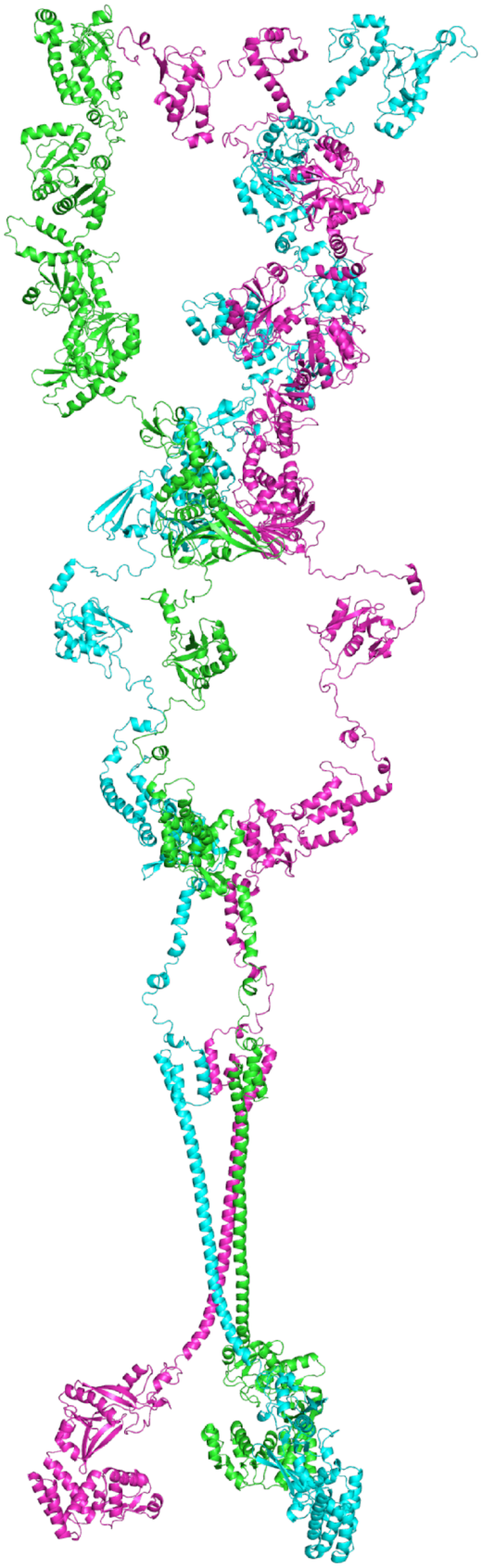


Figure 10. Modelled Homo 3-mer nsp3 structure each protomer coloured differently.

Non-structural protein 4 nsp4

The non-structural protein 4, nsp4, is a transmembrane protein located in the endoplasmic reticulum. It is essential for viral replication and double-membrane vesicle formation. The nsp4 modelled structure was built based on four templates, PDB IDs: 1BCP_F, 3VC8_A, 3A7K_A, and 1T70_A. The transmembrane region is well defined using the Orientations of Proteins in Membranes (OPM) database. The model final MolProbity score is 3.18. (Figure 11)

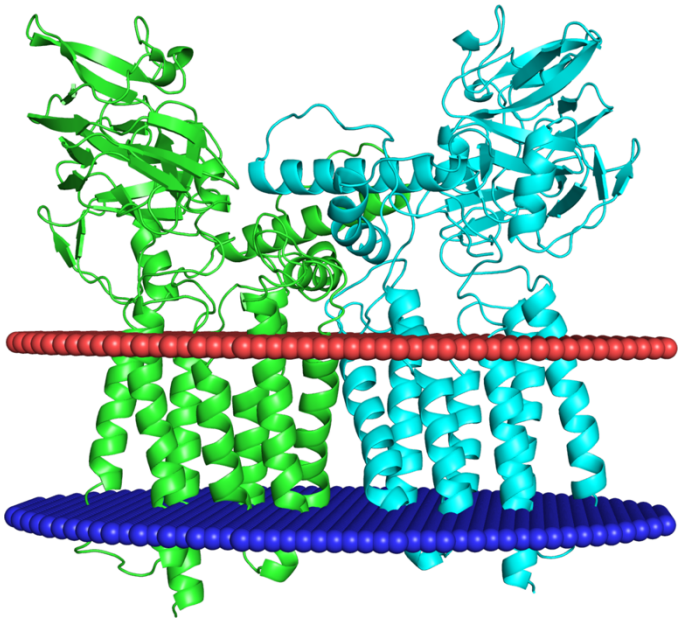


Figure 11. Modelled homodimer nsp4 structure coloured in green and cyan, the transmembrane regions are highlighted between circular blue and red atoms.

Non-structural proteins 7,8, 12, and 13 complexes.

The RTC complex plays an essential role in the SARS CoV-2 life cycle. The modelled nsp7,8,12 and 13 oligomeric complex was built based on PDB ID: 6XEZ, SARS-CoV-2 replication-transcription complex bound to nsp13 helicase - nsp13(2)-RTC. Each protein was processed individually to obtain the complete structure coverage, then superimposed on the solved structure PDB ID: 6XEZ to obtain the full coverage of the mini RTC complex. The final oligomeric complex also includes metal

ions and ligands. (Figure 12) The modelled complex was built with a reasonable quality assessment MolProbity score of 2.7, which indicates that it could be used for further mutational analysis studies and molecular docking

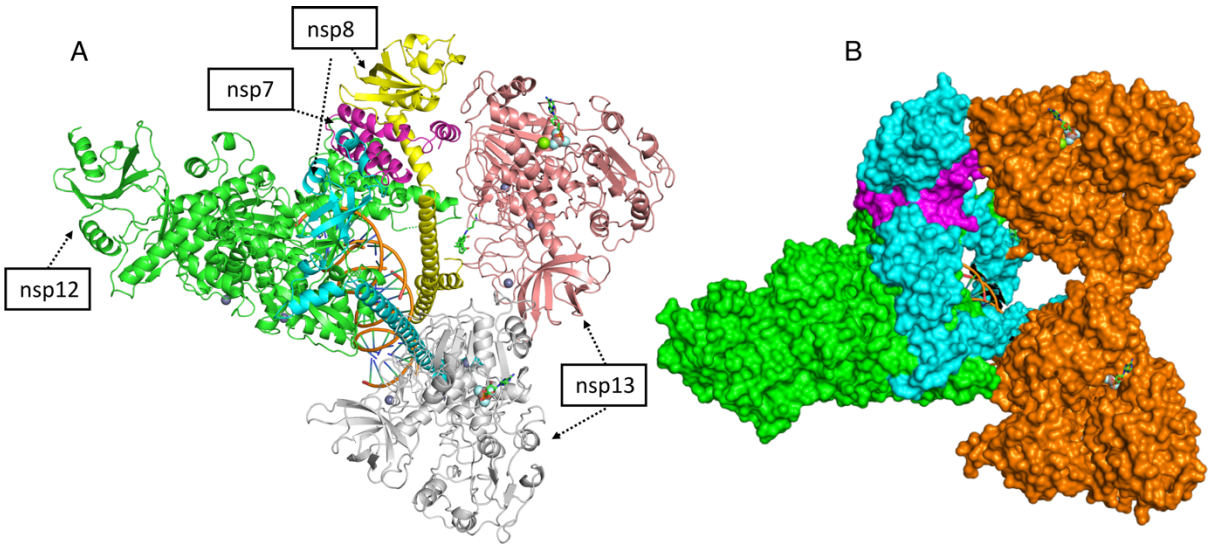


Figure 12. Modelled mini RTC complex. (A) The nsp12 polymerase is shown in green in complex with homodimer of nsp13 helicase shown in white and light red. The homodimer of nsp8 is highlighted in yellow and cyan, whereas nsp7 is shown in magenta. The ligand is shown in stick green, metal ions in spheres and DNA in orange. (B) The top view of the mini RTC is represented as a molecular surface model.

Non-structural protein nsp14 and nsp10 complex

nsp14, one of the proteins most conserved throughout the coronavirus family, is considered an essential target for SAR CoV-2. Moreover, it plays an important role in the RTC complex. The nsp14 includes N-terminal exoribonuclease (ExoN), which interacts with nsp10 and the C-terminal N7 methyltransferase domain. The nsp14 was built as a heterodimer in a complex with nsp10, metal ions, and ligand. Two templates were selected to build the oligomeric modelled structure PDB ID (5C8T, 5C8S) with a MolProbity score of 3.16. A new nsp14 experimentally solved structure has been deposited into the Protein Data Bank as PDB ID (7EGQ) (Yan *et al.*, 2021).

This allows discussion and comparison of the modelled nsp14-nsp10 complex with the experimentally determined structure. Superimposition of the nsp14-nsp10 oligomeric modelled structure to the experimentally solved structure indicates that the same fold is retained and that the predicted nsp14 and nsp10 interface is identical. The TM-score for the two structures is 0.95, meaning that the modelled structure overlaps well with the experimental structure except for the flexible loop in the C-terminal, which can be in different conformations even for experimentally solved structures (Figure 13).

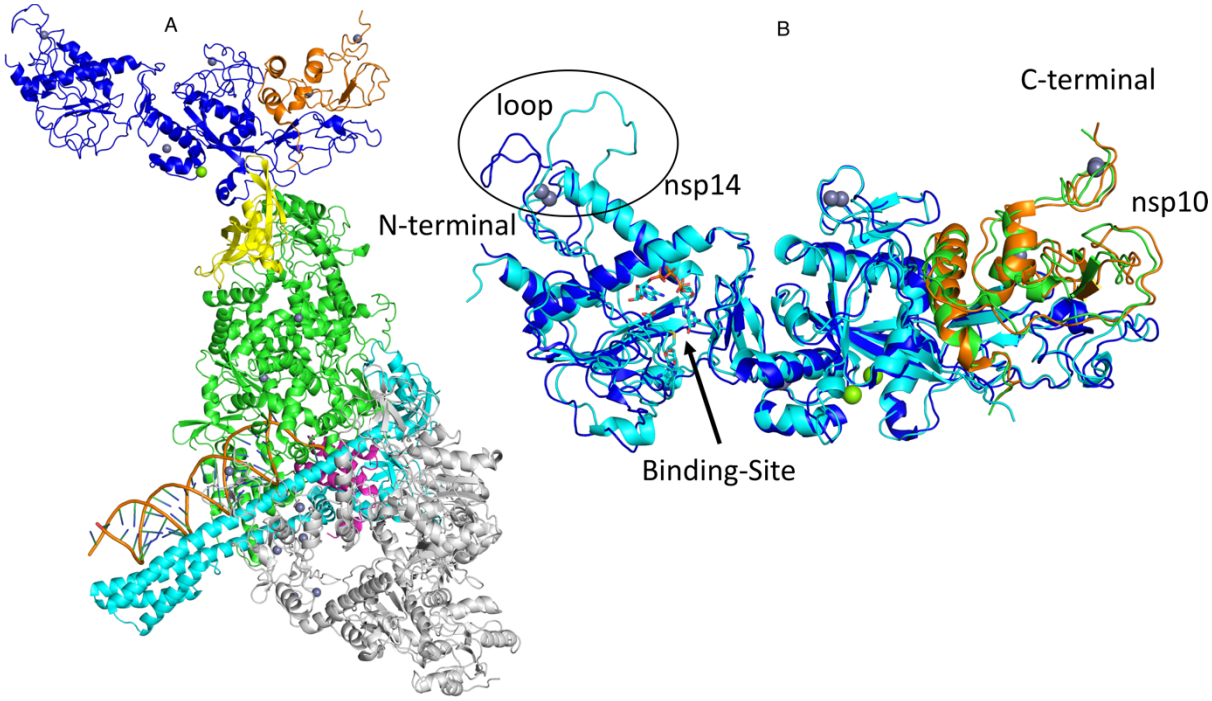


Figure 13. Comparing the modelled and experimental structure of nsp14. (A) Solved nsp14 coloured in blue with experimentally mini RTC shown in figure 12. (B) Modelled nsp14-nsp10 complex in cyan and green with experimentally solved nsp14 coloured in blue and nsp10 coloured in orange.

Envelop small membrane protein E

The envelope protein, one of the smallest SARS CoV-2 transmembrane proteins, forms a homo 5-mer cation channel. Inhibition of E protein reduces the viral loads and budding. The topology domain of E protein consists of virion surface residues 1-

13, transmembrane region residue 14-34, and intravirion region residue 35-75. Newly NMR solved structure PDB ID (7K3G) includes residues 8-38. The E protein was modelled using a ProtCHOIR automated pipeline to build homodimer structures, covering the amino acid region between 8-65. The selected template was from SARS CoV-1 PDB ID (5X29), and the model MolProbity score is 3.22. Comparing the two structures, the overall fold is retained. However, the TM-score between the E modelled protein, and experimentally NMR structure is 0.48. Usually, less than a 0.5 TM-score indicates the two structures do not overlap well. The NMR structure shows the five-helix diameter narrower than the predicted model structure. (Figure 14)

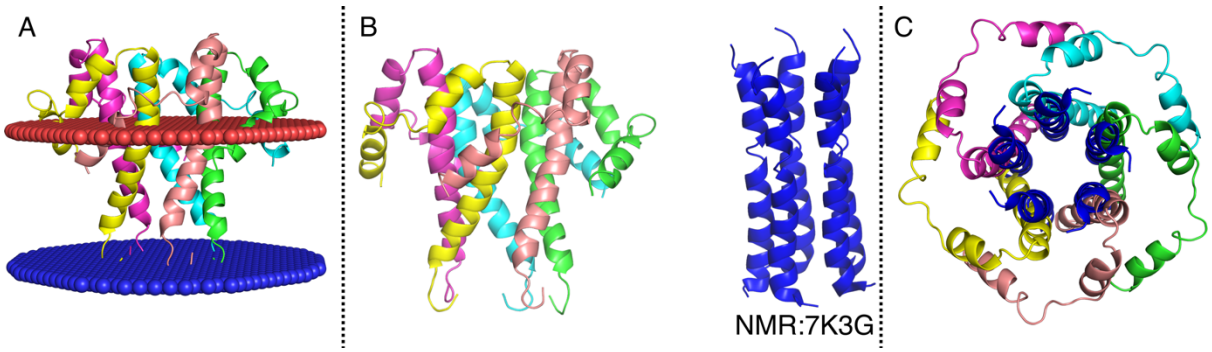


Figure 14. Comparing modelled and experimental structures of E protein. (A) modelled homo 5-mer E protein-based on SARS-1, the membrane region highlighted between circular red and blue atoms. (B) Side view of experimentally solved E protein in blue and modelled protein in SARS CoV2 3d D database. (C) Top view of overlaid modelled structure on the experimental structure, coloured as in B.

Membrane protein M

Membrane protein M in SARS CoV-2 is located between the spike proteins in the viral envelope. It is one of the most abundant proteins in SARS CoV-2, facilitating virus budding. The M protein in SARS CoV-2 is closely related to SARS CoV-1 with 98.6% sequence similarity. It includes three transmembrane helices and short amino acid sequences outside the envelope. Unfortunately, the M protein is still not solved

experimentally. The existing modelled structure includes the three transmembrane regions as annotated by UniProt and the long carboxy-terminal. The M protein is modelled as homo 2-mer based on PDB ID (3A7K_A, 5UTT_A, 6SPB_V, and 6XDC), yielding a model with a 3.04 MolProbity score. The final modelled structure covers the amino acid region between 10-203, with 87% structural coverage. (Figure 15)

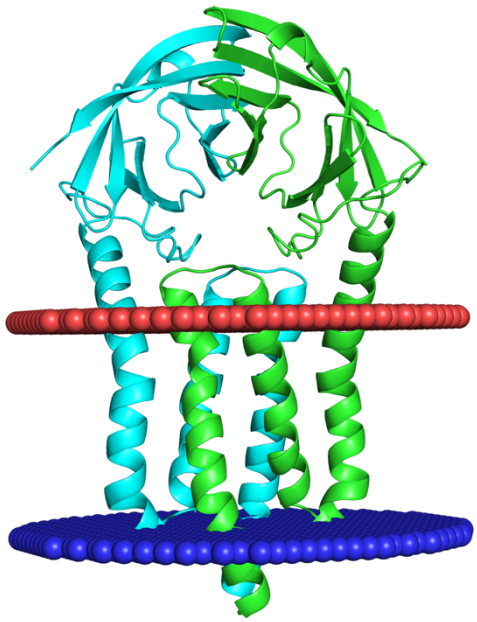


Figure 15. Modelled structure of M protein. The predicted modelled M protein as a homodimer. The transmembrane regions are highlighted between circular red and blue areas.

Accessory protein ORF3a

The ORF3a is one of the putative ion channels in SARS CoV-2 along with E protein. Plasma samples from patients observed ORF3a antibodies. The deletion of ORF3a in mice results in decreasing morbidity (Yan *et al.*, 2021). In addition, the ORF3a has implications in viral release, autophagy inhibition and cell death. The above points

suggest that ORF3a could be an essential target for vaccine or small molecule development.

