MONASH University

Associations between structural topology of cancer drivers, and cancer evolution

Yoann Li Youn Fong

Student ID - 33061920

Supervisor 1

Assoc. Prof. Vivek Naranbhai

Lab Head: Laboratory of Translational Immunology

Supervisor 2

Prof. Tim Dwyer

Embodied Visualisation, Faculty of IT

(WORD COUNT: 7911)

A paper submitted for FIT5210 Masters Thesis at

Monash University in 2024

Faculty of Information Technology

PART 2: The Research Paper

Chapter 1

Abstract

Cancer, driven by somatic mutations, disrupts protein function through driver mutations in oncogenes (OCGs) and tumour suppressor genes (TSGs). This study employs structure-based network analysis, a method to understand the structural topology of proteins, to investigate the functional consequences of mutations in cancer driver genes, with a particular focus on lung cancer. By systematically mapping the network properties of each amino acid residue across cancer driver genes using 3D crystallographic proteins structures, and matching these to protein function and cancer mutation frequency in thousands of patients, we found that structural topology of cancer drivers is strongly linked with cancer evolution. Mutations in highly networked residues were strongly associated with significant functional impairments, highlighting the critical role of central residues in protein stability and function. In lung cancer, distinct mutation patterns differentiate TSGs, which mutate in highly networked residues in order to lose function, from OCGs, which preferentially mutate in poorly networked residues to retain or gain function, thereby driving cancer. A novel visualisation method developed using PyMOL functionalities facilitates intuitive mapping of network scores onto protein structures, revealing regions of high functional importance and providing insights into structural and functional implications of mutations. Future research should aim to expand to all cancer genes. Beyond cancer research, the visualisation tool can adapt to visualise diverse protein metrics and facilitate comparative structural analyses.

Chapter 2

Introduction

2.1 Cancer is complex and hard to predict

Cancer is one of the leading causes of death globally and in Australia [10, 102] . Characterised by uncontrolled cell growth, it can originate in various parts of the body and is primarily caused by somatic mutations in DNA (or other less frequent genomic errors), which occur during life and differ from inherited germline variants. These mutations affect genes, which are segments of DNA encoding proteins essential for various cellular functions. Consequently, mutations can impair protein function, leading to cancer. Driver mutations directly contribute to the development and progression of cancer while passenger mutations arise alongside drivers but do not directly drive cancer [16]. Patients with cancer often have numerous accumulated mutations in their tumours due to genomic instability, a hallmark of cancer [65], but only a fraction are drivers [16].

Two primary mutation types drive cancer: loss-of-function mutations in tumour suppressor genes (TSGs), which inhibit cell growth, and gain-of-function mutations in oncogenes (OCGs), accelerating tumour cell proliferation. A recent study has demonstrated that these mutations have distinct impacts on protein structure, which can influence how they drive cancer progression [40]. It has also been suggested that cancer development follows an evolutionary process [22, 13]. At the functional level, proteins are exploring new mutational space and acquiring mutations that allow cells to grow and evade the immune system. These mutations are subject to selective pressures and constraints, resulting in non-random patterns of mutations called mutational signatures that have been observed in large cohorts of human cancer genomes [5]. Therefore, a key aim of cancer research includes identifying driver events [35] and understanding the evolutionary dynamics of tumours [58], in order to develop effective treatments and predict cancer progression.

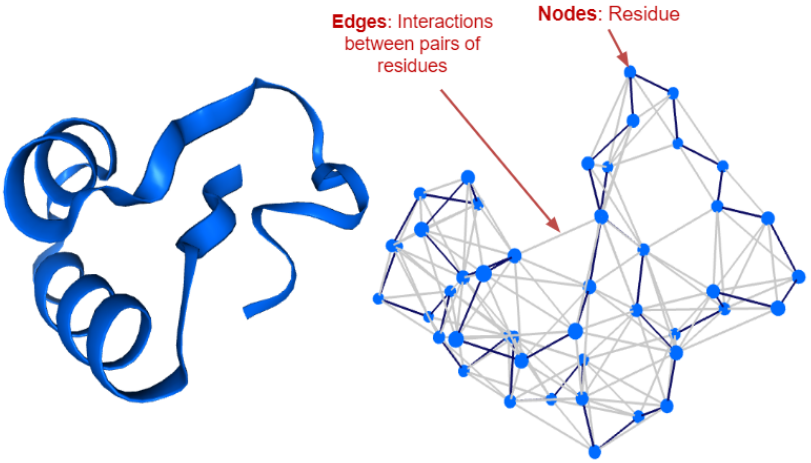

FIGURE 2.1: A protein structure is shown as two representations. Typical representation as a ribbon diagram shown on the left depicts how the protein (sequence of amino acids) folds. Representation on the right as a network shows residues as nodes and interactions between pairs of residues as edges. (Adapted from Chakrabarty & Parekh [26])

2.2 Proteins are three-dimensional structures that can be represented as networks

Proteins are made of a sequence of amino acids which fold into a three-dimensional (3D) structure. There are 20 different amino acids in humans, each with different biochemical and physical properties, so different sequences will fold differently [4]. Various isoforms of a protein may exist, differing in their spatial arrangement, yet they typically perform the same function [45]. Techniques like X-ray crystallography, electron microscopy, and nuclear magnetic resonance spectroscopy are usually employed to ascertain those structures [66]. Importantly, the structure of the protein determines its function - its stability, resilience against mutations, how it interacts with other proteins and molecules, and its responsiveness to drugs [44].

Proteins are typically represented as a ribbon or cartoon diagram, which conveys the fold of the protein [53]. However, multiple studies have extended the representation of proteins as networks of amino acids. In the context of a protein and this paper, the amino acid is called a "residue". The residues are represented as nodes and the interactions between them are represented as edges (Figure 2.1). These networks or graphs have been termed residue interaction networks [7, 43], protein structure networks [42, 94] and protein contact networks [70] and have been used to identify binding sites, functional residues, hotspots of high interactions, understand allosteric effects, and critically, how mutations affect protein structures [82]. We will refer to these methods in general as residue interaction networks for consistency.

2.3 Effective structural molecular biology studies require good visualisation tools

Structural molecular biology has always been dependent on the strength of visualisation tools [68]. Visualisation offers important support for understanding the complex molecular world that is not always intelligible due to its microscopic nature. More specifically, tools are used to be able to create and test hypotheses, usually regarding how molecules interact and function, and to also present findings [60]. By understanding the atomic spatial arrangements and the intermolecular forces, researchers are able to gain insight deeper insight into their functions. The increase in data and availability of structures, such as those in the Protein Data Bank (PDB) [14], underscores the need for effective visualisation. At the time of writing, the PDB contains over 220,000 experimentally determined 3D structures of proteins [76].

2.4 Research aims

Studies have demonstrated that disease-causing variations can significantly affect the structural integrity of proteins [93, 100], highlighting a crucial relationship between molecular structure and the progression of diseases. Therefore, we sought to understand how disruptions of protein structures drive their pathogenic effects in cancer. To advance this understanding, we explored the structural topology of cancer drivers through a particular residue interaction network algorithm, called structure-based network analysis (SBNA), introduced in Section 3.1, and the method explained in Section 4.2.1. **We hypothesised that SBNA will elucidate the patterns observed in cancer evolution, particularly why specific residues mutate more often and how these mutations result in significant functional impacts.** Furthermore, we also aimed to develop enhanced visualisation techniques to enhance our comprehension of cancer mechanisms. Ultimately our efforts of analysis and development of a visualisation tool are to aid in the development of targeted therapies.

Our study is structured as such: we review prior studies using residue interaction networks in broader and cancer-specific contexts, and assess tools for network analysis and visualisation. Our methodology involves identifying cancer-related genes, computing network scores to quantify residue centrality, and integrating datasets on residue function and mutation frequencies. Our analysis focuses on: (1) exploring factors influencing network scores, (2) linking network scores to functional outcomes, (3) studying mutation patterns in lung cancer using network scores, (4) analysing changes in network scores due to mutations, and (5) leveraging visualisation to give deeper insights into our findings.

Chapter 3

Related work

3.1 Residue interaction networks

Network theory, also termed graph theory, is a powerful tool for understanding and representing complex relationships between entities. It has various applications including in computer science, sociology, engineering, physics, biology, and more [104]. In residue interaction networks (RINs), residues are the nodes and the interactions based on the distance or intermolecular energy between them are the edges. The earliest attempts to use this representation found their usefulness in identifying key residues involved in the stability and folding of the protein [51, 95]. Since then, there have been many methods to construct those networks and analyse them through different centrality metrics, leading to numerous applications. As introduced previously, these various applications usually revolve around finding residues or regions that are functionally important in the protein.

RINs can be **constructed in various ways**, especially in how nodes and edges are defined. Amino acids all contain a C_α atom, a C_β atom and a side-chain, except glycine which does not have a C_β atom [1, 2]. Common choices for the nodes include the C_α atom [25, 70], C_β atoms [9], or the side chain [94, 36]. Edges between nodes are formed based on the distance between residues, with a predefined cutoff distance. Research using both C_α and the side chain as nodes has shown that an optimal cutoff distance irrespective of those choices was 7.0 Å [83]. A later study using side chains and energetic forces argued that 5.0 Å was the optimal value [97]. In fact, several studies use different values for their cutoff distances. Moreover, while some earlier applications have used unweighted graphs for simplicity [11, 86], we observed many later studies using weighted approaches, usually based on the inverse distance between residues or interaction strength [25, 31, 98].

Centrality metrics are intended to quantify the relative importance of nodes in the network. Just like network construction techniques are diverse, network analysis methods also vary widely. Importantly, Brysbaert and Lensink [18] offered a comparative analysis of common measures such as betweenness, closeness, degree and more. They concluded that combining those metrics resulted in greater accuracy in identifying key residues, but ultimately the choice of metrics depended on the analysis context [18].

Despite these advances, there has been a lack of application in the field of cancer research. A study performed by Verkhivker [96] used molecular dynamics simulation to study three tumour suppressor genes, TP53, PTEN, and SMAD4, and found that highly mutated residues were structurally stable, high centrality sites, and played crucial roles in the protein. Molecular dynamics, as opposed to crystallographic studies which provide static snapshots of the protein and are used in our study, can explore the time-dependent behaviour and interactions of molecules in a dynamic environment [47]. However, it has not yet been generalised to all genes, and especially to oncogenes.

Recently, a novel residue interaction network algorithm, which we refer to as **structure-based network analysis (SBNA)**, was introduced by Gaiha et al. [36] and was used to shed light on the function of different residues of HIV proteins. Mutations of highly networked residues had a greater impact on the function of the protein and occurred at higher rates. Another study by Hauser et al. [46] applied the same SBNA algorithm to study retinal diseases and found an association between highly networked residues and their pathogenicity. Although the SBNA algorithm has shown promising results in viral and retinal diseases, SBNA has not yet been applied to cancer genes, forming the basis of this study.

It is worthwhile to note that SBNA uses both an energetic network and a distance-based network (the algorithm is explained in Section 4.2.1). We found that previous studies often use either an energetic or a distance-based network, but not both, potentially leaving some aspects uncaptured. It was shown that both methods are effective in presenting small-world characteristics, a feature defining complex networks [38], and allostery [84]. By integrating both types of networks, SBNA represents a promising approach for a more complete understanding of the "networkness" of residues in the network.

3.2 Tools for analysis and visualisation

To perform RIN, many tools are available today that offer both analysis and visualisation simultaneously. For example, NAPS is a widely cited web tool that offers various

ways of creating the network and visualising it [25, 24]. RING 3.0 (and previously RING 2.0), is also another web tool that uses probabilistic models to create a network based on non-covalent bonds and molecular dynamics [29, 73]. Other tools such as PyInteraph2 (previously PyInteraph) [94, 89], and PyProGA [85] work as plugins in PyMOL, a popular general-purpose molecular visualisation tool [107]. PyInteraph2 is used to assess non-covalent interactions between residues while PyProGA is used more to understand protein-to-protein interactions and ligand binding. For a larger review on existing tools, Liang et al. [57] provide another resource.

However, many of the current tools suffer from poor usability and have been made for specific analyses [80]. Moreover, plugin-based tools are restricted to their respective software environments, potentially requiring complex installations and specific system requirements [25]. Importantly, there remains a notable absence of a dedicated method for visualising centrality metrics directly onto the protein structure. Furthermore, most tools do not provide 3D network representations of proteins, except NAPS (as far as current knowledge indicates), albeit NAPS is constrained by its own predefined centrality metrics.

Chapter 4

Methodology

In this study we curated cancer driver genes, their corresponding 3D structures, computed network scores for them, compared the network properties to a range of residue protein properties including function, and developed a new visualisation tool to drive an intuitive understanding of these findings, Figure 4.1 illustrates an overview of the methodology employed for this project for the first four analyses.

4.1 Data sources and preparation

4.1.1 Catalogue Of Somatic Mutations In Cancer (COSMIC)

The COSMIC database contains thousands of somatic mutations that are known to be involved in cancer development [91]. The Cancer Gene Census (CGC) is a key subproject that catalogues genes that are accepted to cause cancer and it has become a global standard in cancer research, used for various applications [87].

From the CGC, we selected 743 cancer genes for our study, with analyses split into two main sets. The first analysis (Analysis 1 in Figure 4.1) focused on investigating residue mutations' functional impacts in three prominent tumour suppressor genes, detailed in Section 4.1.3. The second set of analyses (Analysis 2, 3, 4, and 5 in Figure 4.1) aimed to explore how cancer driver mutations correlate with network scores. Due to the extensive protein structures (14,202) associated with these genes, we concentrated our efforts on lung cancer, the leading cause of cancer-related mortality globally and in Australia [21, 64] and the focus of the host laboratory. Initially, 54 lung cancer genes were identified, but after meticulous curation (described in Section 4.1.2), 34 genes with experimentally resolved protein structures remained.

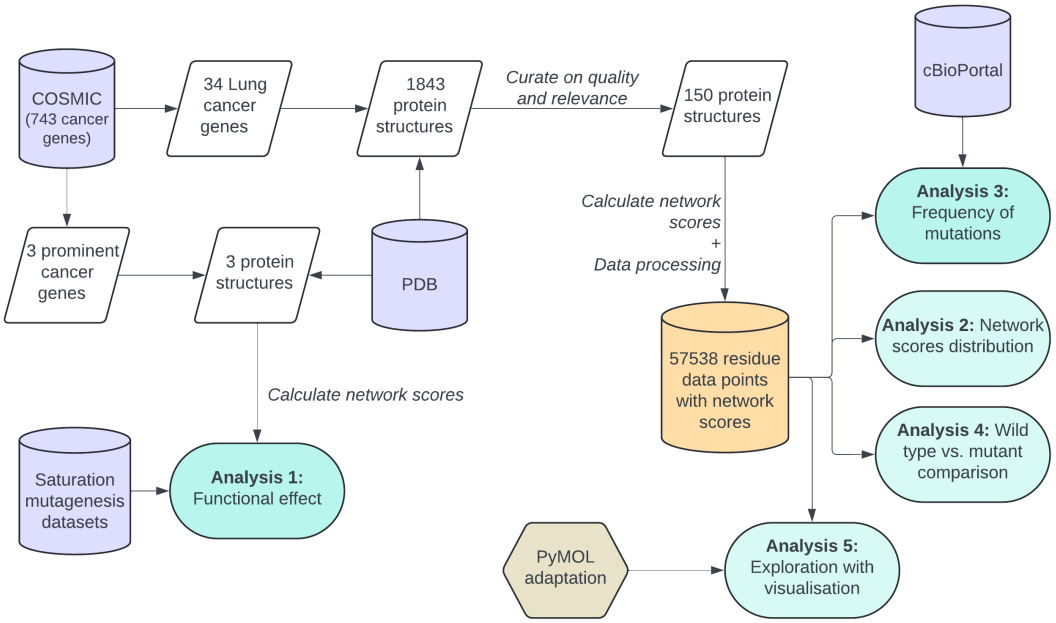

FIGURE 4.1: Overview of workflow from data source to analysis

The lung cancer genes were chosen based on whether the somatic tumour types for that gene contained the word "lung" or "NSCLC" (non-small cell lung cancer). Moreover, a gene may sometimes exhibit multiple roles from tumour suppressor genes (TSGs), oncogenes (OCGs), and fusions (or null-denoting unknown role). For instance, TP53, one of the most commonly mutated genes in cancer [67], is classified under all three categories. This is because while its primary function is tumour suppression, certain mutations can also confer oncogenic properties [77]. To simplify our analysis, we therefore assigned a single role to each gene based on its primary function based on manual curation of biological function by experienced cancer biologists on the team. Specifically, combinations of OCG and fusion were categorised as OCG. For individual genes, TP53 was classified as a TSG, NOTCH1 as an oncogene, and ERBB4 as an oncogene.

4.1.2 Protein Data Bank (PDB)

The Protein Data Bank (PDB) is a database containing experimentally determined 3D structures of macromolecules such as proteins. The data is stored in a .pdb file that contains the spatial coordinates of all atoms in the protein. Each gene in the CGC typically has one or more associated protein structures. For instance, TP53 has 268 experimental structures. Individual PDB accession numbers for the genes were obtained using the MyGene.info API [55, 103, 105], providing easy linkage between gene and PDB accession numbers. These structures may represent various domains and conformations, crystallised with different ligands, sometimes alongside other proteins, and under diverse

experimental conditions and methods. Moreover, some protein structures may exist as a complex of multiple protein subunits which we call the oligomeric state. Therefore, it was important to curate the list of structures for analysis.

We list the conditions for choosing the structures here:

- Resolution (level of detail of the experiment used to determine the structure) is less than or equal to 3.5 Å.
- If multiple oligomeric states were available, only the highest order was chosen. We prioritised the more complex forms in order to capture the full complexity of protein interactions and better understand their functional significance.
- Proteins in complex with proteins of different genes were excluded. This was to ensure that the network construction focuses solely on the proteins of interest, avoiding potential data distortion from other complexes.
- Both wild-type (non-mutated form) proteins and mutant versions were chosen. We distinguish between the two in our analysis.

If no relevant structures remained after applying these criteria, the respective gene was excluded from further analysis. The complete list of structures is listed in Appendix A. Specifically for the saturation mutagenesis analysis, we used similar protein structures (genes PTEN and BRCA1) to those used in previous studies by [46].

4.1.3 Understanding the function of drivers: saturation mutagenesis datasets and cBioPortal database

To understand how the network approach could give insight into cancer drivers and evolution, we used the two widely used methods for identifying driver mutations over passenger mutations: those that predict the functional consequences of mutations and those evaluating mutation frequency [75]. Our objective is to gain a distinct perspective from these datasets, offering new insights into cancer drivers and evolution.

To understand the functional impact of mutations, we used the data from three saturation mutagenesis experiments. These experiments entail deliberately mutating each residue of the protein and measuring the resultant effects on protein function in comparison to the wild-type protein. We obtained data for the three genes TP53 [52], BRCA1 [33] and PTEN [62] which are prominent genes frequently mutated in cancer and extensively studied in this context.