The ORF3a was solved experimentally (PDB ID:6XDC) with structural coverage of 69.8%. The final modelled homo 2-mer structure was deposited into the SARS CoV-2 with 100% structure coverage with a MolProbity score of 2.44. The model was built based on PDB ID (6XDC, 4BKW_A, 5ZBE_A, 5L2F_A). (Figure 16)

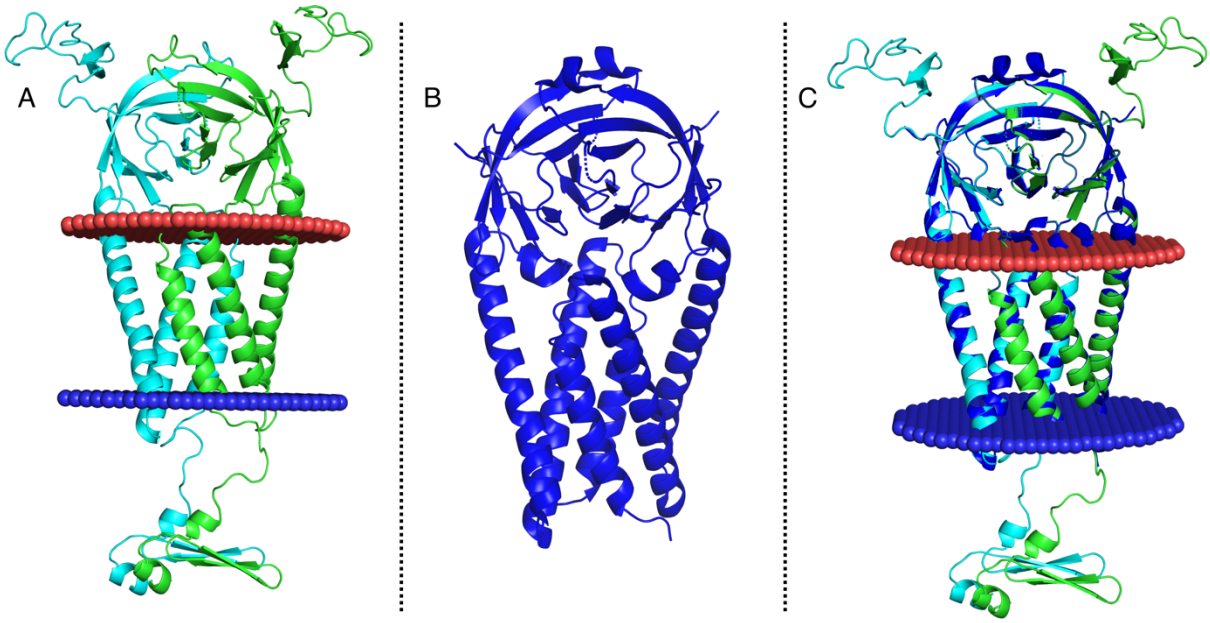


Figure 16. Comparing modelled and experimental ORF3a protein. (A) modelled homodimeric ORF3a protein, the membrane region highlighted between circular red and blue surfaces. (B) Side view of experimentally solved ORF3a protein in blue PDB ID:6XDC. (C) Modelled ORF3a protein in SARS CoV2 3d D database overlaid with the experimental structure in blue.

There are no modelled structures for nsp5, known as 3C-like proteinase or main protease, nsp9, nsp10, nsp15, nsp16, and ORF8 since the complete 3D structures were solved experimentally. Furthermore, AlphaFold did exceptionally well in predicting the nsp6 3D structure. Therefore, the AlphaFold modelled structure is incorporated into the SARS CoV-2 3D database

Mutational analysis example

Single-nucleotide substitutions can result in amino acid changes that impact the protein function and structure. My colleague Christopher A Beaudoin ran several tools such as mCSM-Stability, mCSM-PPI, I-Mutant, Provean, and DeepDDG for mutations in seven SARS CoV-2 targets. All the results are presented on the SARS CoV-2 3D website. Since then, many mutation data have been reported for all SARS CoV-2 targets in the CNCB-NGDC database (https://ngdc.cncb.ac.cn/ncov/variation/annotation?lang=en). For example, looking at SARS CoV-2 structural proteins:

- Envelope protein has 326 unique missense mutations.
- Membrane protein has 831 unique missense mutations.
- Nucleoprotein has 1947 unique missense mutations.
- Spike protein has 5681, of which 53 mutations were frequently reported three times, 368 mutations were frequently reported two times, and the rest were reported only once.

The spike protein is the main target in SARS CoV-2 and receives a lot of attention computationally and experimentally. The spike protein so far has 594 solved structures in PDB. However, the complete structure with the transmembrane region is still not solved. Therefore, the final complete modelled structure of the spike includes both open and close confirmation.

There are 5462 mutations reported for the spike protein with structural representation in the SARS CoV-2 database. The mCSM-Stability, SDM, and mCSM-PPI were used to characterise and evaluate putative destabilising and

stabilising mutations. Most spike mutations are predicted to be relatively neutral or be mildly deleterious. However, there are 168 mutations predicted to be highly destabilised by mCSM-Stability, whereas no mutations are predicted to be highly stabilising. On the other hand, 309 mutations are highly destabilising, and 42 mutations are highly stabilised by SDM. In addition, mutations impacting the spike homo 3-mer interface were predicted using mCSM-PPI, and there are 35 mutations predicted to be highly destabilising. (Figure 17)

Due to different software methods used, i.e. mCSM is machine learning-based, and SDM is a statistically-based method, many spike mutation predictions disagree. However, 75 mutations are predicted to be high destabilising by both mCSM-Stability and SDM, 9 mutations are highly destabilising by both mCSM-PPI and mCSM-stability, and one mutation is predicted to be high destabilising by both mCSM-PPI and SDM.

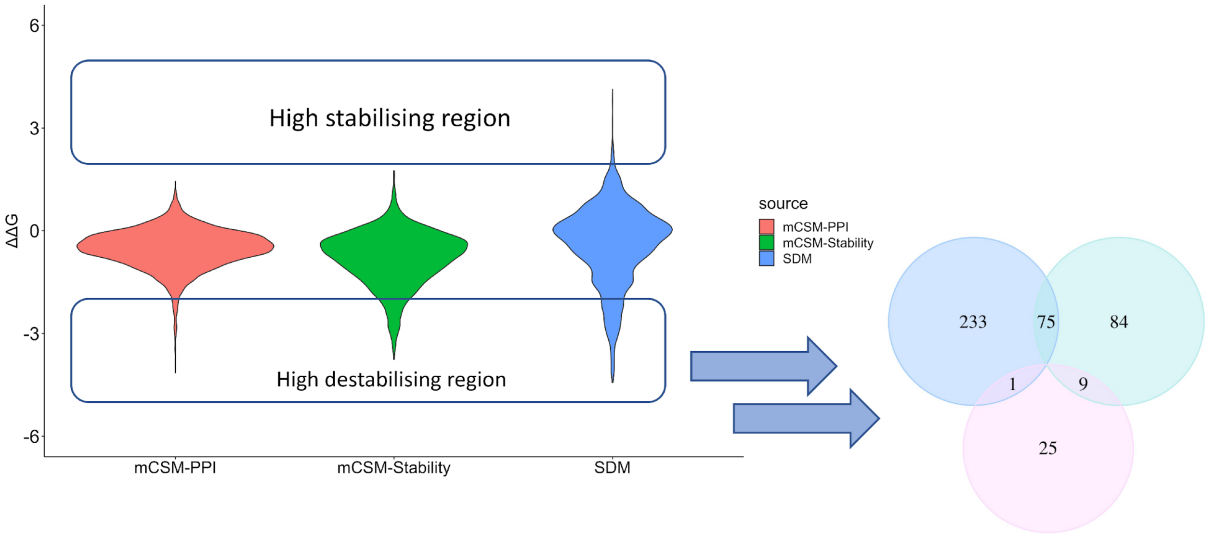


Figure 17. Analysis of spike mutations. The diagram represents 5462 mutations coloured differently. The software used is shown in the x-axes, and the y-axis represents the $\Delta\Delta G$ value. The highly destabilising mutations are drawn in the Venn diagram with the same colours: mCSM-PPI light red, SDM light blue, and mCSM-stability light green. Mutations predicted to be highly destabilising by different tools are represented between the circles.

All frequent mutations (412 mutations) are predicted to be moderate. However, most of these mutations appear on the interface. Although the impact of these mutations cannot be explained by considering the spike protein itself, solving the spike structure in a complex with other proteins will undoubtedly determine the effects of these mutations on binding interaction. For example, PDB ID (7DHX) shows the Interaction between AEC-2 human and RBD spike SARS CoV-2. Mutations reported three and two times are mapped to the RBD solved domain (Figure 18). Mutations tend to work in a synergic way. It is unlikely that one mutation will predominate and give a selective advantage. However, a few mutations, such as D614G, have a virus phenotype effect, i.e. increasing pathogenicity and transmissibility. Mutations that impact immune recognition by weakening antigenic interaction require immediate attention. Putative spike mutations shown in (Table-1 supplementary) need further experimental studies to confirm which mutations influence antigen recognition.

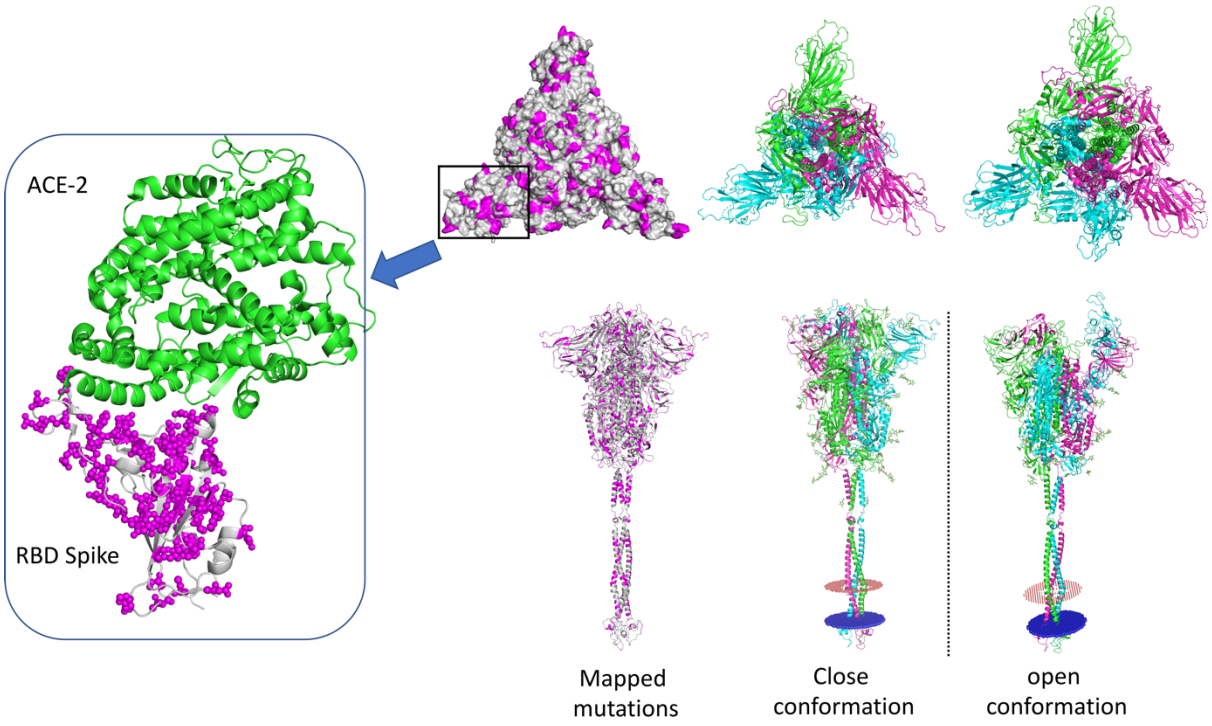


Figure 18. modelled Spike protein in open and closed conformations. Each domain is coloured differently, and the transmembrane region is highlighted between circular red and blue atoms. All frequent mutations are mapped to the spike closed conformation coloured in magenta. Mutations at the RBD and RBM are shown in magenta spheres.

Molecular docking example

Drug repurposing is one of the fastest approaches to bringing anit-SARS CoV2 compound to the market since these drugs have already been shown to be safe. All the drug docking deposited into the SARS CoV-2 3D database was performed by Dr Sundeep Chaitanya Vedithi. The docking focused on two critical stages of the infection: virus-cell interaction and virus replication. SARS CoV-2 targets such as Nsp3, Nsp5, Nsp12, Nsp14, Nsp15, Nsp16 and S were selected to perform molecular docking. Drugs with the top docked Glide score were selected and represented on the website. All docked PDB files are available to be downloaded from the website.

Furthermore, the interaction between the docked ligand and protein can be viewed by clicking on the ligand on the MolStar viewer. For example, the top docked hit for mini RTC complex is Mitoxantrone, an anticancer drug used to treat Acute Lymphoblastic Leukemia, also known to intercalate in DNA and RNA. (Figure 19). Unfortunately, there is no experimental structure with a ligand to compare the docked ligand pose. In addition, there are no data to validate that the docked ligands are effective against virus replication. Therefore, all docked drugs presented were proposed as a hypothesis to help and guide the potential discovery of new hit or lead compounds.

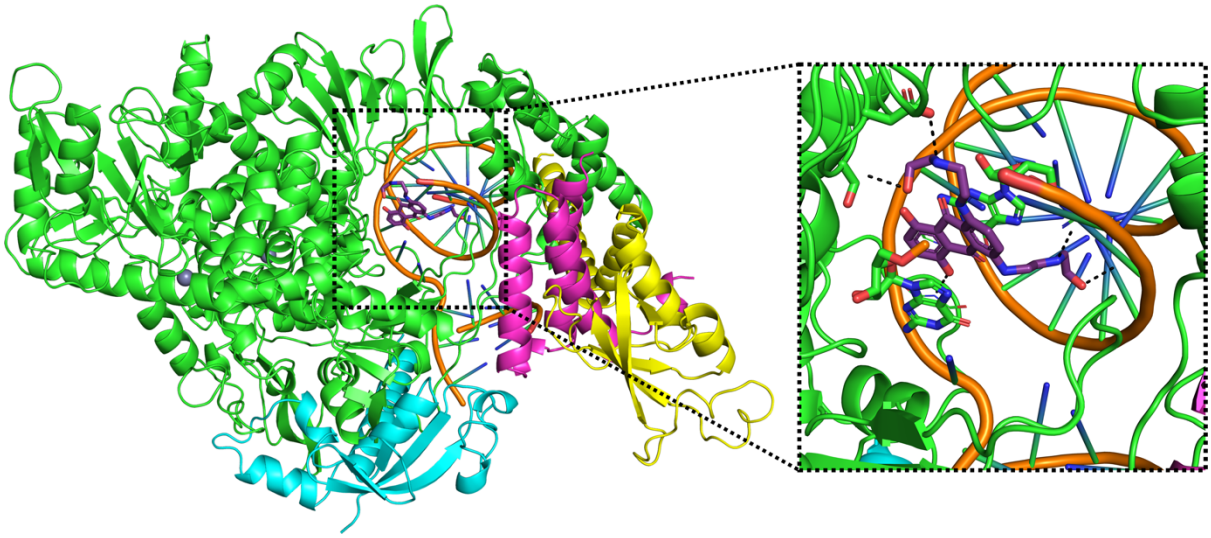


Figure 19. Protein-ligand docking. The docked Mitoxantrone into a mini RTC complex was performed using PDB ID: 7C2K. The nsp12 is coloured in green, nsp8 in yellow and cyan, whereas nsp7 is in magenta, the DNA has shown orange. The docked Mitoxantrone inserted between the DNA base pairs interaction with the neighbouring residues shown in black dashes.

Protein-protein docking example

Identifying essential interactions between human and virus proteins is another way of developing drug repurposing. Since these interactions could perturb the host network interaction or use the host machinery to proliferate. Arian R Jamasb performed all the human-viral interaction docking. As a result, there are 308 experimentally confirmed viral-human interactions. These interactions can be viewed in the protein-protein interactions section at the end of the result page. There are 200 poses for each docked structure. Only structures with a low score of CHARMM22 were selected and presented on the website. However, the rest can be downloaded from the help page and viewed locally, as explained in the Methods section. One of the protein-protein docking examples is nsp14, which interacts with two drug targets,

alpha-galactosidase (GLA) and Inosine-5'-monophosphate dehydrogenase 2 (IMPDH2). (Figure 20).

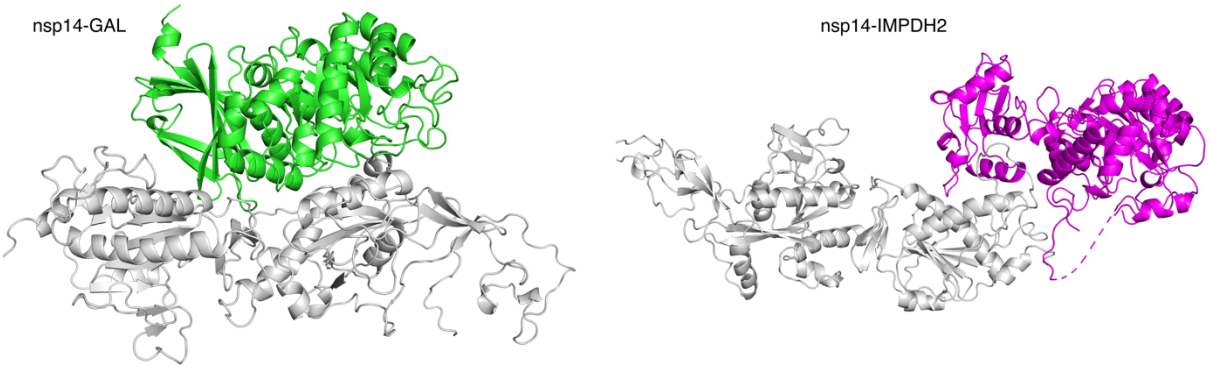


Figure 20. virus human protein-protein docking. Docked structure of nsp14-GAL and nsp14-IMPDH2. The lowest energy poses of GAL and IMPDH2 were shown here and coloured in green and magenta accordingly.

Discussion

A lot of scientific research has focused in the past year on SARS CoV-2 infections, mechanism, function, and treatment. Multiple databases have been updated or created a specific repository for SARS CoV-2 data. Proteomic databases such as RCSB PDB provide a specific archive for 3D structures experimentally solved by X-ray, Cryo-EM, and NMR. Furthermore, a domain-based database such as CATH and SCOP classified experimental SARS CoV-2 structure into domains and families. UniProt has been updated to include SARS CoV-2 annotations such as function, interaction etc., from a different resource. Finally, the GISAID database updates the new SARS CoV-2 lineage.

Multiple successful databases have been developed using Vivace's in-house pipeline tools, such as Chopin (Ochoa-Montaño, Mohan and Blundell, 2015), for the *M.tuberculosis* proteome and Mabellini for *M.abscessus* (Skwark *et al.*, 2019). While the SARS CoV-2 3D (Alsulami *et al.*, 2021) was being developed, other databases that informed modelling were developed, such as I-TASSER, which provides a set of monomeric models for the SARS CoV-2 proteome. Coronoavirus3D (https://coronavirus3d.org/#/) mapped a set of mutations to experimental structures in PDB. COVID-3D using I-TASSER models and performed mutational analysis (Portelli *et al.*, 2020). D3Targets-2019-nCoV is a molecular docking site to perform virtual screening docking SARS CoV-2 targets (Shi *et al.*, 2020)

The SARS CoV-2 3D database is complementary to all the previous works. It includes modelling mutations analysis, pocket detection, protein and ligand docking. In addition, it is the first oligomeric modelling database for SARS CoV-2 proteome with high structural coverage and quality assessment score. The SARS CoV-2

oligomeric modelled includes ligand cofactor clear transmembrane annotations. SARS CoV-2 3D database has most of the computational predictions as well as experimental structure in one database. As a result, the end-user can navigate visualised and analysed results in a straightforward manner to build a new hypothesis.

The most challenging part in protein oligomeric and multidomain protein modelling is predicting the relative orientation of multidomain protein built with multiple templates, for example, nsp2. However, this issue can often be overcome with the use of the cryo-EM, which often helps indicate the relative position of domains, even at low resolutions. Protein-ligand docking can be enhanced by considering other tools such as MolSoft to compare predicted poses of ligands on the target active site. Protein-protein docking could be enhanced by looking at conservation and hydrophobic residues on the interface and directly docking towards these regions.

Conclusion

The SARS CoV-2 3D is a comprehensive resource for the SARS CoV-2 proteome, including proteome modelling, mutational analysis, protein-protein docking, protein-ligand docking, and experimentally solved structures. Since the SARS CoV-2 has a very small proteome with 31 genes, all the modelled genes are built with explicit annotations such as transmembrane. Furthermore, oligomeric models were built when templates were available. SARS CoV-2 3D database also includes links to other modelling resources such as SWISS-MODEL, ITASSER, and AlphaFold to compare built models.

The entire new website based on Node.js was built with new functionality such as MolStar viewer to view 3D structures, ProtVisa to visualise domains and other UniProt annotations, and 2D viewer to visualise the human virus protein interaction, and RESTFUL API to retrieve data programmatically. The goal of the SARS CoV-2 3D website is to be computationally comprehensive to help the drug discovery process to identify and evaluate potential SARS CoV-2 drug targets.

Reference

Abdelrahman, Z., Li, M. and Wang, X. (2020) 'Comparative Review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A Respiratory Viruses', *Frontiers in Immunology*, 11(552909), pp. 1–14.

Alexandersen, S., Chamings, A. and Bhatta, T. R. (2020) 'SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication', *Nature Communications*, 11(1), pp. 1–13.

Alsulami, A. F. *et al.* (2021) 'SARS-CoV-2 3D database: Understanding the coronavirus proteome and evaluating possible drug targets', *Briefings in Bioinformatics*, 22(2), pp. 769-780.

Andersen, K. G. *et al.* (2020) 'The proximal origin of SARS-CoV-2', *Nature Medicine*, 26(4), pp. 450–452.

Ban, T., Ohue, M. and Akiyama, Y. (2018) 'Multiple grid arrangement improves ligand docking with unknown binding sites: Application to the inverse docking problem', Computational Biology and Chemistry, 73(2018), pp. 139–146.

Boni, M. F. *et al.* (2020) 'Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic', *Nature Microbiology*, 5(11), pp. 1408–1417.

Brooks, W. H. *et al.* (2008) 'Computational validation of the importance of absolute stereochemistry in virtual screening', *Journal of Chemical Information and Modeling*, 48(3), pp. 639–645.

Cao, H. *et al.* (2019) 'DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks', *Journal of Chemical Information and Modeling*, 59(4), pp. 1508–1514.

Capriotti, E., Fariselli, P. and Casadio, R. (2005) 'I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure', *Nucleic Acids Research*, 33(suppl_2), pp. 306–310.

Carrat, F. *et al.* (2021) ‘Evidence of early circulation of SARS-CoV-2 in France: findings from the population-based “CONSTANCES” cohort’, *European Journal of Epidemiology*, 36(2), pp. 219–222.

Cavalcanti, A. B. *et al.* (2020) ‘Hydroxychloroquine with or without Azithromycin in Mild-to-Moderate Covid-19’, *New England Journal of Medicine*, 383(21), pp. 2041–2052.

Chen, G. *et al.* (2020) ‘Clinical and immunological features of severe and moderate coronavirus disease 2019’, *Journal of Clinical Investigation*, 130(5), pp. 2620–2629.

Choi, Y. and Chan, A. P. (2015) ‘PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels’, *Bioinformatics*, 31(16), pp. 2745–2747.

Cui, J., Li, F. and Shi, Z. L. (2019) ‘Origin and evolution of pathogenic coronaviruses’, *Nature Reviews Microbiology*, 17(3), pp. 181–192.

Dai, W. *et al.* (2020) ‘Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease’, *Science*, 368(6497), pp. 1331–1335.

Douguet, D. (2018) ‘Data Sets Representative of the Structures and Experimental Properties of FDA-Approved Drugs’, *ACS Medicinal Chemistry Letters*, 9(3), pp. 204–209.

Duran-Frigola, M. *et al.* (2020) ‘Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker’, *Nature Biotechnology*, 38(9), pp. 1087–1096.

Gao, Y. *et al.* (2020) ‘Structure of the RNA-dependent RNA polymerase from COVID-19 virus’, *Science*, 368(6492), pp. 779–782.

Gaulton, A. *et al.* (2017) ‘The ChEMBL database in 2017’, *Nucleic Acids Research*, 45(D1), pp. D945–D954.

Ghosh, S. *et al.* (2020) ‘β-Coronaviruses Use Lysosomes for Egress Instead of the Biosynthetic Secretory Pathway’, *Cell*, 183(6), pp. 1520-1535.e14.

Gootenberg, J. S. *et al.* (2017) ‘Nucleic acid detection with CRISPR-Cas13a/C2c2’, *Science*, 356(6336), pp. 438–442.

Gorbalenya, A. E. *et al.* (2020) ‘The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2’, *Nature Microbiology*, 5(4), pp. 536–544.

Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) ‘Fpocket: An open source platform for ligand pocket detection’, *BMC Bioinformatics*, 10, pp. 1–11.

Han, N. *et al.* (2021) ‘Identification of SARS-CoV-2–induced pathways reveals drug repurposing strategies’, *Science Advances*, 7(27), pp. 1–14.

Harvey, W. T. *et al.* (2021) ‘SARS-CoV-2 variants, spike mutations and immune escape’, *Nature Reviews Microbiology*, 19(7), pp. 409–424.

Huang, C. *et al.* (2020) ‘Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China’, *The Lancet*, 395(10223), pp. 497–506.

Jin, Z. *et al.* (2020) ‘Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors’, *Nature*, 582(7811), pp. 289–293.

Laimer, J. *et al.* (2015) ‘MAESTRO - multi agent stability prediction upon point mutations’, *BMC Bioinformatics*, 16(1), pp. 1–13.

Lecker, S. H., Goldberg, A. L. and Mitch, W. E. (2006) ‘Protein degradation by the ubiquitin-proteasome pathway in normal and disease states’, *Journal of the American Society of Nephrology*, 17(7), pp. 1807–1819.

Ledford, H. (2021) ‘COVID antiviral pills: what scientists still want to know’, *Nature*, 599(7885), pp. 358–359.

Li, G. and De Clercq, E. (2020) ‘Therapeutic options for the 2019 novel coronavirus (2019-nCoV)’, *Nature reviews. Drug discovery*, 19(3), pp. 149–150.

Li, Shiqin *et al.* (2021) ‘SARS-CoV-2: Mechanism of infection and emerging technologies for future prospects’, *Reviews in Medical Virology*, 31(2), pp. 1–16.

Liu, Y. et al. (2021) ‘SARS-CoV-2 Nsp5 Demonstrates Two Distinct Mechanisms Targeting RIG-I and MAVS To Evade the Innate Immune Response’, mBio, 12(5), pp. 1–21.

Lopez-Leon, S. *et al.* (2021) ‘More than 50 long-term effects of COVID-19: a systematic review and meta-analysis’, *Scientific Reports*, 11(1), pp. 1–12.

Lu, R. *et al.* (2020) ‘Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding’, *The Lancet*, 395(10224), pp. 565–574.

Madhavi Sastry, G. *et al.* (2013) ‘Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments’, *Journal of Computer-Aided Molecular Design*, 27(3), pp. 221–234.

Michel, C. J. *et al.* (2020) ‘Characterization of accessory genes in coronavirus genomes’, *Virology Journal*, 17(131), pp. 1–13.

Ochoa-Montaño, B., Mohan, N. and Blundell, T. L. (2015) ‘Chopin: A web resource for the structural and functional proteome of Mycobacterium tuberculosis’, *Database*, 2015, pp. 1–10.

Peng, Q. *et al.* (2020) ‘Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2’, *Cell Reports*, 31(11), pp. 1-10.

Perlman, S. and Netland, J. (2009) ‘Coronaviruses post-SARS: Update on replication and pathogenesis’, *Nature Reviews Microbiology*, 7(6), pp. 439–450.

Portelli, S. *et al.* (2020) ‘Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource’, *Nature Genetics*, 52(10), pp. 999–1001.

Press, S. (2015) *LigPrep 3.6*. Avaliable at: http://gohom.win/ManualHom/Schrodinger/Schrodinger_2015-2_docs/ligprep/ligprep_user_manual.pdf.

Schubert, K. *et al.* (2020) ‘SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation’, *Nature Structural and Molecular Biology*, 27(10), pp. 959–966.

Sedova, M. *et al.* (2020) ‘Coronavirus3D: 3D structural visualization of COVID-19 genomic divergence’, *Bioinformatics*, 36(15), pp. 4360–4362.

Semper, C., Watanabe, N. and Savchenko, A. (2021) ‘Structural characterization of nonstructural protein 1 from SARS-CoV-2’, *iScience*, 24(1), p. 101903.

Shang, J. *et al.* (2020) ‘Structural basis of receptor recognition by SARS-CoV-2’, *Nature*, 581(7807), pp. 221–224.

Shi, Y. *et al.* (2020) ‘D3Targets-2019-nCoV: a webserver for predicting drug targets and for multi-target and multi-site based virtual screening against COVID-19’, *Acta Pharmaceutica Sinica B*, 10(7), pp. 1239–1248.

Skwark, M. J. *et al.* (2019) ‘Mabellini: A genome-wide database for understanding the structural proteome and evaluating prospective antimicrobial targets of the emerging pathogen Mycobacterium abscessus’, *Database*, 2019(1), pp. 1–16.

Stokes, A. C. *et al.* (2021) ‘COVID-19 and excess mortality in the United States: A county-level analysis’, *PLoS Medicine*, 18(5), pp. 1–18.

Su, Y. C. F. *et al.* (2020) ‘Discovery and genomic characterization of a 382-nucleotide deletion in ORF7B and orf8 during the early evolution of SARS-CoV-2’, *mBio*, 11(4), pp. 1–9.

Thoms, M. *et al.* (2020) ‘Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2’, *Science*, 369(6508), pp. 1249–1256.

Torres, P. H. M., Rossi, A. D. and Blundell, T. L. (2021) ‘ProtCHOIR: a tool for proteome-scale generation of homo-oligomers’, *Briefings in Bioinformatics*, 22(6), pp. 1–15.

Wacharapluesadee, S. *et al.* (2021) ‘Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia’, *Nature Communications*, 12(1), pp. 1–9.

Wang, X. *et al.* (2020) ‘Rapid and sensitive detection of COVID-19 using CRISPR/Cas12a-based detection with naked eye readout, CRISPR/Cas12a-NER’, *Science Bulletin*, 65(17), pp. 1436–1439.

Waterhouse, A. *et al.* (2018) ‘SWISS-MODEL: Homology modelling of protein structures and complexes’, *Nucleic Acids Research*, 46(W1), pp. W296–W303.

Weinstein, J. N. *et al.* (2013) ‘The cancer genome atlas pan-cancer analysis project’, *Nature Genetics*, 45(10), pp. 1113–1120.

WHO (2018) Infection prevention and control of epidemic- and pandemic-prone acute respiratory infections in health care, Infection prevention and control of epidemic- and pandemic-prone acute respiratory infections in health care. Avaliable at: https://www.who.int/publications/i/item/infection-prevention-and-control-of-epidemic-and-pandemic-prone-acute-respiratory-infections-in-health-care

WHO (2020) Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), The WHO-China Joint Mission on Coronavirus Disease 2019. Available at: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

WHO (2021) COVID-19 weekly epidemiological update, World Health Organization. Available at: https://www.who.int/publications/m/item/covid-19-weekly-epidemiological-update.

Wishart, D. S. *et al.* (2018) ‘DrugBank 5.0: A major update to the DrugBank database for 2018’, *Nucleic Acids Research*, 46(D1), pp. D1074–D1082.

World Health Organization (2020) ‘Weekly Operational Update on COVID-19 September 27, 2020. World Health Organization’, pp. 1–10.

Wrobel, A. G. *et al.* (2021) ‘Structure and binding properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2’, *Nature Communications*, 12(1), pp. 1–6.

Wu, F. *et al.* (2020) ‘A new coronavirus associated with human respiratory disease in

China', *Nature*, 579(7798), pp. 265–269.

Xue, Y. *et al.* (2021) 'Database resources of the national genomics data center, china national center for bioinformation in 2021', *Nucleic Acids Research*, 49(D1), pp. D18–D28.

Yan, L. *et al.* (2020) 'Architecture of a SARS-CoV-2 mini replication and transcription complex', *Nature Communications*, 11(1), pp. 3–8.

Yan, L. *et al.* (2021) 'Coupling of N7-methyltransferase and 3′-5′ exoribonuclease with SARS-CoV-2 polymerase reveals mechanisms for capping and proofreading', *Cell*, 184(13), pp. 3474-3485.

Yang, J. *et al.* (2020) 'Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor', *Nature Communications*, 11(1), pp. 1–10.

Yang, J. and Zhang, Y. (2015) 'I-TASSER server: New development for protein structure and function predictions', *Nucleic Acids Research*, 43(W1), pp. W174–W181.

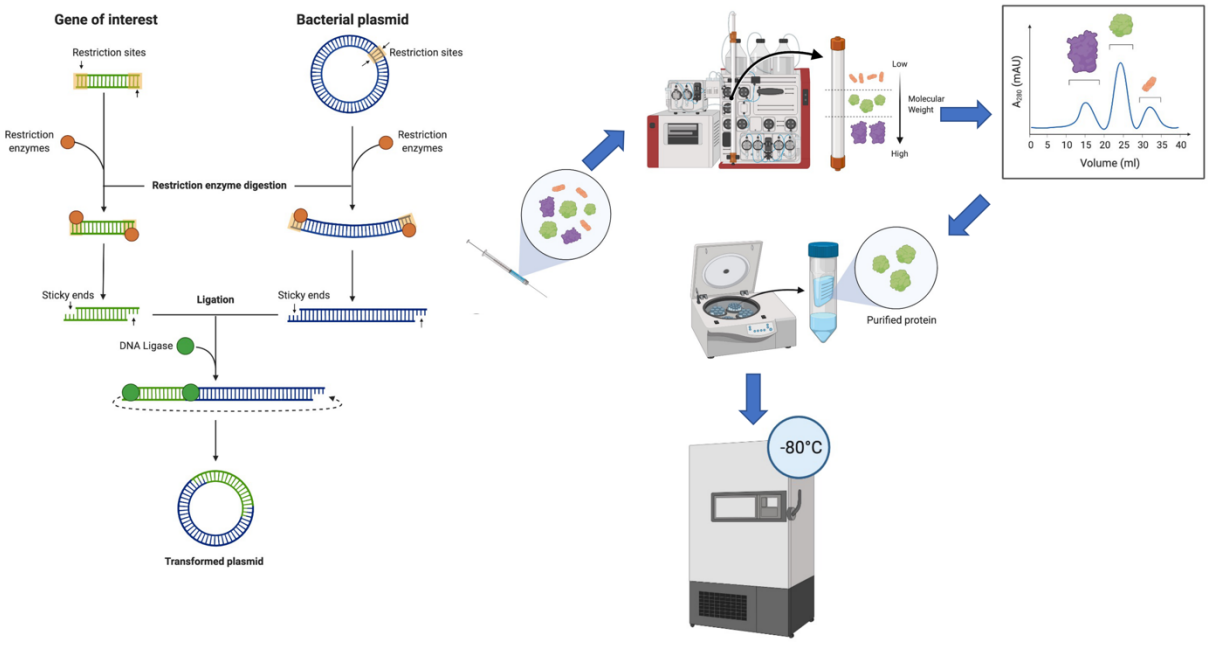
Zhang, Q. *et al.* (2021) 'Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy', *Signal Transduction and Targeted Therapy*, 6(1), pp. 1–19.

Zumla, A. *et al.* (2016) 'Coronaviruses-drug discovery and therapeutic options', *Nature Reviews Drug Discovery*, 15(5), pp. 327–347.

Chapter III: Studying the impact of mutations on NRAS.

Drug discovery target evaluation of SARS CoV-2 nsp13.