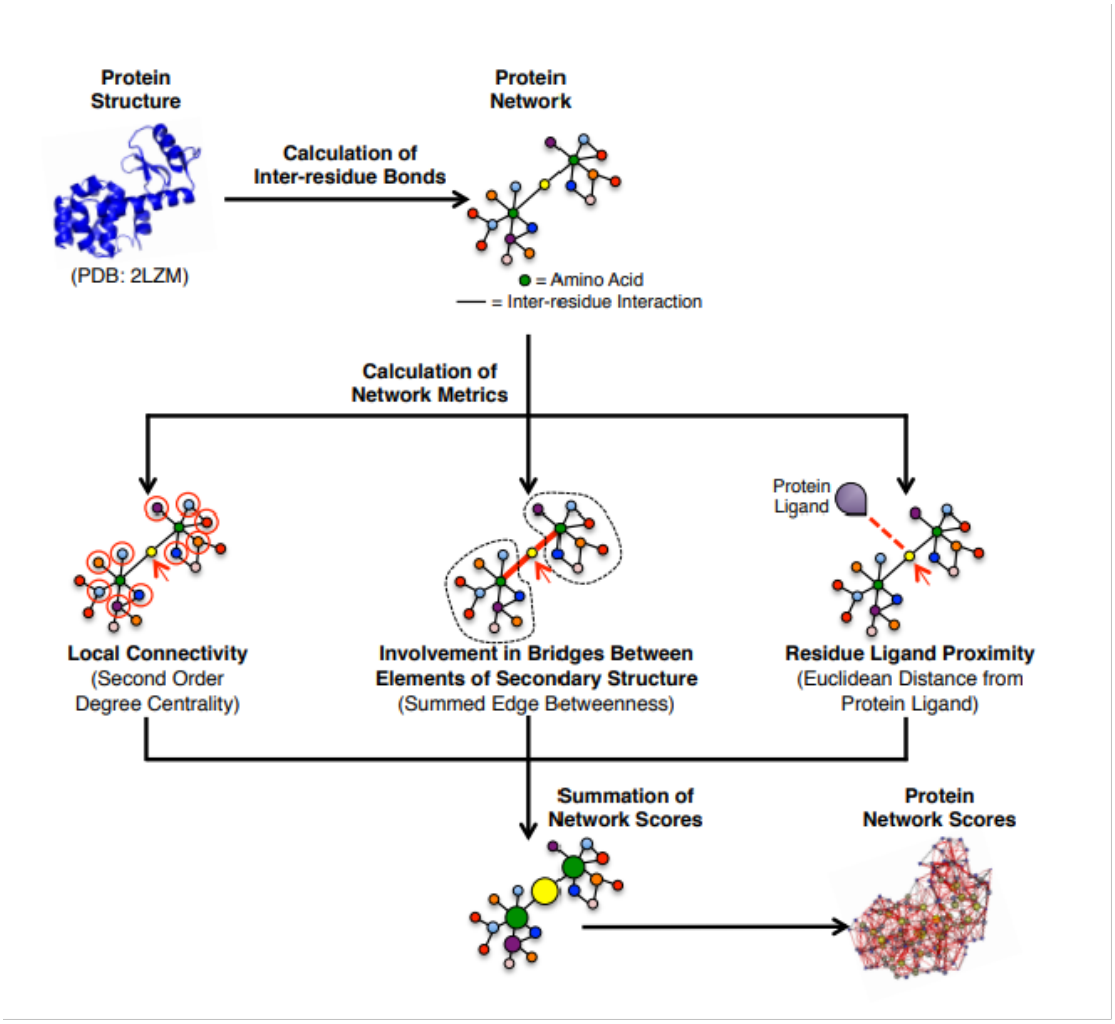

FIGURE 4.2: Atomic coordinates from PDB files are used to identify inter-residue interactions through energy forces, angle and distance thresholds, and distances between side-chain centers of mass. Centrality measures - such as second-order degree centrality, summed node-edge betweenness centrality - and residue ligand proximity, are calculated, as illustrated in the network schematic for the central node (yellow). These measures are converted into Z-scores, which are then summed to produce composite network scores for each amino acid residue in the protein. The size of each residue visually represents these scores. This image is adapted from the supplementary materials of Gaiha et al., [36].

The cBioPortal database provides data from large-scale cancer genomics studies [17, 23, 37]. We used it to extract the frequency of mutations of 14262 patients for the 34 lung cancer genes.

4.2 Network score calculation and integration

4.2.1 Network score calculation

To perform our network construction and analysis, we used the same SBNA algorithm developed by Gaiha et al., [36] which was effectively used in the viral and retinal disease contexts [46]. It calculates a network score for each residue in the protein and is a measure of how central it is in the network. We used the code for the algorithm provided by the Walker Lab [92].

To construct a network for a particular protein, its structure is obtained from the PDB, containing atomic coordinates for the protein. Both energetic forces (energetic network) and geometric distances (centroid network) are considered to create two network representations. In the energetic network, the nodes are the residues and weighted edges are created representing cumulative non-covalent bonds. The centroid network uses the center of mass of the amino acid sidechain as node, and the edges are unweighted and are created if two pairs of residues are within 8.5 Å. A network score is then calculated using network parameters as such

$$\text{Network Score} = SD + NEB - LD$$

where SD is the second-order intermodular degree (number of second-order interactions between residues in different higher-order structures), NEB is the node-edge betweenness (frequency a node's edges are used as shortest paths in the network, weighted by edge weight), and LD is the Euclidean distance from a residue's centroid and the ligand's center of mass. Each network parameter is normalised into Z-scores before summing to derive the final network score. This metric provides a comparative assessment of residue centrality within the protein network, rather than an absolute quantification. We abbreviate "network score" to NS in this paper.

4.2.2 Integrating results with other datasets

Aligning the sequences was an important part of our work for integrating the network scores with other datasets. PDB structures often present challenges due to misalignment with UniProt sequences, and the SBNA algorithm also complicates the process of aligning. The sequence of a protein comprises the sequence of amino acids and their position in the sequence but it can sometimes be in relation to that specific protein or to the whole gene. This alignment process also becomes complex when dealing with multiple structures for the same gene. Different authors may use varied methods and

conventions, leading to inconsistencies in numbering. Furthermore, the chain denotation in the PDB file, labelled by the author, often differs from the one used by the PDB itself. SBNA returns values without accounting for different chains overlapping in number, which means we could have two network scores for the same residue without distinguishing between them. Additionally, PDB files may contain multiple separate sequences for the same protein.

To address these issues, sequences were aligned with UniProt, which serves as a reference point for cBioPortal, COSMIC, and saturation mutagenesis datasets. We used the RCSB PDB Data API [78] to map author-labelled chain names to the PDB chain names. We also used the API to extract the start and end of each separate sequence if multiple sequences were present. Alignment was performed using Biopython `Bio.Align.PairwiseAlignment` method in global alignment mode, mapping PDB residue numbers to UniProt identifiers. Once the mapping was done, we could proceed with further analyses. We provide the code for this alignment method in Appendix B.

Notably, all calculations, data processing and integration were done in Python and pandas. Statistical analysis such as correlation coefficients and ANOVA were performed using the SciPy library.

We also obtained groupings for the amino acids from Wikipedia [1].

4.3 Visualisation of network scores using PyMOL

We developed a Python program that generates a PyMOL script leveraging the latter's flexibility and functionalities to enhance protein structure visualisation by colouring the structures by network scores (or any score), creating a network representation and offering side-by-side comparisons of different structures. We previously explained the limitations of using PyMOL as a program for visualisation, being limited by the need for installation and system requirements (Section 3.2) - however, it is a widely-used tool among biologists for structural analysis [30, 63, 79] and we argue that its powerful features and flexibility outweigh these constraints. PyMOL accepts scripts with commands that perform various actions within its application, enabling customised visualisations, rotations, zoom and more. Our program takes in the PDB accession number and the NS for each residue and outputs the script to be entered in PyMOL.

To **colour the protein structures**, we adapted the `pdb_color_generic.py` script from the `pdbcolor` GitHub Project [34]. We acknowledge his work and the foundation it provided for our enhancements. This adaptation allows for multiple structure handling by adjusting the colour range. Considering that network scores are relative and not

absolute, we used a divergent colour palette where the lowest scores are coloured blue, and the highest scores are coloured red. This palette is centred around the middle of the score range for network scores. Importantly, it also allows the visualisation of different scores with different palettes and methods for colouring, providing flexibility for future works.

To **create the network representation**, we manually created sphere objects for each residue based on the spatial coordinates of that residue's C_α atom. The C_α atom is one of the most widely used nodes for network representation since it "captures very well the 3D topology of protein structure." [25]. Edges are created using bond objects when the centers of mass of the side chains of two residues are within 8.5 Å of each other, which is provided from the results of the SBNA algorithm.

The final visualisation is generated through a `.pml` script that is fed into PyMOL. This script is a sequence of PyMOL commands that automate the visualisation process. The code for this visualisation is provided in C.

Chapter 5

Results and Discussion

5.1 Understanding network scores distribution

5.1.1 Different experimental conditions and mutations give rise to different network scores

Proteins are not static, but dynamic [72]. They are highly dependent on their environment. Experimental conditions, such as temperature, pH, and the presence of ligands, can lead to differing conformations captured in crystallography experiments. Mutations of amino acids in the sequence can also introduce changes in the structure [90], and even in regions other than the specific mutation site through allosteric effects [41]. Thus, the same gene can have multiple structural variations, affecting the NS calculated from static snapshots of the protein.

We therefore compared NS for the same gene across multiple structures. Figure 5.1 provides a series of figures showing the observed differences in NS. Panel (A) shows a correlation matrix for nine different protein crystal structures for the TP53 gene. We note a range of correlations between 0.63 and 0.91, indicating structural diversity. Panels (B) and (C) comparing pairs of structures with different mutations show that differences in network scores arise because of the presence of a ligand or not, but also the node-edge betweenness and to a lesser extent, second-order intermodular degrees. For example, panel (B) shows the R273C mutant exhibiting enhanced spatial connectivity compared to the wild type. In Panel (C), comparing V272M and R280K mutants, one subunit displays significant movement. While some disparities may stem from inherent protein flexibility [8], NS offer insights into these conformational changes.