Graphical Abstract



GTPase NRas (NRAS)

RAS Introduction

The RAS subfamily comprises essential proteins that control cell proliferation. It is ubiquitously expressed in humans and consists of three small RAS GTPase proteins (KRAS, NRAS, and HRAS), which are 82-90 % identical (Downward, 2003). Kirsten rat sarcoma virus (KRAS) consists of two splice variants, KRAS4A and KRAS4B. The RAS proteins possess GTPase activity that influences a signalling pathway by switching between" on" and "off" states (Figure 1). This process happens through RAS conformational states, allowing interchange between GDP and GTP. Upon external stimuli to cell-surface proteins such as the EGFR receptor, the receptor dimerises and activates guanine nucleotide exchange factor (GEF) components such as SOS1, which lead to the exchange between GDP/GTP forming the active RAS protein. The reverse hydrolysis of GTP to GDP happens when the GTPase activating protein (Ras-GAP), such as neurofibromin, down-regulates RAS signalling pathways, enabling the GTP to GDP exchange.(Cullen and Lockyer, 2002)(Reuther and Der, 2000)

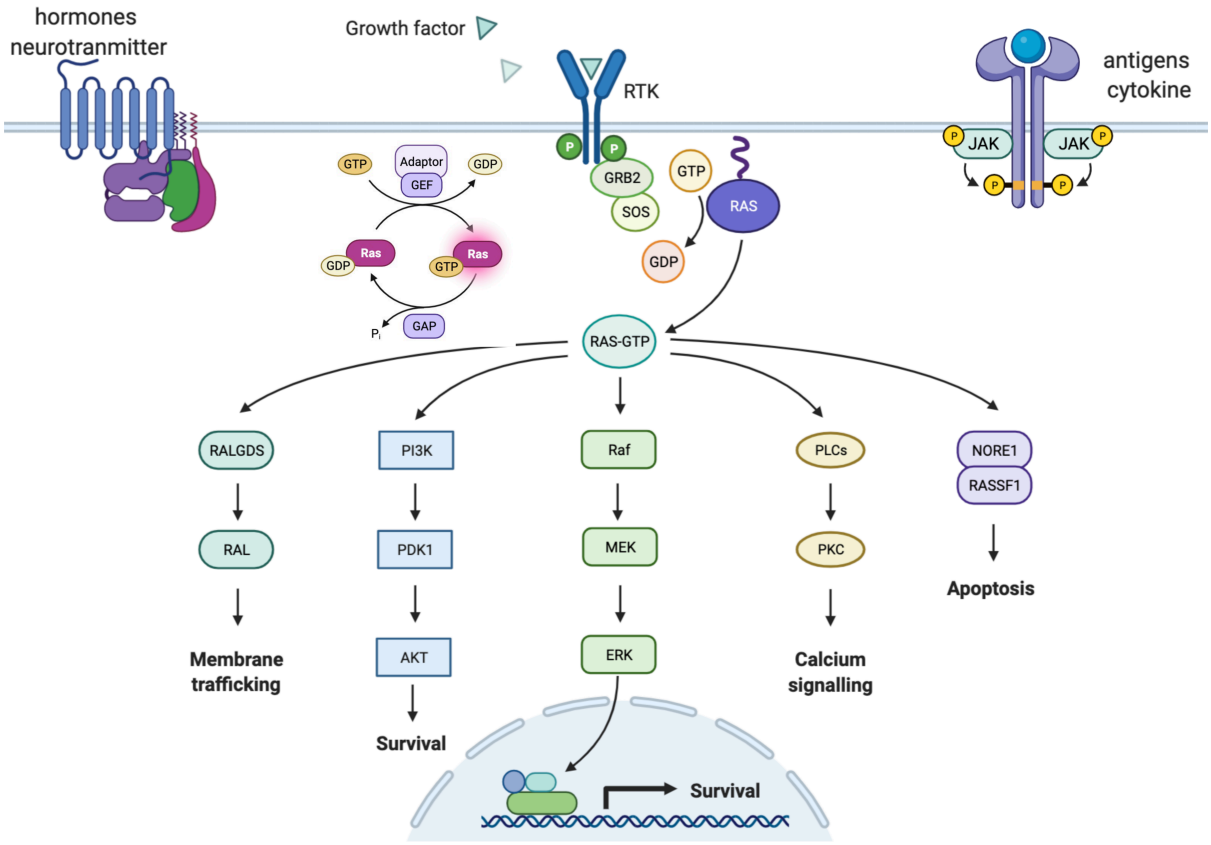


Figure 1 (made by biorender App). The two-state of the RAS cycle. Switching between active-GTP and inactive-GDP states is regulated by GEFs and GAP proteins. Activation of RAS leads to multiple cell functions such as apoptosis, gene regulation, and cell survival. There are five key RAS pathways shown in this diagram.

RAS proteins are close homologues, except in the C-terminal region known as the hypervariable region (HVR) (Figure 2D). This consists of 24 to 25 residues with a specific function for each RAS isoform that facilitates membrane interactions (Goswami *et al.*, 2020). RAS is a well-known oncogene and highly mutated gene, ranked among the top three causing cancer and death. Most RAS gain-of-function missense mutations occur around the GDP/GTP binding site, resulting in persistent RAS activation. Mutation data from (the COSMIC database) showed KRAS is highly mutated among RAS isoforms. It has 45,834 reported mutations, followed by NRAS with 7,622 mutations and HRAS with 2,186 mutations (Prior, Lewis and Mattos, 2012). Oncogenic RAS, which is not predominant in all cancers, is highly mutated in

pancreatic cancer and rarely mutated in breast cancer. RAS isoform mutation frequency varies in different cancer types. For example, KRAS is mutated in 85% of all cancers, NRAS 12%, and HRAS 3%. Oncogenic KRAS is found in patients with pancreatic carcinoma, lung malignancies, and colorectal tumours. Oncogenic NRAS is found in patients with melanomas, whereas oncogenic HRAS is found in head and neck cancers (Prior, Lewis and Mattos, 2012).

RAS localisation

Multiple mechanistic studies show that each RAS isoform displays specific downstream signalling through post-translational modification at the C-terminal hypervariable region. These post-translational modifications allow RAS isoforms to anchor to different subcellular membranes activating multiple signalling pathways (Ahearn *et al.*, 2012). RAS proteins mainly anchor to the plasma membrane but are also found in the endoplasmic reticulum and Golgi apparatus. RAS isoforms express differently in different cellular compartments. For example, KRAS is primarily found in the endoplasmic reticulum, whereas NRAS and HRAS are abundant in the Golgi pool. Thus, the RAS isoform exists in different cellular compartments activating different cellular pathways. HRAS is strongly associated with the RAF/MAPK signalling pathway, whereas KRAS is more closely related to the RAF1 activator than NRAS/HRAS. In addition, the existence of RAS isoforms in different cellular compartments can lead to other cellular functions. For example, anchoring of KRAS to the endoplasmic membrane leads to cellular transformation, while KRAS in mitochondria triggers apoptosis. In order for RAS to transmit signals and promote cell proliferation, it has to anchor to a membrane through HVR by covalently adding C15 farnesyl isoprenoid lipid. This is a very vulnerable aspect of the function of RAS,

i.e. in the absence of this process, RAS cannot be anchored to the plasma membrane and cannot function (Ahearn *et al.*, 2012).

Overall RAS mutations

Mutations in RAS have been known for a long time, during which many studies have been carried out to validate which RAS mutants are oncogenic. Mutations in RAS isoforms are found in approximately one-third of all human cancers. Each isoform has different predominant mutations. Codon 12, 13, and 61 mutations are the most frequent among isoforms, for example, G12 to (D, V, C, A, S, R), G13 to (D, C, S, R) and Q61 to (H, R, L, K).

In KRAS, G12(D, V, C, A) and G13D are the most frequent residues, whereas, in NRAS, Q61(R, K) is the most frequent residues. Similarly, HRAS Q61R and G13R are the most frequent residues. Since these mutations occur in a highly similar region among the RAS isoforms, it has been postulated that they function in the same way (Prior, Lewis and Mattos, 2012). However, multiple studies show that different amino acid substitutions in other RAS carried out distinct GTPase activities and manifest biologically as mutations with RAS preference for specific signalling pathways, cancer types, etc.

The RAS isoforms possess multiple mutations, raising questions about whether these mutations have the same effects in all isoforms and the outcome of these mutations on the signalling pathways? One of the challenges in predicting the impacts of mutations has been data variability when different cell lines were used. For example, (Voice *et al.*, 1999) compared the effect of G12V mutations among RAS isoforms using the NIH3T and Rat-1 cell lines to study the impact of the mutant to induce transformation and stimulate cell motility. Each cell line gave different

outcomes. This variation suggested that different cell types could have different outcomes. In addition, similar mutations in other RAS isoforms could lead to different effects. For example, (Haigis *et al.*, 2008). use genetically engineered mouse models to assess the impacts of mutations on RAS isoforms in colonic epithelium stimulated hyperproliferation. The G12V mutant in NRAS did not alter the growth properties of the epithelium. However, it confers resistance to apoptosis. In contrast, G12V in KRAS leads to an increase in the number of stem cells within the tumour epithelium. Thus, there is substantial evidence that the KRAS mutant alone in pancreatic ductal adenocarcinoma (PDAC) can maintain the malignant disease. However, in PDAC tumour suppressor genes such as Tp53 are commonly found to accelerate cancer metastasis. (Bournet *et al.*, 2016)

All frequent RAS mutations appear around the GDP/GTP binding site affecting the exchange rate between GDP and GTP. These mutations may alter natural ligand affinity by altering the receptor-ligand interactions. This has raised a similar question to that mentioned above: Will these mutations have the same effect on the natural ligands for all isoforms? (Der, Finkel and Cooper, 1986) analysed 17 different mutations around the HRAS GDP/GTP binding site and found that Q61L shows no impact on the ligand-binding site. The ligand binds to both wild-type and mutant-type with the same affinity. In addition, multiple mutations for KRAS have been studied, including Q61H, which showed a 6-fold difference between the wild-type and the mutant-type. G12D and G13D were shown to bind to GTP with the same affinity. In contrast, other G12C/V and G13C showed affinities to GTP of up to 2-fold higher (Der, Finkel and Cooper, 1986). The kinetic exchange between the GDP/GTP wild-type KRAS and mutant showed a similar GDP/GTP rate, except that G13D

suggested this mutant may contribute to more aggressive biology. The NRAS G12D and Q61R mutants showed similar GDP exchange rates. However, different GTP exchange rates, with Q61R being the slowest exchange rate (Hunter *et al.*, 2015). Homologue isoforms of RAS proteins make them functionally redundant. However, RAS isoforms functionally differ in multiple cell and tissue types. For example, knockout KRAS results in mice dying during embryogenesis (Koera *et al.*, 1997). Therefore, KRAS protein signalling is essential during embryogenesis and cannot replace another RAS isoform. Another factor that can keep the RAS protein continuously activated is the RAS-GAP regulator, such as SOS1 and NF1, which results in congenital disabilities like Noonan syndrome (Karnoub and Weinberg, 2008). Furthermore, mutations constantly activate cell membrane receptors, such as EGFR and tyrosine kinases. Finally, amplification of RAS wild type in cancer cells leads to more signalling activation (Stolze *et al.*, 2014).

Ras effector proteins

RAS proteins bind to multiple proteins, such as the RAF protein that when phosphorylated, initiates downstream phosphorylation to MEK that in turn phosphorylates ERK and MYC (Figure 1). The RAS protein also binds to class I phosphoinositide 3-kinases (PI3-Ks), which convert phosphatidylinositol (4,5)-bisphosphate (PIP2) into phosphatidylinositol (3,4,5)-trisphosphate (PIP3), promoting AKT phosphorylation which in turn phosphorylates mTORC1 (Shahbazian *et al.*, 2010). The third protein that binds to RAS is the RAL GEF, which phosphorylates the TBK1 associated with NF-*k*B. Therefore, controlling RAS signalling is essential since active RAS will lead to multiple downstream outputs (Doll *et al.*, 2017). Moreover, all

these proteins are highly mutated in cancer, making it even more complex, i.e. independent of RAS activation.

RAS as a drug target

RAS proteins have been known for a long as oncogenic and driving cancer progression. So why, when RAS has been known as an oncogene, is there no approved FDA drug until 2021? Is the recently approved drug effective against all RAS isoforms? Many factors need to be considered:

- Not all RAS proteins are the same, as explained in the above sections
- Many clinical trials went wrong because each isoform is dominant in a different type of cancer
- The signal transduction is not a linear pathway as in the drawing. Instead, it is a very complex network. Therefore, the RAS signalling pathways are not fully understood.

The literature reported four ways that RAS is targeted (L. Bryant, D. Cox and J. Der, 2018). The first one is preventing RAS membrane anchoring by targeting the farnesyl transferase (FTIs), leading to loss of function and cell death. *In vivo* experiments on mice treated with a 27-day FTIs inhibitor led to rapid cancer regression. Unfortunately, a clinical trial showed FTIs were ineffective against human pancreatic cancer for two reasons. First, FTIs inhibitors prevent only HRAS from anchoring to the plasma membrane (Berndt, Hamilton and Sebti, 2011). However, HRAS is the least prominent isoform, i.e. not a driver gene in cancer. Secondly, KRAS in the presence of FTIs can be covalently linked to geranyl isoprenoid C20 lipid that facilitates the attachment of KRAS to the plasma membrane cell, i.e. KRAS

starts to be active again and transmit signalling, and cancer grows. Other compounds that inhibit RAS membrane association are 2-bromopalmitate (2-BP) and Deltarasin. However, none of these drugs is selective for RAS (Zhong *et al.*, 2021).

The second way of inhibiting RAS is through RAS effector signalling. However, the RAS effector pathways are enormous (Figure 2). Which RAS pathway is essential to drive cancer forward and can be targeted? Multiple studies show that PI3-K and RAF pathways are crucial (Waldmann *et al.*, 2004). Furthermore, both pathways have kinase catalytic domains, which are very well understood and more tractable than other GTPase targets, which can be more challenging. A third way of inhibiting RAS is through synthetic lethal interactors, i.e. RAS interactome (Downward, 2015). Finally, inhibition of RAS driven metabolism represents a promising therapeutic vulnerability. Unlike normal cells, cancer cells get addicted to a particular metabolism or specific pathways (Hanahan and Weinberg, 2011). For example, cancer cells depend a lot on glycolysis. Chloroquine is a very well-known drug to treat malaria. It has multiple effects on metabolism. Therefore, it has been tested with another chemotherapy for cancer treatment (Wolpin *et al.*, 2014).
None of these approaches targets RAS directly since GTP binds tightly, not allowing pockets for a new molecule to bind. However, another allosteric pocket has been identified, and this will be explained more in the results section. Some of the compounds inhibit RAS function, such as DCAI, which interfere with blocking RAS activation SOS1. The Kobe 0065 compound inhibits RAS from transmitting downstream signalling by interfering with effector activation such as RAF (Welsch *et al.*, 2017). Compounds such as ARS-853 bind to RAS-G12C mutant form at allosteric pockets that do not present in the GDP wild-type with low micromolar

potency (Patricelli *et al.*, 2016). This compound was the first to change the idea that RAS is not a draggable target. Since then, a combination of a micromolar inhibitor of RAS with an upstream activation protein such as EGFR was hypothesised to be a more robust approach than finding a potent drug inhibiting RAS protein alone to avoid drug resistance. Recently on the 28th of May, the FDA has approved the first KRAS G12C mutant drug called sotorasib (Lumakras) for non-small cell lung cancer (NSCLC) which used to be treated with chemotherapy (Canon *et al.*, 2019). This astonishing development brings the question of the effect of this drug in combination with another drug? Moreover, is sotorasib only effective for this G12C mutant or other RAS isoform mutants such as NRAS Q61K or Q61L?

Materials and methods

Protein production

The experimental material described here focuses on molecular biology cloning, expression, and purification of NRAS wild-type and mutant forms. Amongst the most frequent NRAS mutants reported in the COSMIC database are Q61K/L, G12D, and G13D. The genes coding for residues 1-166 of NRAS wild type and four mutants were ordered from Thermo Fisher Scientific with code optimisation for expression in *E. coli*. The forward and reverse primers were designed for NRAS wild-type and mutant using multiple primer analyser tools from the Thermo Fisher Scientific website to check primer melting temperature, GC% content, etc. The final forward sequence was (ATATATGGATCCATGACCGAATATAAACTGGT), and the final reverse primer sequence was (ATACCCAAGCTTTTAATACTGGCGGATTT). In addition, random nucleotides were added before BamHI, and restriction digestion enzymes were highlighted in red to increase the efficiency of restriction enzymes binding at the restriction digestions stage. Next, the genes were amplified using a standard polymerase chain reaction (PCR). All the required reagents in Table 1 were added and mixed. Four PCR tubes were made for each gene and loaded into the PCR machine. The following thermocycling conditions were applied: Initial-1 for two minutes and initial-2 for one minute at 95 ^{0}C. Then 35 cycles for denaturation at 95 ^{0}C. Annealing-1 step for 20 seconds and annealing-2 for 40 seconds between 65-70 ^{0}C depending on the primer melting temperature. The final step is an extension for 20 min at 70 ^{0}C. The amplified DNA was evaluated by gel electrophoresis. The Gel Extraction Kit miniprep from Thermo Fisher Scientific was used to purify the DNA

fragment from the agarose gels. Genes concentration was measured using the Nanodrop, and all amplified gene yields were above 100 ng/ul.

Table 1. PCR reaction components for a single PCR tube.

Components	Stock Con	Finial Con	Single PCR (50uL)
dNTP Mix	10 mM	0.2mM	1 uL
PCR Buffer	10X	1X	5 uL
F primer	10 uM	0.3 uM	1.5 uL
R primer	10 uM	0.3 uM	1.5 uL
MgSO4	25 mM	1.5mM	3 uL
DMSO	100%	5%	2.5 uL
Polymerase	5 Units/uL	2.5 units	1 uL
Template DNA	10 ng/ul	50 ng (1ug/ul)	5 ul
Water			29.5

The plasmid is amplified by transformation, 1 ul of the purified pET28a-SUMO plasmid into a DH5-Alpha (Invitrogen) competent cell. First, heat shock at 42 ^{0}C for 40 seconds, then 250 ul of S.O.C media was added. Next, the cells were grown for 1 hour at 37 ^{0}C, and 50 ul was transferred from the grown cells into the premade agar plate. Finally, the plate was incubated at 37 ^{0}C overnight. After the colonies were grown on the agar plate, picked colonies were transferred into a 10 ml tube with 5 ml LB (Luria Bertani broth) media with 5 ul Kanamycin antibiotic (50mg/ml) and left shaking at 37 ^{0}C for 18 hours. Plasmid miniprep from Thermo Fisher Scientific was used to isolate grown plasmid for the next stage.

The amplified, purified genes and plasmid were processed to the restriction digestion stage. The enzyme sites (BamHI, Hind III) highlighted in red in the Materials and methods section above were cleaved, forming a sticky end insert. The reaction component is mixed in a 50 ul tube and incubated at 37 ^{0}C for 5 hours. On the other hand, purified concentrated plasmid restriction digestion was performed using a 100 ul tube (Table 2) for 5 hours. High fidelity enzymes were used to avoid star activity. The digested genes/plasmid gave a single intense band on the agarose gels. The standard protocol from Gel Extraction Kit miniprep was used to get pure concentrated cut-genes/plasmid.

Table 2. Restriction digestion reaction components

Component (gene insert)	Reactions
10 x CutSmart buffer	5 ul
BamH1 HF	1.5 ul
Hind III HF	1.5 ul
DNA (PCR product) 1000 ng/ul	10 ul
Water	To 100 ul
Component (pET28a-SUMO)	Reactions
10 x CutSmart buffer	10 ul
BamH1 HF	3 ul
Hind III HF	3 ul
DNA (vector) 2000 ng/ul	50 ul
Water	To 100 ul

Before the ligation stage, all the digested genes/plasmid concentrations were measured using the Nanodrop. The T4 DNA ligase was ordered from the BioLabs (https://international.neb.com/). The ligation molar ratio was selected as a 1:3 vector to insert. The mixture components were left for 3 hours at room temperature, then heat-inactivated at 70 ^{0}C for 10 minutes and placed on ice for 10 minutes. Finally, 5 ul of the reaction were transformed into 50 ul of DH5-Alpha competent cells. The transformation was carried out as described above. The ligated genes were checked by sequencing. All genes, wild type and mutant, were successfully cloned into pET28a-SUMO.

Table 3. Ligation reaction components

Component	Reactions
T4 DNA ligase buffer	2 ul
Digested pET28a-SUMO	4
Digested genes insert	1
Water	to 20 ul
T4 DNA Ligase	1 ul

Protein expression and purification

The NRAS wild-type and mutant residues 1-166 were cloned into the pET28a-SUMO vector with Ubl-specific protease-1. One μl of the plasmid was transformed into a Rosetta (DE3) cell. Multiple colonies were selected and transferred into 250 ml LB with 250 kanamycin (50mg/ml) for 16 hours at 37 ^{0}C. This is known as the primary culture. 120 ml of the primary culture was transferred into 6L LB with 1 ml kanamycin 50 mg/ml. The cultures were grown at 37 ^{0}C for approximately 3 hours until OD_{600nm}

0.5, followed by induction with 1mM of Isopropyl ß-D-1-thiogalactopyranoside (IPTG). The cultures were shaken at 200 rpm, and the temperature was reduced to 30 ^{0}C for 6 hours. The cells were harvested by centrifugation at 3000 rpm for 30 minutes. The cells were resuspended in 200 ml resuspended buffer (20 mM Tris, pH 8.0, 5 mM MgCl2, 50 mM NaCl, 5% glycerol, 20 mM GDP, tablet proteases cocktails inhibitors (Roche)) and frozen at -80 ^{0}C.

The following day, the cells were lysed by four rounds of sonication using EmulsiFlex-C5 homogeniser pressure between (1500-2000 psi). The sonicated lysate was clarified by centrifugation at 14,000 rpm for 20 min at 4 °C. Next, polyethyleneimine (PEI, 0.02% w/v) was added to precipitate contaminating nucleic acids and proteins in the lysate supernatant and stirred for 30 minutes at 4 °C. The precipitation solution was formed and cleared by centrifugation at 30000 rpm for 30 minutes. Finally, the protein supernatant was passed through a 0.45 mm pore membrane using syringe filtration to remove the remaining cell debris and large protein aggregates.

To increase the protein yield, the clear supernatant was incubated with Ni-NTA Agarose overnight at 4 °C to increase the protein yield. The supernatant with Ni-NTA Agarose resin passed through the manual purification column collecting all Ni-NTA Agarose beats. Next the resin was washed with 40 ml washing buffer (20 mM Tris, pH 8.0, 5 mM MgCl2, 50 mM NaCl, 5% glycerol, 20 mM imidazole) and the protein was eluted with elution buffer (20 mM Tris, pH 8.0, 5 mM MgCl2, 50 mM NaCl, 5% glycerol, 300 mM imidazole). The protein was loaded onto an anion exchange HiTrap Q HP 5 ml column again using an ÄKTA machine and eluted with ((20 mM Tris, pH 8.0, 5 mM MgCl2, 1 M NaCl, 5% glycerol, 20 mM GDP, 1 mM DTT). Ubl-specific protease 1, SUMO protease, was added to the purified protein and left

overnight dialysis with buffer (20 mM Tris, pH 8.0, 5 mM MgCl2, 50 mM NaCl, 5% glycerol, 1 mM DTT). The next day the protein was passed through the Hitrap HP 5ml column. The cleaved SUMO-His bound to the column and the proteins were collected through the flow-through. The collected protein was concentrated by centrifugal filtration (Sartorius Vivaspin), aliquoted, flash-frozen with liquid N2, and stored at 80 ^{0}C.

Results

NRAS structure annotation

The NRAS structure has been solved experimentally seven times: four times with GDP (PDB IDs: 3CON, 6WGH, 6ZIO, 6ZIR) and three times with GTP (PDB IDs: 5UHV, 6E6H, 6ZIZ). All the experimental structures are missing the hypervariable region. RAS proteins are well known as a monomer. However, multiple studies suggest RAS dimerisation. NRAS has been solved as a homodimer in two PDB IDs: 6WGH and 6ZIO.

Four experimental structures of mutants of NRAS have been solved: PDB IDs: 6E6H (G13D), 6ZIZ (Q61R), and 6ZIO, 6ZIR (C118S). Most KRAS proteins with frequently mutated residues are experimentally solved. However, there are multiple frequent mutations in NRAS, such as Q61K, Q61L, G12D and G12C, which have not been solved. For this reason, NRAS was selected as a target to study the impact of these mutations and compare it to another RAS isoform that has already been solved. In addition, the objective was to develop small molecules specifically bind to NRAS frequent mutants.

The three RAS isoforms have high sequence similarity, except in the HVR region, where there is around 8% sequence identity (Figure 2D). The G domain GTP bound-NRAS has critical residues that facilitate the active site's dynamic movement, i.e. the conformational changes of the P-loop (10-17), switch I (25-40) and switch II (60-74). The gain-of-function mutations disrupt the movement of the active site loop, preventing the hydrolysis of GTP and keeping the protein in the active state (Figure 2A/B). However, a protein structure is not static as we tend to describe it, but rather is continuously moving between multiple conformational states. Fpocket for NRAS

active states (PDB ID; 6E6H) shows a new potential pocket that could be targeted which could result in altering the protein conformation or enhancing the hydrolysis of GTP (Figure 2C). Since the natural ligand (GDP) binds to RAS with a picomolar affinity and cannot easily be replaced, other pockets were considered. One pocket shown in blue between switch I and switch II has been already characterised. However, four pockets still have not been targeted. (Figure 2C)

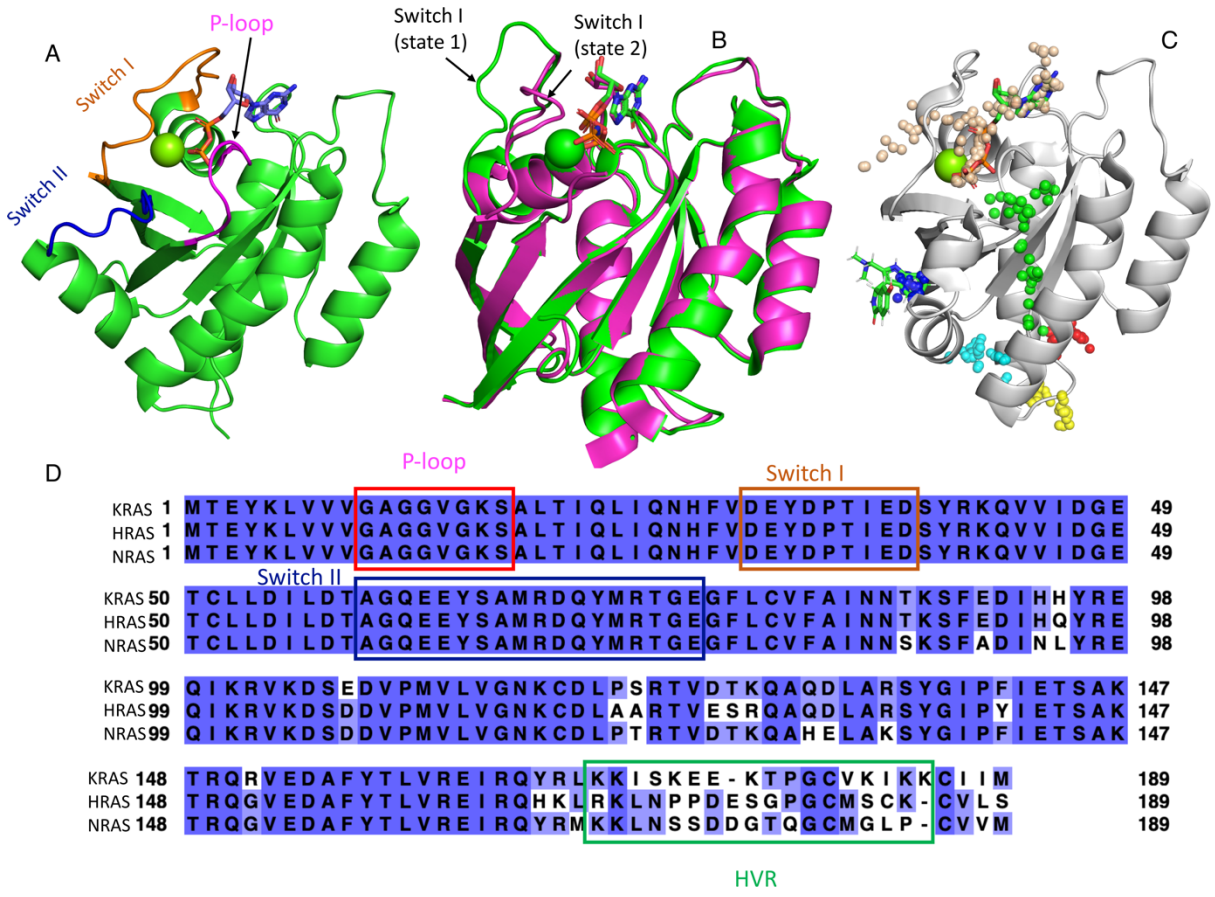


Figure 2. (A) Three-dimensional structure of the wild type residues (1-166) GDP ligand shown in stick, the NRAS essential loops coloured differently. (B) Overlay of wild-type 2 NRAS structures (active GTP) in green PDB ID (6E6H) and (inactive GDP) in magenta, ligands represented in stick blue and green (3CON). (C) The Fpocket prediction of NRAS PDB ID (6ZIZ) showed six pockets that can be targeted, each pocket coloured differently. Two pockets are occupied in wheat with GTP ligand and blue with compound code (EZZ). (D) multiple sequence alignment of RAS isoforms each loop represented in (A) highlighted with the same coloured. The low sequence identity hypervariable region is shown in square green.

There are 7622 NRAS mutations reported in the COSMIC database. The most frequent residues in NRAS are Q61R/K/L. It is essential to determine each structure since each mutant site might form a new pocket that could be targeted. For example (Figure 3) compares the NRAS-GTP wild-type to NRAS-GTP mutant Q61R. There is a structural difference between NRAS wild-type and mutant; the switch II helix in the mutant Q61R moves away, creating new pockets not present in the wild-type structure. In addition, the Q61 in the wild type has hydrogen bonding with water molecules and residue R68, allowing the switch I loop movement, hence allowing the NRAS to exchange between the GTP and GDP states. On the other hand, the Q61R mutant form shows the wild-type R68 hydrogen bonding broken, allowing the helix to move away and start to form a new hydrogen bonding with the switch I loop T35. As a result, the loop is fixed and prevents the exchange between the GTP and GDP.

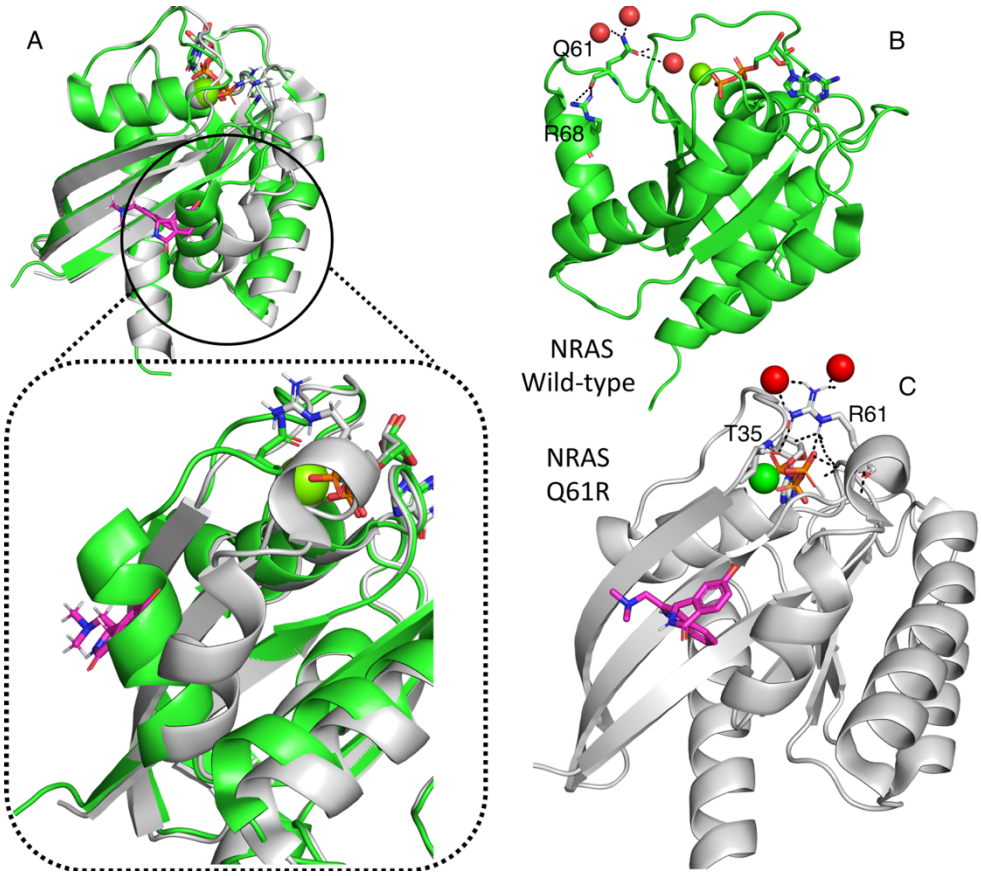


Figure 3. The difference between the NRAS wild-type and mutant interactions. (A) NRAS wild type coloured in green and Q61R coloured in white. Conformational changes in the helix connected to switch II are represented in a circle. (B) Q61 NRAS wild type interaction with R68 and surrounding water molecules is represented in sphere red. (C) R61 NRAS mutant interacts with the switch I loop residue T35 and surrounding water molecules.

Are other Q61K/L frequent mutants going to have the same effect? Are we going to see a new potential pocket? Unfortunately, the NRAS structures for these mutations have not been solved. The mutant modelled structures indicate Q61K mutant has a similar effect to Q61R due to the positive charge created by both amino acids. However, whereas Q61L has less impact on the structure, the pocket created by Q61R/K mutants is not observed for Q61L mutants, indicating this mutant has a less structural impact. (Figure 4).

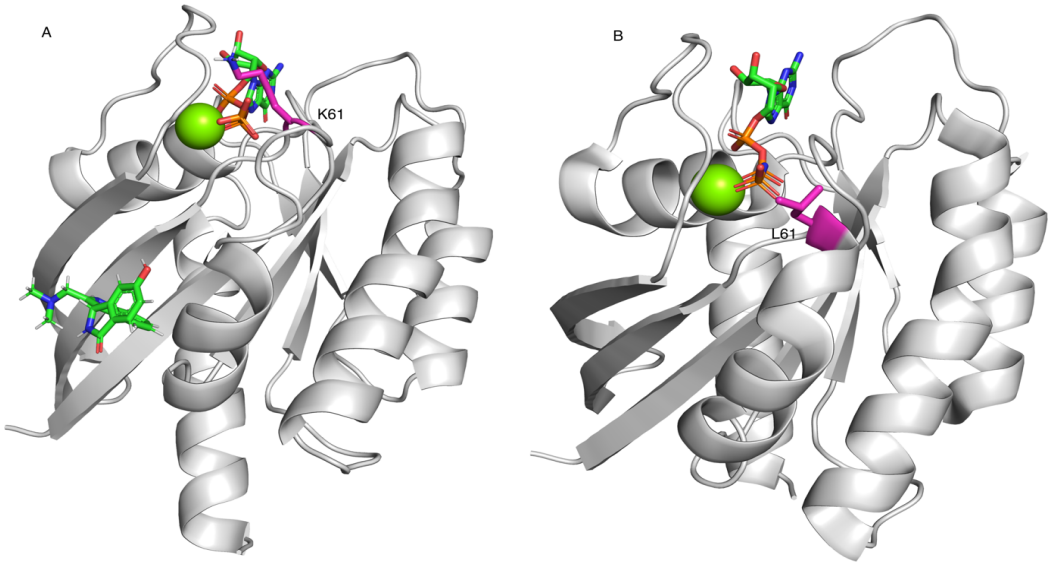


Figure 4. Mutant modelled structures. (A) The modelled K61 mutant structure, the K61 represented in stick magenta, GTP and other ligands shown in a green stick. (B) The modelled L61 mutant structure with the L61 represented in stick magenta, GTP represented in a green stick.

The experimentally solved structure KRAS G12C with approved FDA drug Sotorasib PDB ID: 6OIM is also of interest. After multiple structural analyses of RAS isoforms, the newly approved FDA drug Sotorasib appears very selective for KRAS G12C for the following reasons.

Sotorasib irreversible drug has a Michael acceptor (the substituent groups on an activated unsaturated compound, for example, ketone and nitro groups):

- which is susceptible to a nucleophilic attack, in this case, KRAS cysteine mutant (Figure 5B)
- Sotorasib fits into a pocket which in NRAS Q61R is moved towards it and narrowed. (Figure 5C)
- There is glycine residue at position 12 in NRAS, which cannot act as a nucleophile to attack the designed Michael acceptor in Sotorasib. Therefore, Sotorasib is considered very specific for KRAS G12C only.