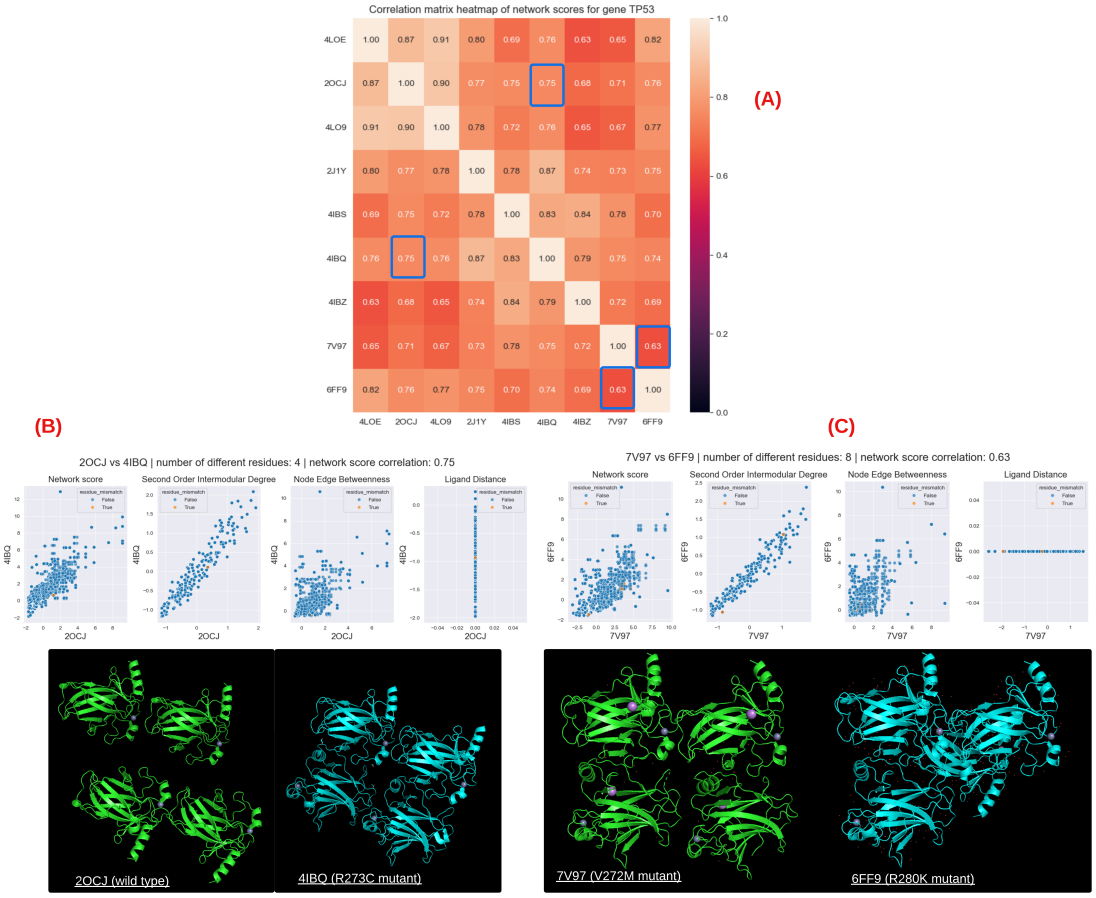

FIGURE 5.1: Analysis of network scores and structures for TP53 gene. **(A)** Correlation matrix of network scores TP53 gene. Blue boxes highlight the correlation between 2OCJ and 4IBQ, and 7V97 and 6FF9. **(B)** Comparing PDB 2OCJ (wild type) and 4IBQ (R273C mutant). Scatterplots show the network score, second-order intermodular degree, node-edge betweenness and ligand distance for the two structures. The bottom image shows the crystal structures. **(C)** Comparing PDB 7V97 (V272M mutant) and 6FF9 (R280K mutant)

5.1.2 Network score differs by amino acid type

Next, we explored how the NS for each residue vary by amino type, categorised by the physico-chemical properties of their side chains into hydrophobic, polar uncharged, polar charged, polar charged negative and special cases and plotted their distributions (Figure 5.2). Amino acids' side chains significantly influence their properties and interactions within the protein structure [59].

Our analysis showed that hydrophobic amino acids tend to have higher NS, consistent with their clustering in the protein's interior to avoid water, which is central to maintaining structural integrity [4]. Conversely, polar side chains are typically found on the exterior of the protein, explaining their lower NS. Statistical analysis using ANOVA for these five groups yielded a p-value of 8.47e-197, indicating that the differences in NS

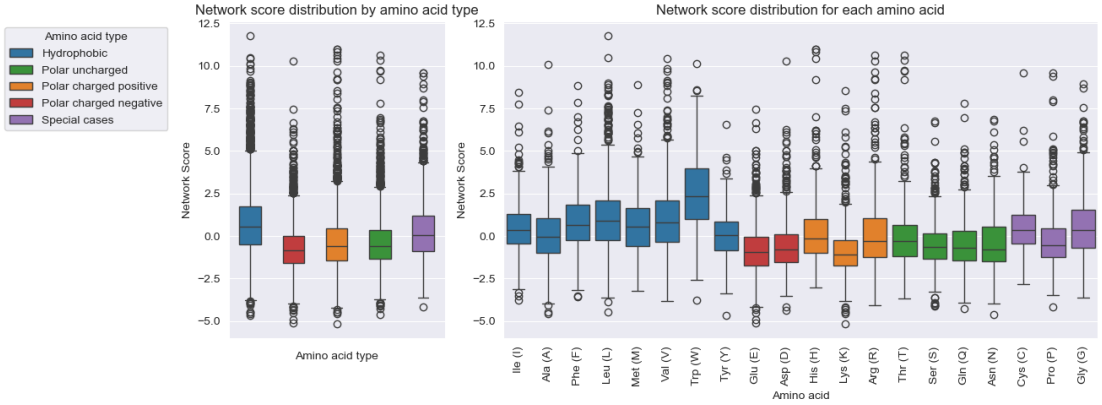

FIGURE 5.2: Distribution of network scores by amino acid type and individual amino acids. Statistical test for difference between for these five groups (ANOVA) showed a p-value of 8.47e-197.

between these groups are statistically significant. Moreover, tryptophan in particular exhibited significantly higher NS compared to all other amino acids. It has the unique properties of having the largest hydrophobic area and "bioenergetically the most expensive amino acid to produce", meaning it is used selectively in proteins, and only appears at crucial sites essential for the protein's function and structure [12], explaining its high NS. These data help rationalise how amino acid properties may differentially affect network scores. Moreover, they suggest that there may be a relationship between NS and function, which is studied in the next section.

5.2 Higher network score is associated with higher functional effect

To explore how NS relates to protein function, we examined functional scores obtained from saturation mutagenesis against the NS of three tumour suppressor genes BRCA1, PTEN and TP53 (see Figure 5.3). Our results show a linear association between mutations in highly networked residues and greater functional impact - particularly a loss of tumour-suppressing abilities in BRCA1 and PTEN genes. This is consistent with previous studies that have suggested a relationship between central residues and function [56, 86].

For TP53, the relative fitness score (RFS) measures cell growth in the presence of mutated residues compared to the wild type (non-mutant). Therefore, Kotler et al. [52] suggested that high RFS could indicate a possible disruption in tumour-suppressing function and low RFS, a retention of normal function. Moreover, they also observed a strong positive correlation between RFS and evolutionary conservation score (ECS).

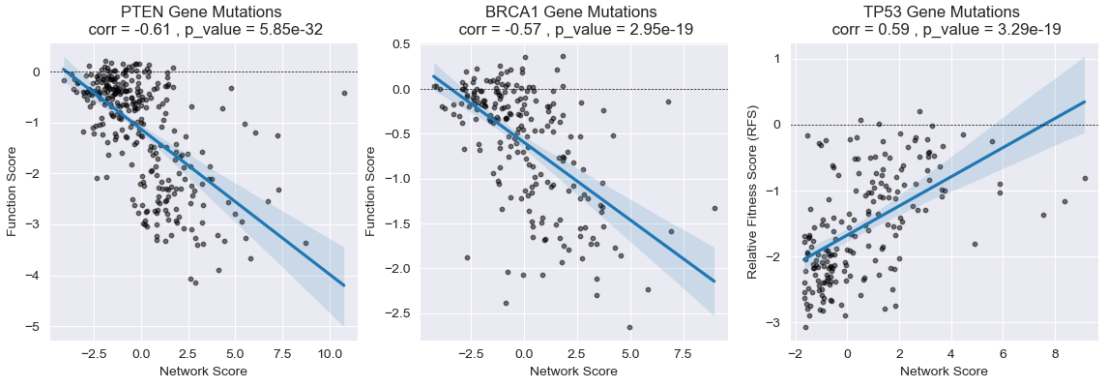

FIGURE 5.3: Functional consequences of mutating residues in BRCA1 and PTEN genes against network score. The function score is a relative measure of the mutant protein's effectiveness compared to wild-type. Mutations in highly networked residues lead to greater functional impairment. The third plot shows the relative fitness score (RFS) for TP53 against NS, where higher RFS indicates increase tumour cell growth, hence greater loss of tumour-suppressing activity and is linked to higher networked residues.

The ECS for a given amino acid position quantifies how preserved it is within a protein family by comparing sequences from multiple species [81]. It is well established that natural selection "conserves functionally important residues in proteins in order to preserve biological activity" [74]. Comparing NS to RFS for TP53 we found that mutations at highly networked residues (that likely inhibit TP53 tumour suppressive function) are associated with greater cell growth.

Importantly, these results suggest that the NS calculated with SBNA offers a valuable way of understanding the function of a particular residue. It can potentially reveal an important mechanism in how cancer evolves.

5.3 Network score constrains lung cancer evolution

Given the saturation mutagenesis results, we hypothesised that cancer targets highly networked residues in TSGs to induce loss of function. Conversely, OCGs would likely mutate in a way that leads to a gain of function. Hence, we filtered for 96 wild-type structures out of the 150 total structures we obtained as described in Section 4.1.2, calculated their network scores, obtained the mutation frequencies for each residue in lung cancers from 14262 patients obtained from cBioPortal, and averaged mutation frequencies by network score bin. The counts for the gene and crystal structures are shown in 5.1 and the results are depicted in Figure 5.4. It is worthwhile to note the distinction between using sum aggregation versus average aggregation, as the NS calculation

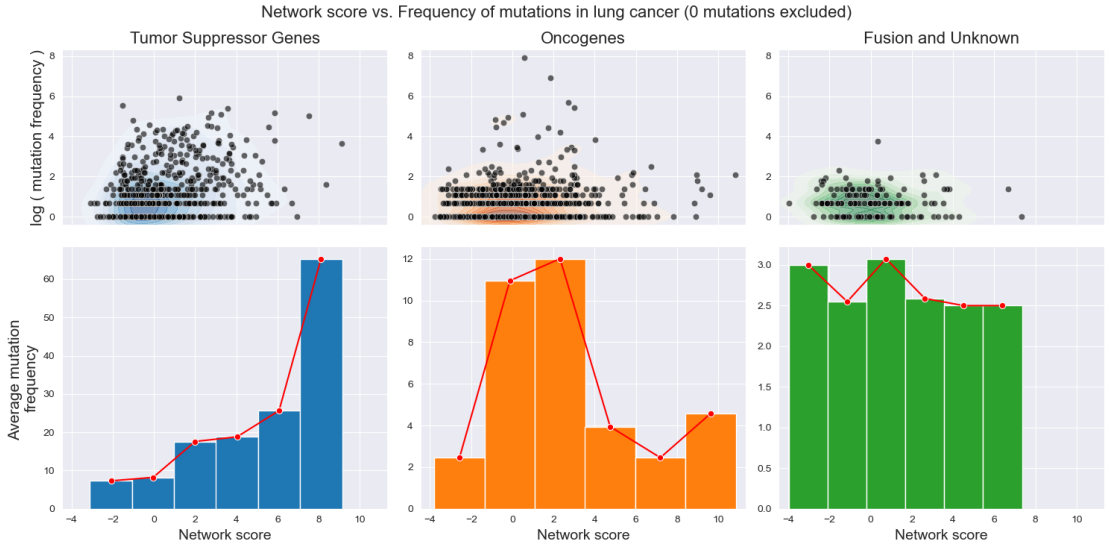

FIGURE 5.4: Network score against (1) log (mutation frequency) as a scatterplot where each point represents a residue position (2) average mutation frequency binned by NS. There is a clear disparity between TSGs, OCGs and fusions/unknown roles. TSGs mutate in highly networked residues, while OCGs mutate in poorly networked ones. Fusion genes and genes with unknown roles serve as control.