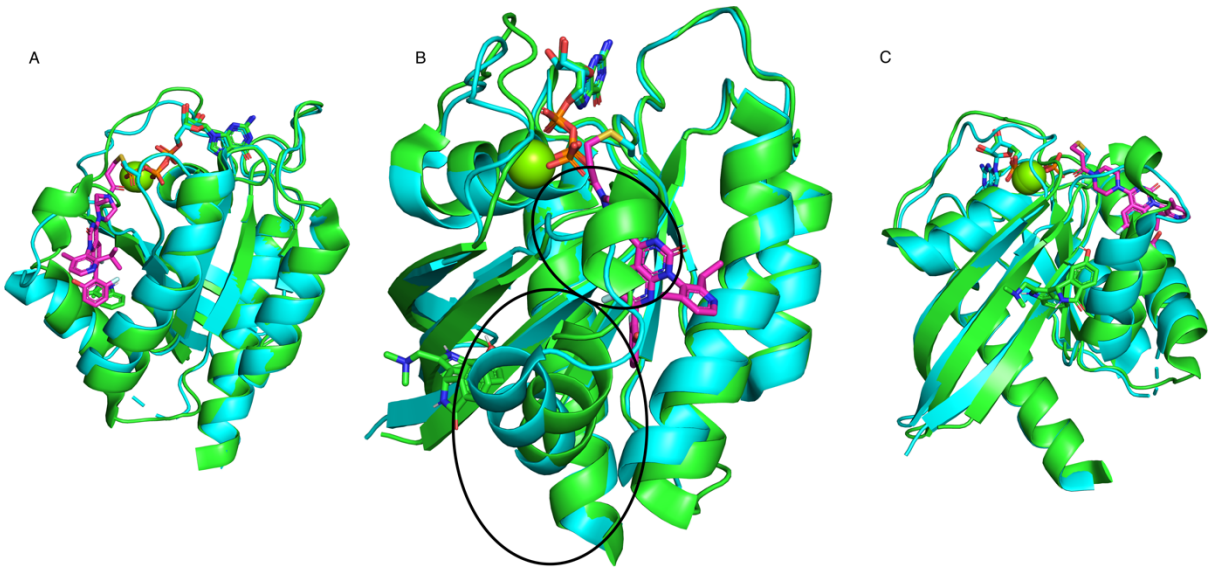


Figure 5. comparison of KRAS G12C mutant PDB ID: 6OIM and NRAS mutant Q61R PDB ID:6ZIZ (A) Side view of the overlay of the NRAS mutant coloured in green and KRAS coloured in cyan with Sotorasib represented in a magenta stick. (B) structural alignment indicated Sotorasib binding is not available in NRAS. (C) Side view of the overlay of the NRAS mutant coloured in green with ligand represented in green and KRAS coloured in cyan.

Synthesis and purification of NRAS Q61K/L mutant.

Solving lysine mutants in most proteins is challenging since lysine often affects protein stability. The NRAS Q61R mutant has been solved previously, whereas the lysine mutant is still unsolved, although both amino acids are basic and can form ionic and hydrogen bond interactions. In addition, arginine has a guanidine group that interacts in three possible directions providing protein stability, whereas lysine geometric structure has one direction to interact. Therefore, the NRAS Q61K was very challenging to obtain compared to NRAS Q61L. (Figure 6)

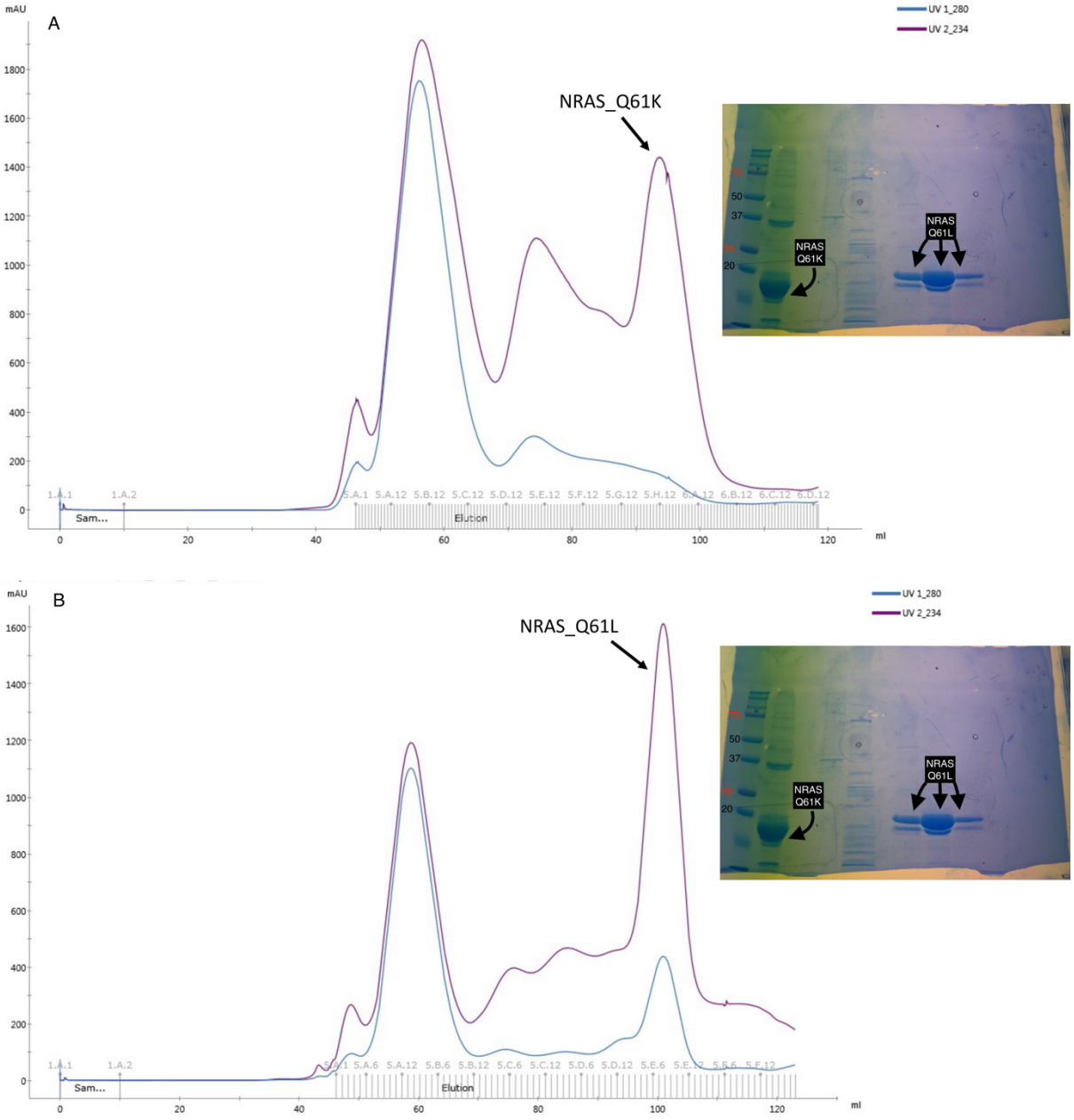


Figure 6. Elution profiles and SDS PAGEL gel for NRAS after size exclusion. (A) the NRAS Q61K mutant, the UV1_280 in purple, shows a shallow peak. However, looking at UV_234 allows detection of the amount of protein purified. (B) the NRAS Q61K mutant. More protein was purified for this mutant, the protein was detected by both UV, and more pure samples were obtained.

Crystallisation condition

Multiple commercial crystallisation plates have been tried, including BCS, JCSG+, and Morpheus. In addition, 40 μl of concentrated protein is loaded into mosquito robots that allow fast, accurate pipetting. Protein against reservoir ratios 2:1 and 2:2 was used. Unfortunately, no crystal has been grown in the BCS, JCSG+ from these commercial conditions. However, six crystals have been grown from the Morpheus (Figure 7), and these crystals will be fished and sent to the Diamond Light Source to check the diffraction patterns.

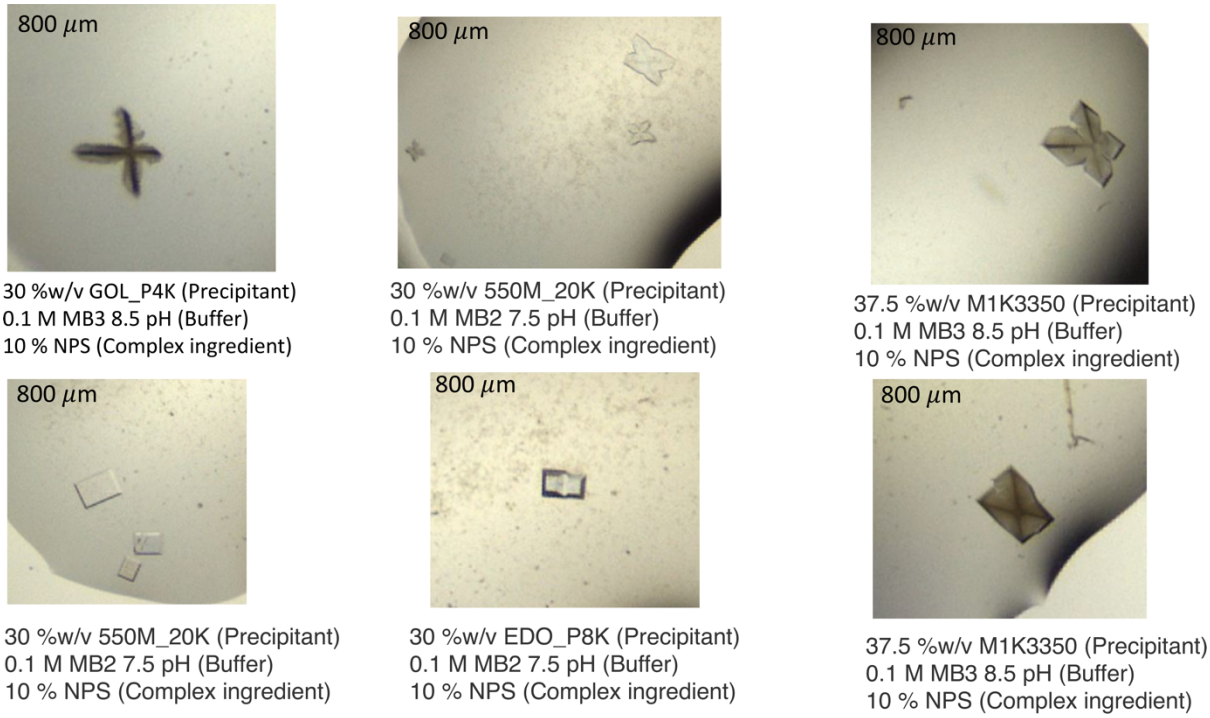


Figure 7. A different crystal form was grown for NRAS Q61L mutant using Morpheus crystallisation commercial plate.

SARS CoV-2 Non-structural protein 13 (NSP-13)

Nsp-13 introduction

The nsp-13 is known as helicase. One of the essential proteins in SARS CoV-2 plays a critical role in the viral life cycle (Vazquez *et al.*, 2021). It is highly conserved throughout the coronavirus family, making it an attractive target for developing a new inhibitor (Jia *et al.*, 2019). Only one amino acid differs (V570I) between SARS CoV-2 and SARS CoV-1. Developing a high potent compound could potentially prevent the emergence of coronavirus. Recent structural analysis of nsp13 showed two highly conserved druggable pockets in nsp13 (Newman *et al.*, 2021). SARS CoV-2 helicase belongs to the helicase superfamily IB. It is a motor protein that utilises triphosphate (ATP) hydrolysis to catalyse the unwinding of DNA and RNA double helix from a 5' to 3' direction into one RNA/DNA strand (Tanner *et al.*, 2003). The involvement of nsp-12 enhanced the catalytic unwinding process by The SARS CoV-2 helicase acting on RNA (Chen *et al.*, 2020). However, *in vivo* activity shows that helicase also acts on DNA and involves multiple biological processes such as chromatin remodelling (Mickolajczyk *et al.*, 2021). The specific function of nsp-13 in SARS CoV-2 is not very well defined. However, multiple experimental structures showed that SARS CoV-2 helicase interacts with other SARS CoV-2, such as RNA polymerase nsp-12 forming the mini RTC complex (nsp-7, nsp-8, nsp-12, nsp-13), which is essential for the virus replication life cycle.

Nsp-13 is 67 kDa, consisting structurally of five domains: zinc-binding domain at the C-terminal, stalk domain, beta-barrel 1B domain, and two 1A/2A domains coordinate with each facilitating the unwinding and the hydrolysis processes (Chen *et al.*, 2020). The nsp-13 five domains are also found in other viruses, such as nsp-10 from

Equine arteritis virus (EAV), interestingly in humans such as regulator of nonsense transcripts 1 (UPF1) (Tang *et al.*, 2020). The TM-align score shows that the structural similarity between nsp-13 and human protein is 0.7, which could be considered a disadvantage in targeting nsp-13. There were several efforts to find compounds that inhibit nsp-13 and display cellular activity. However, structural information of the ligand binding modes was missing, preventing further ligand development. The Gileadi lab recently solved 152 nsp-13 experimental structures with different fragment molecules in collaboration with the Diamond Light Source and the XChem team (Newman *et al.*, 2021). In addition, multiple Cryo-EM structures show nsp-13 in high order assembly with other SARS CoV-2 proteins (Chen *et al.*, 2020)(Chen *et al.*, 2020).

After multiple fragments have been solved with nsp-13 structure, the goal is to identify new lead compounds that bind to nsp-13 with higher μM and elicit inhibition of anti-viral activity by taking advantage of solved fragments.

Materials and methods

Protein expression and purification

The full length of the nsp-13 gene was amplified and cloned with the same procedures in section (Pag-161). The nsp-13 was cloned into two plasmid vectors pET28a-SUMO, and pET28a, to see which plasmid gave better expression. The nsp-13 gene was expressed in the Rosetta (DE3) cell. Multiple colonies were transferred into 250 ml 2XYT media with 250 μL Kanamycin (50 mg/ml) for 16 hours at 37 ^{0}C. Next, 20 ml was transferred to each 6L 2XYT media with 1 ml Kanamycin (50 mg/ml) until OD_{600nm} reached 0.8. The culture was cold to 16 ^{0}C before the induction of 2 mM IPTG. The cultures were shaking at 200 rpm for 18 hours at 16 ^{0}C. The cells were harvested by centrifugation at 3000 rpm for 30 minutes. The cells were resuspended in 250 ml resuspended buffer (50 mM HEPES-NaOH, pH 7.5, 500 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol, 20 μM ATP, 2 tablets proteases cocktails inhibitors (Roche). and frozen at -80 0C. The cells were lysed by four rounds of sonication using EmulsiFlex-C5 homogeniser pressure between (1500-2000 psi). After centrifugation at 30,000 rpm, the clear supernatant was incubated with Ni-NTA Agarose overnight at 4 °C.

The resin collected into manual column washed with 200 ml of (50 mM HEPES-NaOH, pH 7.5, 1 M NaCl, 4 mM MgCl2, 5% (v/v) glycerol) then eluted with elution buffer (50 mM HEPES-NaOH, pH 7.5, 500 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol, 300 mM Imidazole). The nsp13 protein was eluted at five mM of imidazole in the first batch. Therefore, no imidazole was included in the lysis and washing buffers.

After eluting the protein from the Ni-NTA Agarose resins, the protein was dialysed overnight in 2L buffer (50 mM HEPES-NaOH, pH 7.5, 500 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol) to remove imidazole then loaded into HisTrap HP 5 ml column using an ÄKTA machine. It was then eluted with slow gradient buffer A (50 mM HEPES-NaOH, pH 7.5, 200 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol) and buffer B elution buffer.

After elution of the protein from HisTrap HP 5 ml column, the protein was dialysed overnight in 2L buffer (50 mM HEPES-NaOH, pH 7.5, 50 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol) to remove high salt, then loaded into Q column using ÄKTA machine. It was then eluted with slow gradient buffer A (50 mM HEPES-NaOH, pH 7.5, 0 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol) and buffer B elution buffer with 1 M NaCl

The purified fractions were collected and loaded onto a HiLoad 16/600 Superdex column which was equilibrated with (50 mM HEPES-NaOH, pH 7.5, 500 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol), and eluted with (50 mM HEPES-NaOH, pH 7.5, 500 mM NaCl, 4 mM MgCl2, 5% (v/v) glycerol, 0.5 TCEP, 20 μM ATP).

Results

Preliminary expression and purification tests.

The pET28a-nsp13 initially transformed into *E. coli* Bl21 (DE3). However, colonies did not grow. Furthermore, different ranges of IPTG concentration between 1 and 5 mM were used, and various temperatures 15-30 ^{0}C and times 3-16 hours were used. The media were cooled to 15 ^{0}C before adding 200 mM of IPTG, then left shaking at 200 rpm for 16 h. Including ATP in the lysis buffer is essential to obtain stable protein.

Synthesis and purification of nsp13.

The expression and purification of nsp-13 are explained in detail in the materials and methods section. (Figure 8) shows elution profiles of the QHP column, HiLoad 16/600 Superdex column, and SDS PAGE gels of nsp13 (67 kDa). The Q column shows more contaminated samples. After size exclusion chromatography, the protein becomes a cleaner homogenous sample.

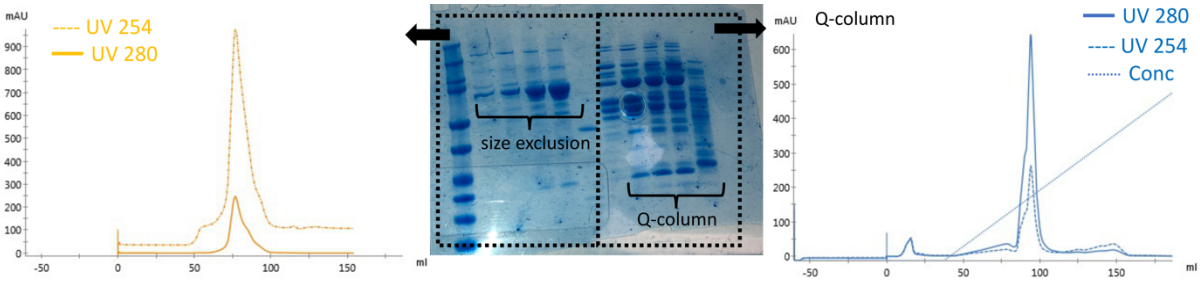


Figure 8. Elution profiles and SDS PAGEL gel for nsp13 protein after Q column and size exultation. The nsp13 size-exclusion shown in yellow shows a shallow peak. Whereas the gradient elution from the Q column is shown in blue.

Structure analysis of nsp13

The nsp-13 overall shape is a triangular pyramid consisting of 5 domains (Figure 9). The nsp-13 domains are linked with flexible linkers, for example, 30 amino acid linkers between 1B and 1A domains. This indicates that the nsp-13 structure is dynamic, and domains can move slightly. The nucleotide-binding site is between the 1A and 2A domains. The Fpocket detections for nsp-13 showed multiple pockets. Most of these pockets can be accommodated with fragments. The question is, which pockets should be targeted? When there are multiple pockets in the structure, the protein function can be decided which pockets should be targeted to disrupt the target function. In this instance, the preferred target sites are the DNA/RNA binding site or the allosteric site.

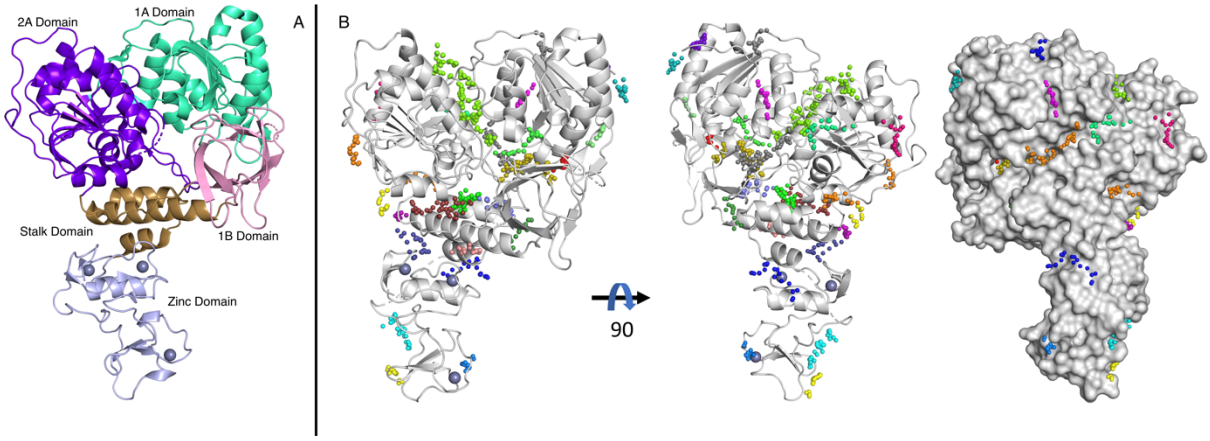


Figure 9. Experimentally solved nsp13 structure. (A) five domains coloured differently. (B) nsp13 Fpocket prediction of ligand and fragments binding site. Each pocket is coloured differently and represented in a small sphere.

Fragment growing and linking

Fragment-based drug design (FBDD) has become a leading method in drug discovery. The nsp13 protein has been solved with 52 fragments by the *Opher Gileadi lab* at the University of Oxford. All these fragments are very low molecular weight allowing growing, linking and merging fragments to obtain a lead-like compound. The fragment library usually has multiple chemical features such as chiral centres, hydrogen bond acceptors and donors, which are very important for fragment growing and fragment linking. Therefore, experimentally solved fragments could be enhanced computationally by screening a large fragment library. Working with solved fragments will be highly beneficial for moving the project forward quickly. Two areas will be illustrated here.

- Finding fragments with similar chemical properties could replace experimentally solved fragments with better binding. This process is known as scaffold hopping. (Figure 11)
- Fragment linking, joining two fragments together, forming a new lead compound that could be a potential inhibitor. (Figure 10A)

Fragments screening and linking are typically done by screening a database of small fragments library. Two fragments that occupied different regions were selected from an experimentally solved structure for fragments linking (Figure 10C) PDB ID (5RME, 5RLI). Most algorithms have proven to work for large ligands but not with low molecular weight fragments. Therefore, it is essential to consider the type of software to analyse fragments. The Cresset tools (Spark and Forge) were used to perform the fragment linking and growing (Stroganov *et al.*, 2008). Fragments were joined by the closest points highlighted in (Figure 10C). It is crucial to maintain the electrical fields

like the original fragments. Therefore, all linked fragments' electrostatic fields have a similar shape to the original fragments (Figure 10B).

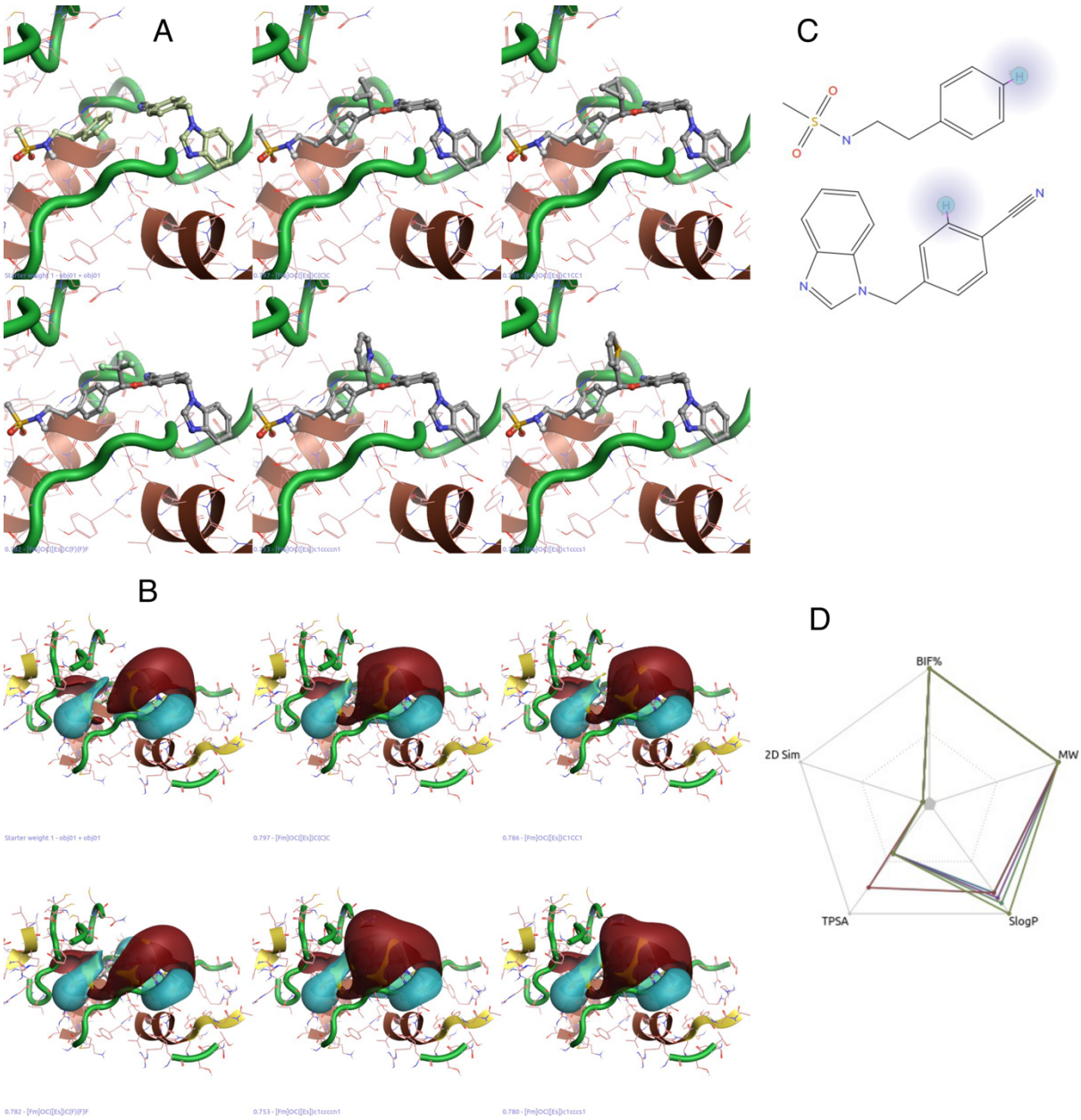


Figure 10. Fragments growing using Cresset tool. (A) two fragments joined with cyclic and non-cyclic linkers using an extensive fragments library. (B) the electron density map of joined fragments negative charges is represented in cyan, whereas the positive charge is represented in purple. (C) the original fragments, the highlighted region with sphere hydrogen is where the two fragments joined. (D) radial plot graph shows new compounds' physio-chemical properties, such as total polar surface area (TPSA) and molecular weight.

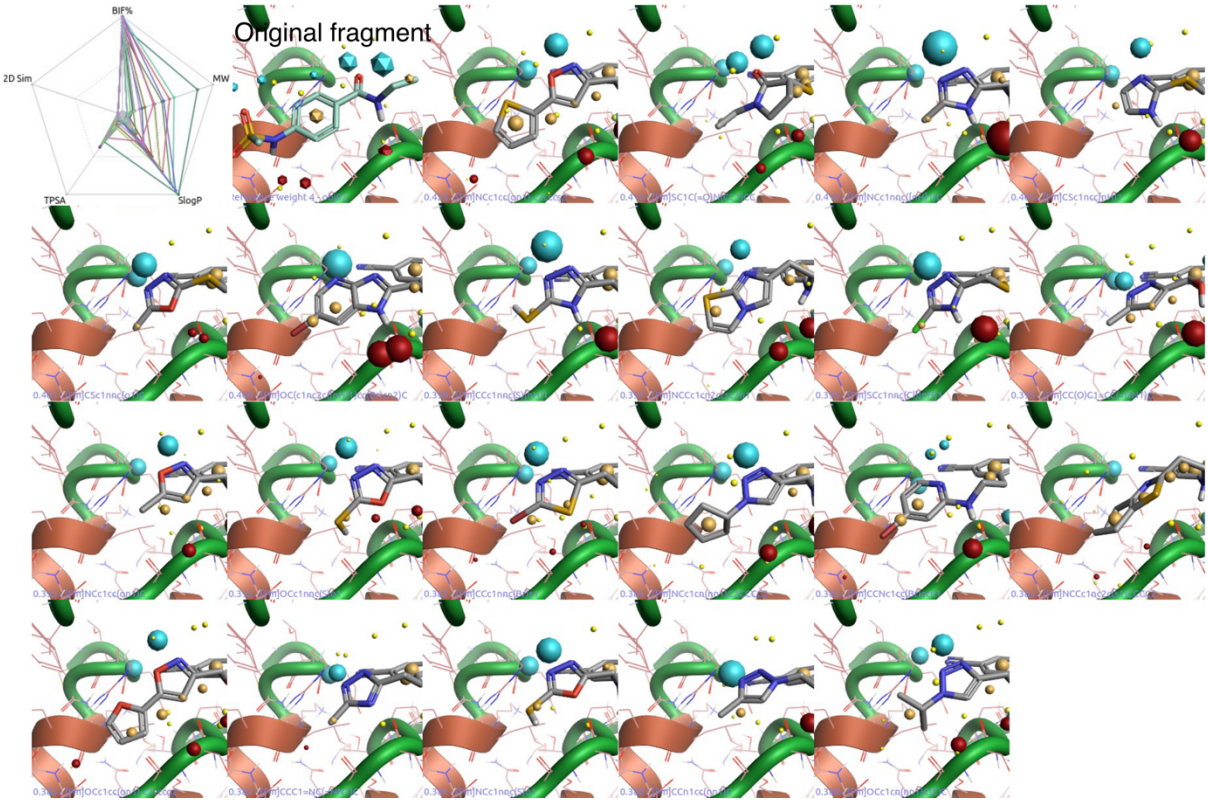


Figure 11. Isostere fragments are fragments that have similar chemical and physical propriety to the original fragment.

Specific fragments have preferred interaction with a particular pocket. Unfortunately, most of the available fragments follow the chemical combinatorial rule, i.e. not synthetically challenged. However, these fragments usually miss 3D complexity, which is needed for shallow pockets. Therefore, fragment growing is the best strategy to fulfil a small pocket's shape, size, and phytochemical properties (Figure 12). The newly grown compound should enhance receptor-ligand interaction and increase the drug likeness. Furthermore, two allosteric pockets can be targeted via fragments growth. This could potentially inhibit the nsp13 function by restricting the nsp13 dynamic movement in the stalk domain.

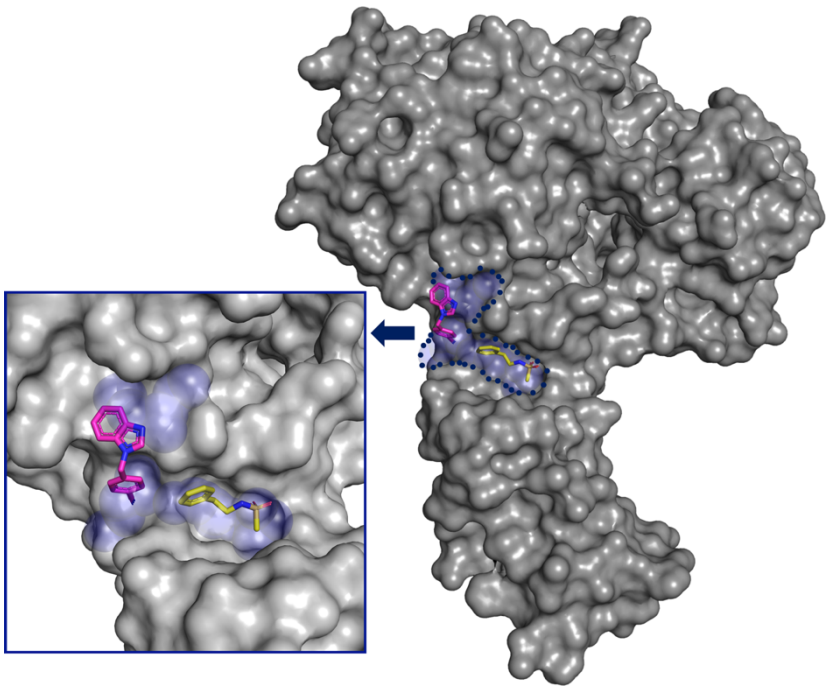


Figure 12. Two fragments from experimentally solved structures PDB ID (5RLI, 5RME) are yellow and magenta. The small pockets around the fragments shown in transparent blue allow both fragments linking/growing.

Discussion

RAS small GTPase is an essential protein controlling cell proliferation and survival. The RAS proteins have high sequence identity in the G-domain and low sequence identity in the hypervariable region. RAS proteins are dynamic switches of conformational states to exchange between GDP and GTP. Oncogene mutations varied between RAS isoform and stabilised and destabilised different conformational states, specifically switch II and ATP hydrolysis. The most frequent mutations on NRAS isoform found in patients are Q61R, Q61K, and Q61L. Only the Q61R mutant has been solved experimentally, whereas the other two mutants have not been solved previously for any RAS isoforms. Early work on RAS focused on solving mutant/wild types of structures and the hydrolysis of the ATP to GDP. However, the focus in solving the NRAS mutant structures is to identify new allosteric pockets for a potential selective NRAS inhibitor.

NRAS protein is very small in size 19 KDa, and cannot be studied experimentally using a technique such as Cryo-EM, which can capture different dynamic states of the protein. The NRAS previous mutant/wild structures have been solved without the hypervariable region. In addition, the mutant expression level of NRAS is not sufficient to accomplish drug discovery projects. These limitations directed most NRAS studies to solve 3D structures without further elaborating. Most proteins in the living system are assembled in a high order state. Therefore, we tend to think of the problem individually, not collectively. Solving the entire length of NRAS mutants in complex with other proteins such as mitogen-activated protein kinase 3 (MAPK3) increases the overall size, which allows it to be studied by Cryo-EM.

Furthermore, the binding of mutant-NRAS to other proteins could enhance the detection of allosteric pockets. Typically for drug discovery projects, a large amount

of protein is needed. Therefore, the NRAS mutant expression systems could be enhanced by optimising different systems such as insect cell lines or other plasmids with/without soluble tags such as GFP.

The purified NRAS wild/mutant types are enough to set up a few crystallisation plates to identify the 3D structures that could be the starting point to look for allosteric pockets before doing other biophysical techniques. Optimising the expression system to obtain sufficient NRAS protein is essential for further experiments.

Human drug targets are very challenging, and learning from them will be essential. The SARS CoV-2 nsp13 was selected as a covid-19 drug target. It is conserved throughout the coronavirus family, making it a promising drug target candidate. The nsp13 primary function is to unwind double-strand DNA or RNA. Furthermore, it is necessary for mini RTC complex formation. There are multiple flexible linkers between nsp13 domains allowing the protein to be dynamic. Therefore, finding fragments pockets around the stalk domain could disturb the nsp13 dynamics and function. The nsp13 protein can be studied using cryo-EM. However, small-molecule fragments can be significantly challenging to be detected using this technique. Unfortunately, most drug discovery targets tend to be investigated individually even when other interacting proteins are known, which sometimes limits our understanding of the target behaviour. The nsp13 solved experimentally with nsp12-nsp8-nsp7 making a drug discovery on the entire complex instead of nsp13 itself will give more insight. Since binding partners sometimes could change protein conformation or restrict protein dynamic.

Conclusion

Two proteins were selected to validate modelled structures, and to understand the impacts of mutations on protein structures. The NRAS was chosen from the COSMIC cancer gene census 3D, whereas the nsp13 was selected from SARS CoV-2 3D databases. The NRAS wild type Q61K/L mutants and nsp13 were successfully cloned, expressed, and purified. The NRAS frequent mutant Q61K is expected to form an allosteric pocket similar to experimentally solved structure Q61R. In contrast, Q61L is expected to show no effect on the ligand-binding site as predicted by SDM and mCSM tools. Fragments merging and hopping show promising lead candidates to be tested experimentally. Future work will include multiple crystallisation conditions and optimisation of the expression system to obtain a large amount of protein will be carried out. In addition, multiple biophysical techniques will be used, such as ITC and SPR, to test new fragment binding affinity and merged fragments. Cryo-EM technique will be used for nsp13 target with merged compounds after promising binding affinity result from biophysical technique ITC, and SPR

References

Ahearn, I. M. *et al.* (2012) 'Regulating the regulator: Post-translational modification of RAS', *Nature Reviews Molecular Cell Biology*, 13(1), pp. 39–51.