Cancer Role	Unique gene count	Unique wild-type PDB count
TSGs	10	24
OCGs	17	57
Fusion and Unknown	7	15

TABLE 5.1: Count of genes and wild-type PDBs per cancer role

involves normalisation, leading to an inherent paucity of residues that are highly networked. Additionally, we excluded residues with no mutations to avoid skewing the average mutation frequency.

Crucially, we found a novel and important observation - there is a clear dichotomy in the distributions between TSGs and OCGs. TSGs' normal function is to suppress cancer. So a mutation that drives cancer requires TSGs to lose function. This is reflected by the greater average frequencies in higher network residues. As discussed in Section 5.2, we showed that those residues lead to greater loss of function. On the other hand, OCGs accelerate the growth of cells when mutated. Therefore, they mutate in poorly networked residues in order to retain or gain function, hence driving cancer. To validate our findings, we included fusion genes and genes with unknown roles in cancer as controls. Unlike TSGs and OCGs, these genes showed no difference in average mutation frequencies across all network scores. This distinct behaviour supports our conclusion and highlights the different evolutionary pressures on TSGs and OCGs. **Importantly, this provides a novel way to understand how cancer evolves - it is constrained by the structural topology of cancer drivers**.

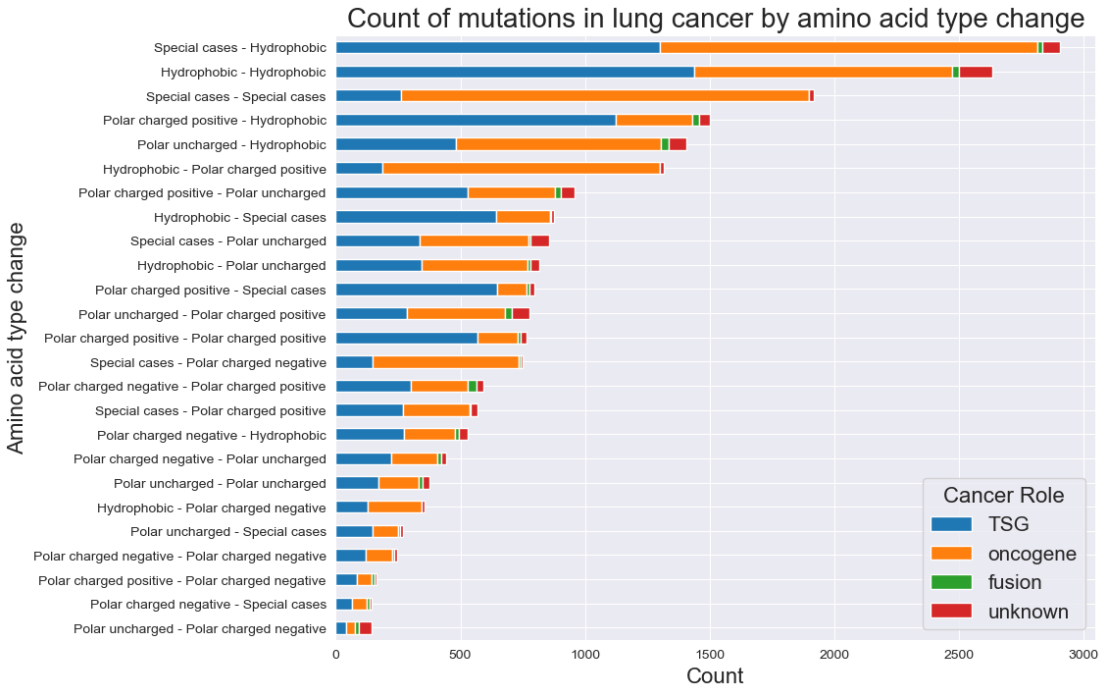

FIGURE 5.5: Counts of mutations in lung cancer by change in amino acid grouping . The label at left details the original amino acid and the mutated amino acid separated by a hyphen (-). The count of amino acid mutations was calculated from genome sequencing of 14262 patients with lung cancer in cBioPortal.

5.4 Mutations affect amino acid properties and therefore network scores

5.4.1 Mutations alter the amino acid type

We analysed the frequency of mutations in 14262 lung cancer patients from cBioPortal by the type of amino acid change (Figure 5.5). Our results show that mutations involving hydrophobic amino acids are very common. This tendency aligns with our previous findings that hydrophobic amino acids tend to have higher network scores and are more likely to be buried within the protein to maintain structural stability. Therefore, mutations could either aim to enhance stability/burial, increasing NS, or disrupt the structure, decreasing NS. It has been argued that the most harmful variation sites are less surface-exposed [71], indicating that more networked residues are typically more buried in the protein, while less networked residues are more exposed. We further validate this argument in Section 5.5 and it is depicted in Figure 5.7.

Furthermore, most mutations involve a change in amino acid type, which is consistent with statistical analyses indicating that the most harmful disease-causing variations tend to cause significant changes in the physico-chemical properties of the mutation

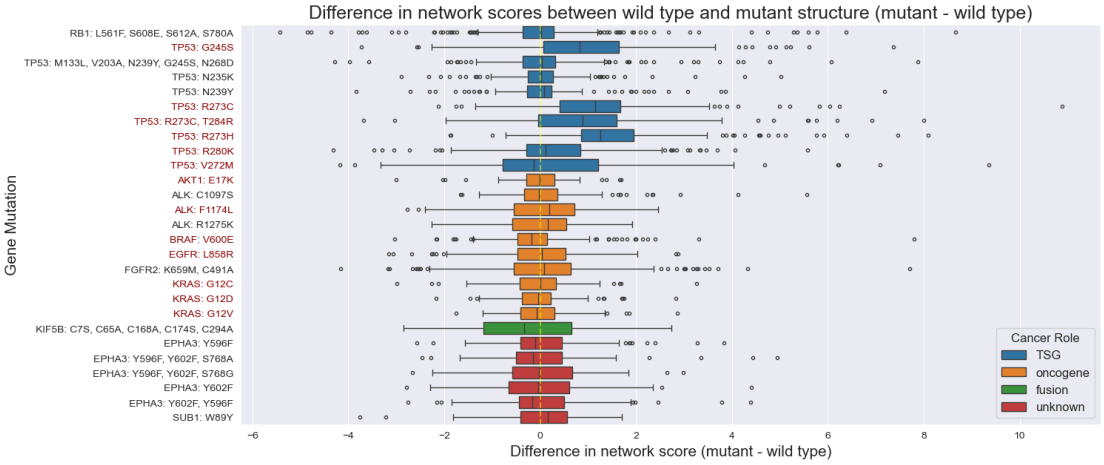

FIGURE 5.6: Difference in network scores between the wild-type and mutant structure. Mutants labelled in red represent mutations that occur in lung cancers according to our cBioPortal data.

site, impacting native hydrogen and salt bridge networks [71]. This supports the idea that disease-causing variations often destabilise or stabilise the protein structure [6, 106].

Mutations involving polar, negatively charged amino acids are also rare. This observation aligns with our previous findings that the network scores of polar, negatively charged residues had a lower distribution (Figure 5.2). These residues may not be structurally critical, so cancer mutations may not target them frequently. Taken together, these data support a model in which cancer mutations drive amino acid preferences that selectively affect network score and hence protein function.

5.4.2 Mutations affect network scores

Since we understand that disease-causing mutations tend to alter the physico-chemical properties of amino acids and the protein's structural integrity, we wanted to understand how the overall NS varies between the wild-type and mutant proteins. Fortuitously, many protein crystal structures have been solved for both the wild type and corresponding to cancer-causing mutant (the list of PDB is given in Appendix A). We therefore calculated the difference in network scores for all the residues between the wild-types and mutant structures, aggregating the scores by position (i.e., combining all network scores for wild-type residues at a specific position and comparing them to the corresponding scores in the mutant structures) (see Figure 5.6).

Among the TP53 mutations, R273 and G245S are among the most common in all types of cancers [28] and were the only ones in TP53 that clearly led to an overall increase in NS. These mutations have been suggested to possibly be oncogenic, indicating a gain of function [99]. The increase in NS suggests that these mutations might enhance protein

stability and shift the protein's role from a TSG to an OCG. These highlight how NS may help identify divergent mutational impacts within a single protein, explaining how mutations in TP53 can both disrupt its tumour-suppressive functions and potentially endow it with new oncogenic properties. On the other hand, OCGs tend to remain more or less conserved upon mutation, which could indicate they retain or gain function to drive cancer.

Importantly, the plot of network score differences also shows a wide range of variations, and this could give insight for further research into how different regions of the protein become more or less stable upon mutations. However, it is crucial to consider the specific contexts in which these mutations occur, as exemplified by mutations like the mutant structures for TP53 G245S (PDB ID: 6FF9) and V272M (PDB ID: 7V97), which occur in the presence of compounds like arsenic trioxide that was said to "rescue" the protein's structure [27, 88]. In such cases, direct comparison of network scores may not be appropriate.

Nevertheless, this analysis connects to the visualisation objective, where visualising network scores can help identify how different regions of the protein are affected by mutations, providing a clearer picture of the structural and functional impacts of these changes. This is explored deeper in Section 5.5.2.

5.5 Novel visualisation technique allows intuitive understanding of network scores and function

Given the extensive analyses conducted in previous sections, it became crucial to develop a visualisation technique to comprehend the structural implications of network scores in a three-dimensional space, and to grasp the functional implications, since structure and function are closely intertwined. Furthermore, we wanted to understand how mutations impact NS, providing an intuitive method to understand structural differences linked to diseases.

The program we developed generates a PyMOL script that can be used in PyMOL's interface, offering several functionalities and capabilities:

- Leveraging PyMOL's intrinsic functionalities, users can rotate, zoom, label residues, align multiple structures, compare multiple structures side by side, and more.
- We created a novel network representation that visualises individual residues and their positions in the network, the method of which is described in Section 4.3.

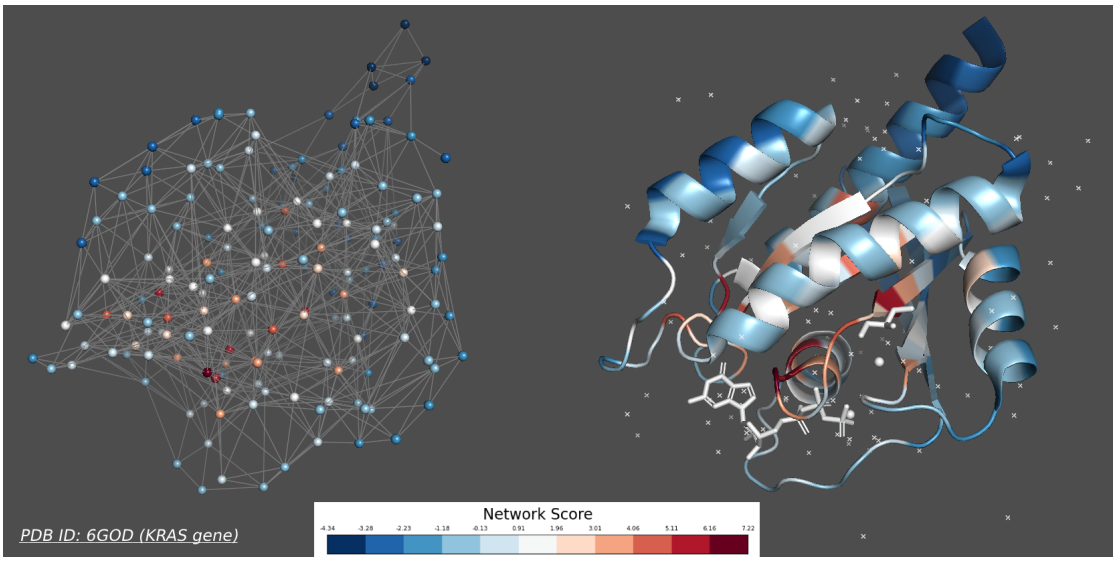

FIGURE 5.7: Side-by-side representations of the residue network and the cartoon representation (depicts fold of protein) automated generated script. The wild-type protein of the KRAS gene is shown and the NS of each residue is mapped onto the structures.

- Users have the flexibility to visualise by colour any selected metric onto the protein structure as well as the network representation. This allows users to visualize metrics like network scores, differences in network scores, or functional scores by residue.

In Figure 5.7 we present an example of the visualisation tool in action as applied to the KRAS protein, an oncogene frequently mutated in lung, colorectal and other cancers [48]. The network representation is displayed alongside a cartoon representation of the protein, depicting the fold of the protein, with residues coloured according to their network scores. This visualisation highlights that highly networked residues tend to be located internally within the protein structure, while poorly networked residues are situated on the protein's surface. This observation aligns with previous findings, where highly networked residues are typically less exposed and mutations in these residues lead to greater functional effects 5.4.1.

In subsequent sections, we will delve into specific case studies and functionalities of the visualisation tool, illustrating how it provides intuitive insights into protein structure, function, and the impact of mutations in cancer.

5.5.1 Highly networked and functional residues in p53 are located near the DNA binding region

The p53 protein, encoded by the TP53 gene, is described as the "guardian of the genome" [54]. It plays a crucial role in protecting our cells by halting the cell cycle (proliferation

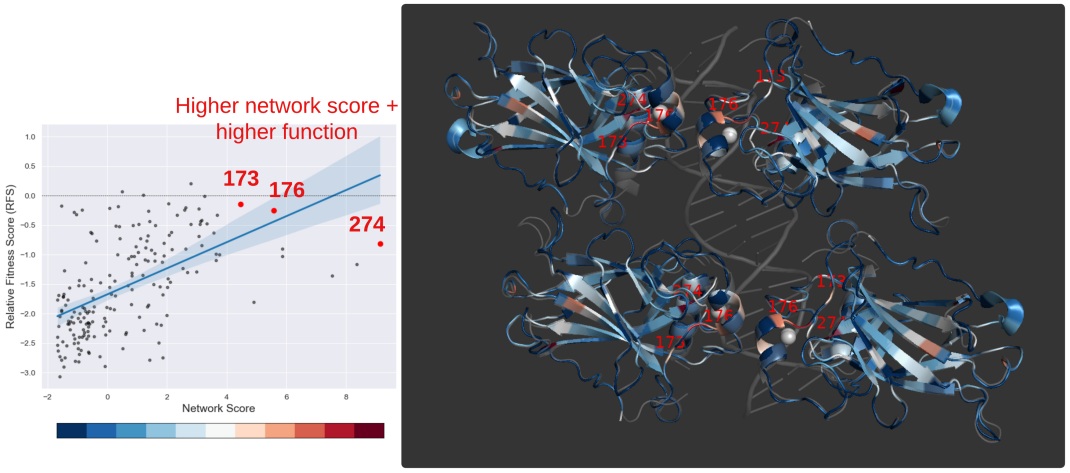

FIGURE 5.8: Network scores of wild-type p53 protein without DNA (2OCJ) (displayed as opaque) aligned onto wild-type p53 protein with DNA (4HJE) (displayed as transparent). p53 works as a tumour-suppressor gene by binding to DNA and so cancer-causing mutations disrupt its ability to bind DNA. Residues that are highly networked and functional (173, 176, 274) were located near the DNA contact region.

program) when DNA damage is detected. It therefore controls cell growth, can trigger cell death, and is crucial to DNA damage repair. Structurally, p53 binds to DNA as a tetramer (4 identical subunits) through its DNA-binding domain, recognising a specific sequence [61]. p53 is one of the most commonly mutated proteins in cancers, and the majority of these mutations prevent p53 from binding to DNA and activating genes that suppress tumour formation [20].

We calculated the NS for the wild-type p53 in the absence of DNA (PDB ID: 2OCJ) and aligned each chain (subunit) to the wild-type p53 with DNA (PDB ID: 4HJE) (Figure 5.8). This allowed us to observe how network scores arrange themselves concerning the location of DNA in the DNA-binding domain. Interestingly, the residues with the highest NS and functional importance (residues 173, 176, 274), as identified from our previous saturation mutagenesis analysis (Section 5.2), were located in close contact with the DNA strand. This observation has significant implications for understanding how structurally important residues play a role in cancer progression. If those residues or residues nearby are mutated, it may disrupt the tetramers ability to bind DNA and regulate gene expression, leading to loss of its tumour-suppressing functions, thereby also validating our conclusion from Section 5.3.

It is also worth noting that although residues located at positions 173, 176, and 274 are not frequently mutated themselves, residues at positions 175 and 273, located nearby, are mutated in high volumes [77].

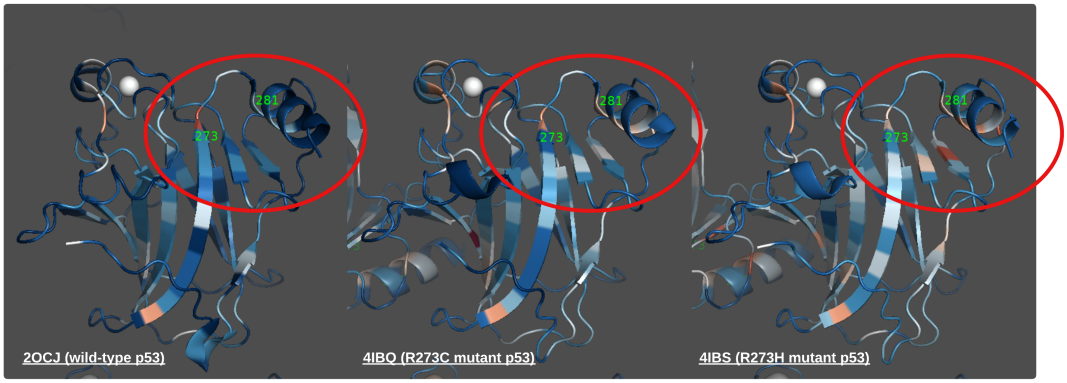

FIGURE 5.9: Comparison of network scores across the wild-type p53, and the R273C and R273H mutants. We highlight the region with significant changes in network scores. Molecular dynamics studies have shown a disruption in the interaction and increased distances between R273 and D281 for R273C and R273H mutations [39].

5.5.2 Network score differences unveil dynamic changes missed by crystallography

We examined structures for the R273C and R273H mutants and calculated NS for the residues (Figure 5.9). The R273C mutation represents a switch from arginine (R), a positively charged and hydrophilic amino acid, to cysteine (C), a special case amino acid, at position 273. Likewise, the R273H mutation replaces arginine with histidine (H), also a positively charged and hydrophilic amino acid. Arginine at position 273 is mutated to cysteine and histidine in 39.1% and 47.5% of all cancers [15] and the mutations have been described as "contact mutations" [32, 99], disrupting DNA binding without affecting the overall protein structure, unlike "structural mutations" which impact protein stability [19, 101]. Indeed, crystallographic examinations have shown that the R273C and R273H mutations do not disturb the overall structure of the protein or neighbouring residues [49, 50].

However, molecular dynamics simulations reveal that these mutations alter the dynamics and stability of the protein-DNA complex, weakening DNA binding by increasing distances between the DNA strand and the protein, and disrupting the salt-bridge interaction between R273 and D281 residues [39]. This indicates that while crystallographic structures appear unchanged, the mutations significantly impact the protein's dynamic behaviour and interaction network, affecting its function.