Berndt, N., Hamilton, A. D. and Sebti, S. M. (2011) 'Targeting protein prenylation for cancer therapy', *Nature Reviews Cancer*, 11(11) pp. 775–791.

Bournet, B. *et al.* (2016) 'KRAS G12D Mutation Subtype Is A Prognostic Factor for Advanced Pancreatic Adenocarcinoma', *Clinical and Translational Gastroenterology*, 7(3), pp. 1-8.

Canon, J. *et al.* (2019) 'The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity', *Nature*, 575(7781), pp. 217–223.

Chen, J. *et al.* (2020) 'Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex', *Cell*, 182(6), pp. 1560-1573.

Cullen, P. J. and Lockyer, P. J. (2002) 'Integration of calcium and Ras signalling', *Nature Reviews Molecular Cell Biology*, 3(5), pp. 339–348.

Der, C. J., Finkel, T. and Cooper, G. M. (1986) 'Biological and biochemical properties of human rasH genes mutated at codon 61', *Cell*, 44(1), pp. 167–176.

Doll, S. *et al.* (2017) 'Quantitative proteomics reveals fundamental regulatory differences in oncogenic hras and isocitrate dehydrogenase (IDH1) driven astrocytoma', *Molecular and Cellular Proteomics*, 16(1), pp. 39–56.

Downward, J. (2003) 'Targeting RAS signalling pathways in cancer therapy', *Nature Reviews Cancer*, 3(1), pp. 11–22.

Downward, J. (2015) 'RAS synthetic lethal screens revisited: Still seeking the elusive prize?', *Clinical Cancer Research*, 21(8), pp. 1802–1809.

Goswami, D. *et al.* (2020) 'Membrane interactions of the globular domain and the hypervariable region of KRAS4b define its unique diffusion behavior', *eLife*, 9(e47654), pp. 1–24.

Haigis, K. M. *et al.* (2008) ‘Differential effects of oncogenic K-Ras and N-Ras on proliferation, differentiation and tumor progression in the colon’, *Nature Genetics*, 40(5), pp. 600–608.

Hanahan, D. and Weinberg, R. A. (2011) ‘Hallmarks of cancer: The next generation’, *Cell*, 144(5), pp. 646–674.

Hunter, J. C. *et al.* (2015) ‘Biochemical and structural analysis of common cancer-associated KRAS mutations’, *Molecular Cancer Research*, 13(9), pp. 1325–1335.

Jia, Z. *et al.* (2019) ‘Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis’, *Nucleic acids research*, 47(12), pp. 6538–6550.

Karnoub, A. E. and Weinberg, R. A. (2008) ‘Ras oncogenes: Split personalities’, *Nature Reviews Molecular Cell Biology*, 9(7), pp. 517–531.

Koera, K. *et al.* (1997) ‘K-Ras is essential for the development of the mouse embryo’, *Oncogene*, 15(10), pp. 1151–1159.

L. Bryant, K., D. Cox, A. and J. Der, C. (2018) RAS Oncoproteins: Therapeutic Vulnerabilities, Tocris Scientific Review Series. Available at: https://www.tocris.com/literature/scientific-reviews/ras-oncoproteins.

Mellema, W. W. *et al.* (2015) ‘Comparison of clinical outcome after first-line platinum-based chemotherapy in different types of KRAS mutated advanced non-small-cell lung cancer’, *Lung Cancer*, 90(2), pp. 249–254.

Mickolajczyk, K. J. *et al.* (2021) ‘Force-dependent stimulation of RNA unwinding by SARS-CoV-2 nsp13 helicase’, *Biophysical Journal*, 120(6), pp. 1020–1030.

Newman, J. A. *et al.* (2021) ‘Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase’, *Nature Communications*, 12(1), pp. 1–11.

Patricelli, M. P. *et al.* (2016) ‘Selective inhibition of oncogenic KRAS output with small molecules targeting the inactive state’, *Cancer Discovery*, 6(3), pp. 316–329.

Prior, I. A., Lewis, P. D. and Mattos, C. (2012) ‘A comprehensive survey of ras mutations in cancer’, *Cancer Research*, 72(10), pp. 2457–2467.

Reuther, G. W. and Der, C. J. (2000) ‘The Ras branch of small GTPases: Ras family members don’t fall far from the tree’, *Current Opinion in Cell Biology*, 12(2), pp. 157–165.

Shahbazian, D. *et al.* (2010) ‘Control of Cell Survival and Proliferation by Mammalian Eukaryotic Initiation Factor 4B’, *Molecular and Cellular Biology*, 30(6), pp. 1478–1485.

Stolze, B. *et al.* (2014) ‘Comparative analysis of KRAS codon 12, 13, 18, 61, and 117 mutations using human MCF10A isogenic cell lines’, *Scientific Reports*, 5(8535), pp. 1–9.

Stroganov, O. V. *et al.* (2008) ‘Lead finder: An approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening’, *Journal of Chemical Information and Modeling*, 48(12), pp. 2371–2385.

Tang, C. *et al.* (2020) ‘Helicase of type 2 porcine reproductive and respiratory syndrome virus strain HV reveals a unique structure’, *Viruses*, 12(2), pp. 1–18.

Tanner, J. A. *et al.* (2003) ‘The severe acute respiratory syndrome (SARS) coronavirus NTPasefhelicase belongs to a distinct class of 5′ to 3′ viral helicases’, *Journal of Biological Chemistry*, 278(41), pp. 39578–39582.

Vazquez, C. *et al.* (2021) ‘SARS-CoV-2 viral proteins NSP1 and NSP13 inhibit interferon activation through distinct mechanisms’, *PLoS ONE*, 16(6), pp. 1–15.

Voice, J. K. *et al.* (1999) ‘Four human Ras homologs differ in their abilities to activate Raf-1, induce transformation, and stimulate cell motility’, *Journal of Biological Chemistry*, 274(24), pp. 17164–17170.

Waldmann, H. *et al.* (2004) ‘Sulindac-Derived Ras Pathway Inhibitors Target the Ras-Raf Interaction and Downstream Effectors in the Ras Pathway’, *Drug Discovery II*, 43(4), pp. 454–458.

Welsch, M. E. *et al.* (2017) ‘Multivalent Small-Molecule Pan-RAS Inhibitors’, *Cell*, 168(5), pp. 878-889.

Wolpin, B. M. *et al.* (2014) ‘Phase II and Pharmacodynamic Study of Autophagy Inhibition Using Hydroxychloroquine in Patients With Metastatic Pancreatic Adenocarcinoma’, *The Oncologist*, 19(6), pp. 637–638.

Zhong, L. *et al.* (2021) ‘Small molecules in targeted cancer therapy: advances, challenges, and future perspectives’, *Signal Transduction and Targeted Therapy*, 6(1), pp. 1–48.

Supplementary section

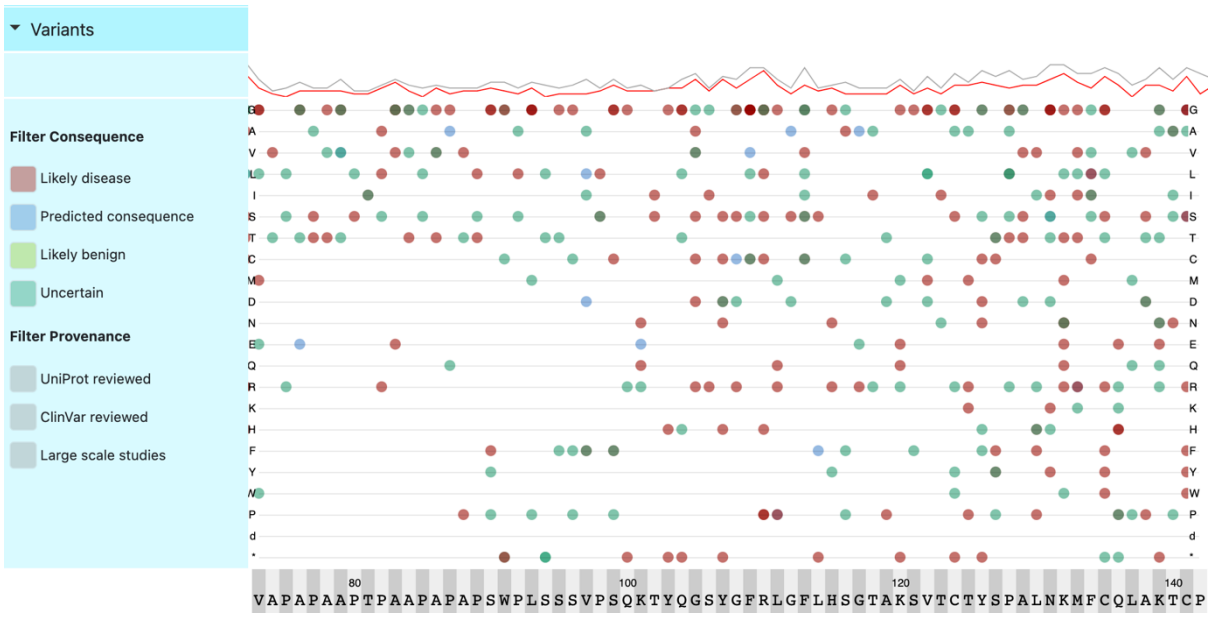


Figure 1. Curated mutations by UniProt. The impact of mutations is coloured differently. Likely disease in red, benign in light green and predicted mutations by computational tools in blue. The panel allow to filer mutations according to mutations impact and curated mutations form different databases such as ClinVar.

Table 1. Putative destabilising spike mutations

mutation	mCSM	SDM
Y508D	-3.736	-2
L962H	-3.07	-2.02
I805T	-2.842	-2.04
V1033G	-3.006	-2.06
F186S	-2.633	-2.1
F65S	-2.986	-2.1
F429S	-2.861	-2.15
Y495D	-2.823	-2.21
I1115T	-2.9	-2.23
F58S	-2.654	-2.28
F800S	-2.798	-2.28
G1251E	-2.861	-2.29
W353S	-3.385	-2.29
L877Q	-2.709	-2.3
G107D	-2.692	-2.32
I128N	-2.974	-2.34

L223S	-3.257	-2.38
F497S	-2.817	-2.47
Y91S	-3.4	-2.48
I410N	-3.171	-2.5
V781D	-3.361	-2.51
Y508N	-3.053	-2.52
F927S	-2.68	-2.54
I742T	-3.005	-2.54
F782S	-3.077	-2.54
R34S	-2.533	-2.55
I119S	-3.031	-2.55
F86S	-3.365	-2.55
W1102S	-2.842	-2.56
V130G	-2.706	-2.57
V193G	-2.876	-2.57
V127D	-2.621	-2.59
I980T	-3.002	-2.59
I402N	-2.584	-2.6
Y266D	-2.836	-2.6
I326T	-2.798	-2.61
Y265N	-2.606	-2.63
I931T	-2.782	-2.64
L425S	-3.37	-2.64
F329S	-2.863	-2.66
V736G	-2.52	-2.69
V722G	-2.521	-2.69
V433G	-2.815	-2.69
F888S	-2.968	-2.69
Y453S	-3.132	-2.69
Y91N	-3.138	-2.73
F135S	-2.733	-2.76
I980N	-2.735	-2.76
F1075S	-3.041	-2.76
F1095S	-3.376	-2.76
Y91D	-3.459	-2.79
I587T	-2.662	-2.8
V781G	-2.715	-2.8
F275S	-2.918	-2.82
Y265D	-3.133	-2.82
Y279N	-3.328	-2.85
A672D	-2.685	-2.86
A1015D	-2.692	-2.91

I418N	-3.012	-2.92
F392S	-3.201	-2.96
F565S	-3.223	-2.96
F718S	-3.273	-2.96
F543S	-3.376	-2.96
F79S	-3.546	-2.96
F140S	-3.583	-2.96
I233S	-2.525	-2.97
L56S	-2.801	-3.07
I997T	-3.001	-3.12
I818T	-3.206	-3.12
I410T	-3.255	-3.12
I1081T	-2.969	-3.13
I923T	-3.009	-3.16
Y265S	-2.75	-3.2
I235S	-2.806	-3.24
L492S	-2.58	-3.28
L229S	-2.902	-3.28
L141S	-3.729	-3.28
I714T	-2.569	-3.29
V1104E	-2.927	-3.38
L650S	-3.403	-3.4
R1000G	-2.75	-3.46
V130D	-2.793	-3.62
L241S	-3.371	-3.63
L277S	-3.755	-3.63
L763S	-2.853	-3.7
L878S	-3.208	-3.7
L959S	-2.805	-3.71
L996S	-3.057	-3.78
I402S	-2.922	-3.83
A903D	-2.563	-3.85
A1022D	-2.997	-3.85
I1081S	-3.414	-4.14

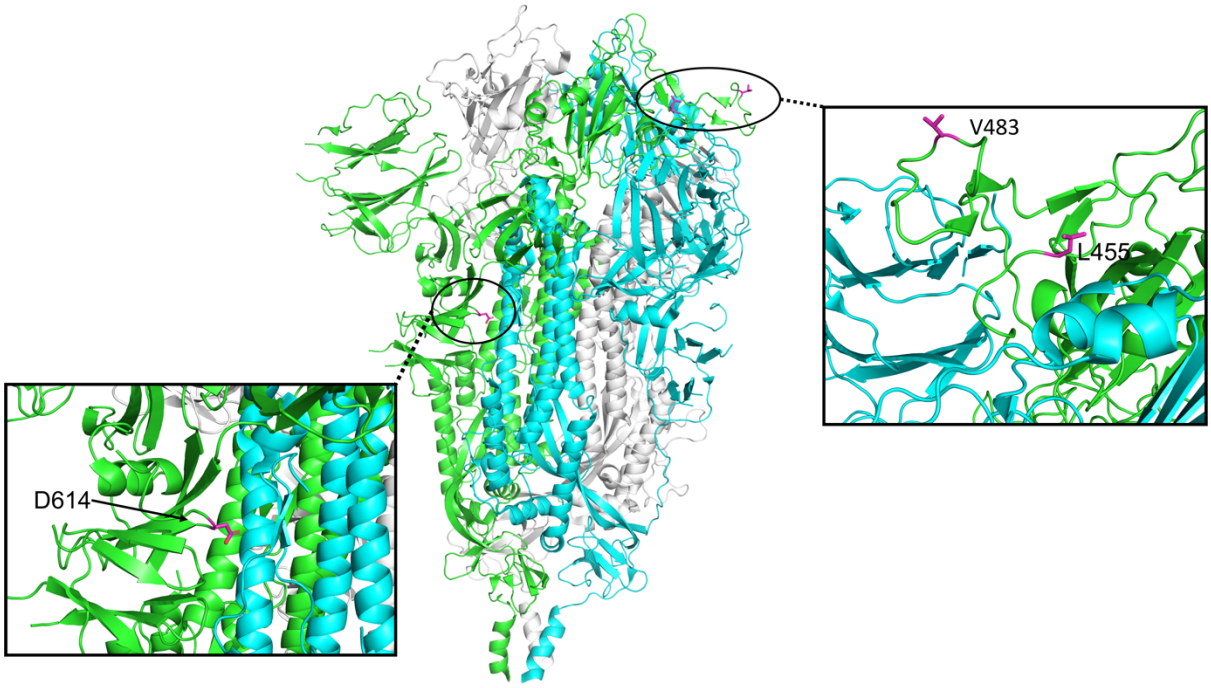


Figure 2. The most proposed detrimental residues in the spike protein. Two of these mutations in the receptor-binding domain are coloured in magenta at the spike-ACE binding site interaction, whereas the third mutant D614 appears close to the furin protease cleaving site.

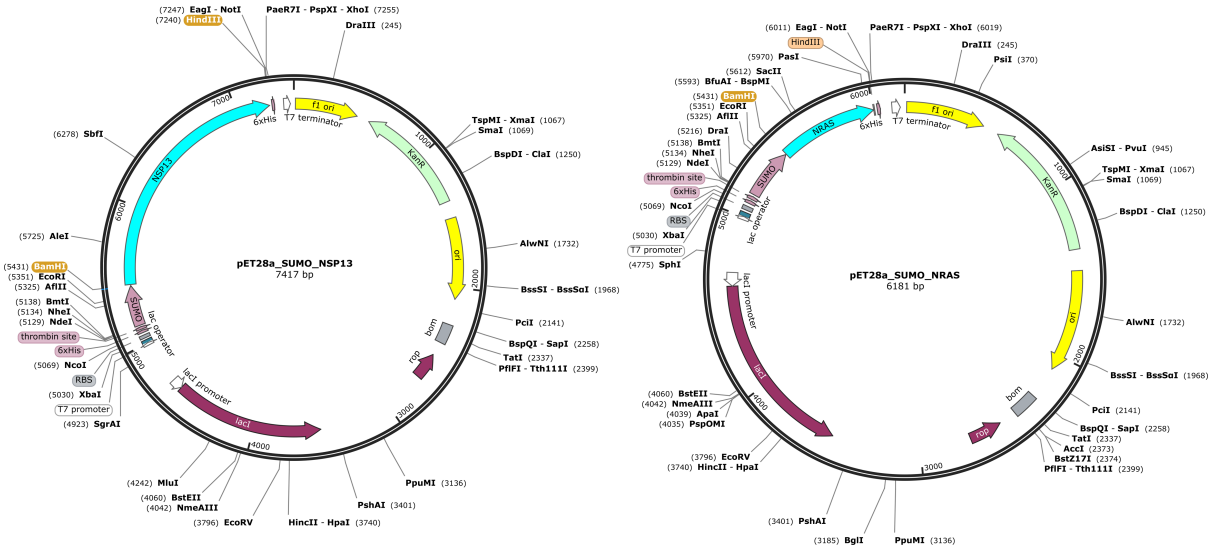


Figure 3. Map of pET28a_SUMO plasmid contains NRAS and nsp-13 genes. The plasmid also contains the antibiotic resistance gene, the bacterial origin of replication and multiple cloning sites.