We therefore plotted the NS for the wild type p53, the R273C (PDB ID: 4IBQ) and R273H mutants (PDB ID: 4IBS), coloured by network scores (Figure 5.9). Interestingly, although the structures appear similar, NS analysis reveals significant changes near

D281, highlighting increased network connectivity. Importantly, this highlights a powerful aspect of visualising NS differences: demonstrating the influence of not only visible 3D structural changes but also energetic forces. In fact, the network creation process in SBNA involves accounting for these energetic forces, allowing for deeper interpretation and insight beyond simple crystallographic analysis. The novel NS visualisation tool developed here may be a promising tool to rapidly capture these complexities effectively, driving intuition on understanding how the structural topology of cancer drivers affects protein structure and function, and explaining their role in cancer evolution.

Chapter 6

Conclusion

Our study leveraged network scores using structure-based network analysis to investigate the functional impacts of mutations in three prominent cancer driver genes and the patterns of mutations in lung cancer genes. Through systematic analysis and novel visualisation techniques, we have revealed several key findings:

1. **Structural topology is linked with protein function:** Mutations in highly networked residues, as observed in BRCA1, PTEN, and TP53 genes through saturation mutagenesis data, lead to significant functional impairments. Higher NS correlates with greater functional importance of that residue, highlighting the crucial role of central residues in maintaining protein function.

2. **Structural topology is linked with cancer driver evolution**: Our analysis reveals distinct mutation patterns in TSGs and OCGs in humans. TSGs tend to mutate in highly networked residues, resulting in loss of function and promoting cancer progression whereas OCGs preferentially mutate in poorly networked residues to enhance or maintain function, driving oncogenesis. This finding provides a novel perspective on the evolutionary pressures that shape cancer development by understanding structural constraints in proteins.

3. **A novel visualisation technique developed here may help interpretation:** The development of a visualisation tool using PyMOL enabled an intuitive understanding of the relationship between NS and protein structure and function. By mapping NS onto protein structures, we can identify regions in the protein with high functional importance. By observing changes in network scores between wild-types and mutants, we also gained insight into the structural, energetic and functional implications of mutations.

6.1 Contributions and future research directions

Our research presents significant contributions to the understanding of structural topology through NS and its implications in cancer biology. Expanding the analysis to include all types of cancers is the evident next step to allow our conclusion to be generalised. Our findings also have practical implications in targeted vaccine design - prioritising highly networked residues in oncogenes may increase their efficacy. For instance, inducing an immune response against these residues could steer cancer evolution away from the most functionally damaging mutations, similar to strategies used in HIV and SARS CoV-2 vaccines [48]. Furthermore, our research may provide a framework for understanding novel cancer driver genes - by analysing their mutation patterns it may help understand whether they are tumour-suppressing, oncogenic, or neither. The visualisation script we developed demonstrates the power of using NS to understand cancer dynamics. Its applications extend beyond cancer research. The script can be adapted to map different metrics and visualise side-by-side comparisons of protein structures automatically.

6.2 Limitations and challenges

The accuracy of NS calculations hinges on the quality and methodology of crystallographic studies from which protein structures were obtained. Variations in experimental conditions, such as the presence of ligands, may significantly influence protein structure and NS comparisons across different structures which may affect the analysis done in this paper. Therefore, it is crucial to acknowledge these contextual factors when interpreting NS data. Moreover, NS may be limited by the inherent limitations of crystallographic studies, which present proteins as static, instead of dynamic entities. New methods that predict structure [3] could be a promising way to overcome limitations in formal crystallographic methods, but will need further evaluation.

6.3 Final statements

I acknowledge the use of ChatGPT [69] to summarise texts and academic papers, and to refine and improve the conciseness of my writing. Any new information was factually checked.

We also address ethical concerns. The primary personal data sources for this study are COSMIC and cBioPortal. COSMIC gathers data from diverse origins, including scientific literature, public databases, and direct submissions. Similarly, cBioPortal

collects genomic and clinical data from numerous cancer patients spanning various cancer types and demographics. We emphasise that the data obtained from these sources have already been de-identified. This research also does not generate new data on the patients themselves.

The word count only includes the abstract, introduction, background, methodology, results and conclusion sections.

Appendix A

Genes and Structures used

We present the list of lung cancer genes obtained from CGC (COSMIC) and PDB structures used in our second set of analyses after curation to understand how network scores constrain lung cancers.

#	Gene	Wild Type PDBs	Mutant PDBs	Total PDBs	Cancer Role
1	ALK	3AOX, 8ARJ, 4CMT, 4CTB, 7R7R, 5IMX, 6CDT	4FNX, 4FNW, 5IUI, 5IUG, 4FNY, 4FNZ, 2YJR	14	oncogene
2	KRAS	6GOD, 6MBQ	4LDJ, 4L8G, 8AFB, 8AZX, 7C40, 6GOF, 8TXG, 4EPR, 6GJ7, 6GOE, 8AZZ	13	oncogene
3	EPHA3	2QOO, 2QO2, 2QOQ, 2QO7, 3FXX, 2GSF, 2QO9	2QOD, 2QOF, 2QOI, 2QOL, 2QOK	12	unknown
4	BRAF	3II5, 3Q4C, 3PRF, 3PPK, 3D4Q, 5FD2, 4.00E+26	5C9C, 3IDP, 4MNF, 4G9R	11	oncogene
5	KEAP1	8EJR, 7K2F, 7K2A, 7K2K, 7K2E, 7K2D	5WFL, 5WHL, 7K29, 5WFV, 5WG1	11	TSG
6	TP53	2OCJ	4LOE, 4LO9, 2J1Y, 4IBS, 4IBQ, 4IBZ, 7V97, 6FF9, 7DHY	10	TSG
7	EED	7SI5, 7SI4, 3JPX, 3JZG, 3K26, 3K27, 5U5H, 5H13, 5U5K		9	TSG
8	EGFR	7SI1, 7UKV, 3POZ, 8F1X	2ITZ, 6JWL, 5X26, 5X27, 5X28	9	oncogene
9	MAP2K1	4LMN, 4U7Z, 3E8N, 3DY7, 3V04, 4U81, 4U80, 3PP1		8	oncogene
10	KDR	1Y6A, 1Y6B, 3VHK, 3VHE, 3VID		5	oncogene
11	NTRK2	4ASZ, 4AT5, 4AT3, 4AT4, 1WWB		5	oncogene
12	AKT1	2UVM, 1UNR, 1UNQ	2UZR, 2UZS	5	oncogene
13	FGFR2	6LVK, 6LVL	4J96, 4J98	4	oncogene
14	KIF5B	1MKJ, 1BG2	5LT0	3	fusion
15	BIRC6	8E2H, 8E2F, 8E2G		3	oncogene
16	ROS1	7Z5X, 7Z5W, 3ZBF		3	oncogene
17	HIP1	3I00, 2NO2		2	oncogene
18	RB1	3POM	4ELL	2	TSG
19	TPR	5TO5, 5TVB		2	fusion
20	SUB1	2C62	4USG	2	unknown
21	SMARCA4		2GRC, 3UVD	2	TSG
22	ERBB4	3U2P, 3BCE		2	oncogene
23	EZR	4RMA, 1NI2		2	fusion
24	NOTCH1	3ETO		1	oncogene
25	ERBB2	2A91		1	oncogene
26	HGF	3HMS		1	oncogene
27	PTPRD	6X3A		1	TSG
28	N4BP2	3FAU		1	TSG
29	RFWD3	6CVZ		1	TSG
30	MB21D2	7LT1		1	unknown
31	SIRPA	2WNG		1	TSG
32	MAP2K2	1S9I		1	oncogene
33	EML4		4CGC	1	fusion
34	PTPN13	1WCH		1	TSG

FIGURE A.1: Lung cancer genes and the respective PDB structures used for the second set of analyses.

Appendix B

SBNA to UniProt alignment

The main function to align from SBNA to UniProt is `align_final_sum_with_uniprot`

```
def map_sequ_sbna_pdb(sbna, pdb):
    # maps the sbna sequence to the pdb sequence
    sbna_to_pdb_mapping = []
    # i starts at where the first letter is found (not -) in pdb_aligned
    i = 0
    for (s, p) in zip(sbna, pdb):
        if s == p:
            sbna_to_pdb_mapping.append(i)
            i += 1
        elif s == '-':
            i += 1
        else:
            sbna_to_pdb_mapping.append("?")
    assert(len(sbna_to_pdb_mapping) == len("".join(aa for aa in sbna if
    aa != '-')))  # verify that the length is correct
    return sbna_to_pdb_mapping


def convert_auth_to_pdb(pdb_id):
    # get the author chain id to pdb chain id mapping
    response = requests.get(f'''
    https://data.rcsb.org/graphql?query={{entry(entry_id:"{pdb_id}")
    {{polymer_entities {{
            polymer_entity_instances {{
                rcsb_polymer_entity_instance_container_identifiers {{
                auth_asym_id
                asym_id
                entry_id
                entity_id
                }}
            }}
```

```
            }}
        }}
    }}
    ''')
    # check response
    if response.status_code != 200:
        print("Failed to fetch data from RCSB: convert auth_asym_id to asym_id")
        return None
    response = response.json()

    auth_pdb_map = {}
    for polymer in response['data']['entry']['polymer_entities']:
        # represent one "macromolecule" section in PDB website
        for chain_instance in polymer['polymer_entity_instances']:
            # represents one chain
            # map author chain id to rcsb pdb chain id
            auth_pdb_map[chain_instance['rcsb_polymer_entity_instance_container_identifiers']['auth_asym_id']] = \
                chain_instance['rcsb_polymer_entity_instance_container_identifiers']['asym_id']
    return auth_pdb_map

def get_alignment_regions(uniprot_id, pdb_id, chain, auth_pdb_map):
    """ returns the aligned sequences like e.g.
    [{'query_begin': 696, 'query_end': 772, 'target_begin': 2, 'target_end': 78}, {'query_begin': 773, 'query_end': 1022, 'target_begin': 82, 'target_end': 331}]
    and the PDB sequence"""

    def tmp(unp):
        return requests.get(f"""https://1d-coordinates.rcsb.org/graphql?query={{
        alignment(
            from:UNIPROT,
            to:PDB_INSTANCE,
            queryId:"{unp.split('-')[0]}",
        ){{
            query_sequence
            target_alignment {{
            target_id
            target_sequence
            coverage{{
                query_coverage
                query_length
                target_coverage
                target_length
```

```
            }}
            aligned_regions {{
                query_begin
                query_end
                target_begin
                target_end
            }}
            }}
        }}
        }}""")

    # maps UniProt - PDBs (one to many)
    response = tmp(uniprot_id)
    if response.status_code != 200 or response.json()['data']['alignment'
    ]['target_alignment'] is None:
        # print("Failed to fetch data from RCSB: get alignment regions")

        # try to get uniprot id another way
        response_tmp = requests.get(f"""https://www.ebi.ac.uk/pdbe/api/
    mappings/uniprot/{pdb_id}""")
        response_tmp = response_tmp.json()
        found = None
        for uniprot_id, data in response_tmp[pdb_id.lower()]['UniProt'].
    items():
            for mapping in data["mappings"]:
                if mapping["struct_asym_id"] == chain:
                    response = tmp(uniprot_id)
                    found = True
                    break
            if found:
                break

    response = response.json()
    for alignment in response['data']['alignment']['target_alignment']:
        # search for correct pdb and chain
        if alignment['target_id'] == f"{pdb_id}.{auth_pdb_map[chain]}":
            target_sequence = alignment['target_sequence']
            aligned_regions = alignment['aligned_regions']
            return aligned_regions, target_sequence
    return None, None


def map_sequ_sbna_pdb(sbna, pdb):
    # maps the sbna sequence to the pdb sequence
    sbna_to_pdb_mapping = []
    # i starts at where the first letter is found (not -) in pdb_aligned
    i = 1
    for (s, p) in zip(sbna, pdb):
```

```
117         if s == p:
118             sbna_to_pdb_mapping.append(i)
119             i += 1
120         elif s == '-':
121             i += 1
122         else:
123             sbna_to_pdb_mapping.append("?")
124 
125     assert(len(sbna_to_pdb_mapping) == len("".join(aa for aa in sbna if aa != '-')))  # verify that the length is correct
126     return sbna_to_pdb_mapping
127 
128 def align_finalsum_with_uniprot(final_sum, pdb_id, chain):
129     """
130     final_sum is the output from SBNA, where each row is a residue in the chain (although not always correct)
131     pdb_id is the pdb id
132     chain is the chain id
133 
134     This function aligns the SBNA sequence with the PDB sequence, and then maps the PDB sequence to the UniProt sequence
135     """
136     data = get_dbref_data(pdb_id) # obtain the DBREF lines details from the pdb file
137     ranges = [range(i['start'], i['end']+1) for i in data if i['chain'] == chain] # get the start and end sequence number in the chain
138     r = [item for sublist in ranges for item in sublist] # flatten the list. r is the list of all sequence numbers in the chain
139     final_sum = final_sum[final_sum['num'].isin(r)].reset_index(drop=True)
140 
141     # get uniprot sequence (use a cache to avoid multiple requests for the same uniprot id)
142     uniprot_id = get_uniprot_id(pdb_id, chain)
143     with open("../cache/uniprot_seqs.json", "r") as f:
144         uniprot_seqs = json.load(f) # saved uniprot sequences
145     if uniprot_id in uniprot_seqs:
146         uniprot_seq = uniprot_seqs[uniprot_id]
147     else:
148         uniprot_seq = get_uniprot_sequence(uniprot_id)
149         uniprot_seqs[uniprot_id] = uniprot_seq
150         with open("../cache/uniprot_seqs.json", "w") as f:
151             json.dump(uniprot_seqs, f) # save uniprot sequences
152 
153     # convert author chain id to pdb chain id
154     auth_pdb_map = convert_auth_to_pdb(pdb_id) # convert author chain id to pdb chain id
```

```
    aligned_regions, target_sequence = get_alignment_regions(uniprot_id, pdb_id, chain, auth_pdb_map)

    # sometimes SBNA has multiple residues for the same number (due to other chains/polymer entities)
    # where there are multiple residues for the same number, make final_sum use the n'th value
    # method is not perfect, but works for most cases
    if any(final_sum.groupby('num')['res_code'].count() > 1):
        print(f"Multiple residues for the same number detected for {pdb_id}{chain}")
        try:
            # non-duplicated rows
            non_dups = final_sum.drop_duplicates(subset='num', keep=False)

            # see how many chains there are for each number
            dbrefs = get_dbref_data(pdb_id)

            # duplicated rows
            dups = final_sum[final_sum.duplicated('num', keep=False)].reset_index(drop=True)

            # unique chains in dbrefs (dbrefs is a dict)
            unique_chains = []
            for chain_entity in dbrefs:
                if chain_entity['chain'] not in unique_chains:
                    unique_chains.append(chain_entity['chain'])
            i = unique_chains.index(chain)

            # get the i'th value for duplicated rows
            # Group by 'num' and create a counter for each group
            dups['counter'] = dups.groupby('num').cumcount()
            # Filter the DataFrame to get the ith occurrence (e.g., n=3)
            nth_items = dups[dups['counter'] == i]
            nth_items = nth_items.drop(columns=['counter'])

            # combine non-duplicated and duplicated rows
            final_sum = pd.concat([non_dups, nth_items])

            # sort by num
            final_sum = final_sum.sort_values(by='num')
        except:
            print("Error in handling multiple residues for the same number, using all values.")
            pass

    sbna_seq = ''.join(final_sum['res_code'].values)
```

```
196
197     aligner = Align.PairwiseAligner()
198     aligner.mode = 'global'
199     alignments = aligner.align(sbna_seq, target_sequence)
200
201     sbna_aligned, pdb_aligned  = alignments[0][0], alignments[0][1]
202
203     mapped = map_sequ_sbna_pdb(sbna_aligned, pdb_aligned)
204
205     # align pdb sequence to uniprot seqeunce using aligned regions
206     pdb_to_uniprot_mapping = {}
207     for region in aligned_regions:
208         pdb_begin = region['target_begin']
209         pdb_end = region['target_end']
210         uniprot_begin = region['query_begin']
211         uniprot_end = region['query_end']
212
213         # map numbers from pdb to uniprot
214         for i in range(pdb_begin, pdb_end+1):
215             pdb_to_uniprot_mapping[i] = i - pdb_begin + uniprot_begin
216
217     # map sbna sequence to uniprot sequence
218     sbna_to_uniprot_mapping = []
219     for i in mapped:
220         if i != "?":
221             try:
222                 sbna_to_uniprot_mapping.append(pdb_to_uniprot_mapping[i])
223             except:
224                 sbna_to_uniprot_mapping.append("?")
225         else:
226             sbna_to_uniprot_mapping.append("?")
227     final_sum['uniprot_num'] = sbna_to_uniprot_mapping
228     final_sum['uniprot_res'] = [uniprot_seq[i-1] if i != "?" else "?" for
     i in sbna_to_uniprot_mapping]
229
230     return final_sum
```

Listing B.1: Code to align sequence from SBNA to PDB

Appendix C

Network scores visualisation script

We first adapted the `pdb_color_generic.py` script from the `pdbcolor` GitHub Project [34]. In part C.1, it is named `pdb_color_generic_v3.py` and the main changes are changing the palette (uniform), simplifying the script, and changing the number of bins. We provide the code to generate a script for the network representation in PyMOL. In part C.2, it is named `pdb_color_generic_v4_comparison.py`, and the main change includes accepting multiple PDB structures, instead of just one, extending the range of colours considering all structures. Similarly, we provide the code to generate a script for side-by-side comparison of multiple structures coloured by the network scores.

C.1 Create network representation

```
1
2  def get_ca_coordinates(pdb_file):
3      # get x, y, z coordinates of alpha carbon atoms
4      parser = PDBParser()
5      structure = parser.get_structure("pdb_structure", pdb_file)
6      ca_atoms = []
7      for model in structure:
8          for chain in model:
9              for residue in chain:
10                 if residue.get_id()[0] == ' ':
11                     for atom in residue:
12                         if atom.get_id() == 'CA':
13                             coords = atom.get_coord()
14                             ca_atoms.append({
15                                 'residue_number': residue.id[1],
```

```
                                    'amino_acid': residue.resname,
                                    'chain': chain.id,
                                    'x': coords[0],
                                    'y': coords[1],
                                    'z': coords[2],
                                })
    return ca_atoms

def create_pymol_script(score_file_path, pdb_id, score_name, palette="RdBu", chain_col=0, site_col=1, score_col=2, reverse_color=True, meaningful_zero=True):
    """
    Create a PyMOL script to visualize the protein structure with the calculated scores.
    """

    pdb_file_path = f"../pdb_files/{pdb_id}.pdb"
    ca_coordinates = pd.DataFrame(get_ca_coordinates(pdb_file_path))
    ca_coordinates['foo_name'] = ["foo"+str(i) for i in range(1, len(ca_coordinates) + 1)]
    ca_coordinates['atom_name'] = ["atom"+str(i) for i in range(1, len(ca_coordinates) + 1)]

    out = "bg_color grey30\n"
    for _, row in ca_coordinates.iterrows():
        out += f"pseudoatom {row['foo_name']}, pos=[{row['x']}, {row['y']}, {row['z']}], name={row['atom_name']}\n"
        out += f"set grid_slot, 1, {row['foo_name']}\n"
        out += f"set grid_slot, 1, {row['atom_name']}\n"

    with open(f"../sbna_results/{pdb_id}/A/{pdb_id}_multimer/Centroid/{pdb_id}_multimer_nowaters_centroidNetSC", "r") as f:
        res1 = []
        res2 = []
        dist = []
        for line in f:
            res1.append(line.split()[0])
            res2.append(line.split()[1])
            dist.append(line.split()[-1])
    tmp = pd.DataFrame()
    tmp['res1'] = res1
    tmp['res2'] = res2
    tmp['dist'] = dist

    # --> there are duplicates where res1 and res2 are just swapped

    # Create a new column with sorted node pairs
```

```
    tmp['sorted_nodes'] = tmp.apply(lambda row: sorted([row['res1'], row['res2']]), axis=1)
    tmp.drop_duplicates(subset='sorted_nodes', inplace=True) # Drop duplicates based on sorted node pairs
    tmp = tmp.drop(columns=['sorted_nodes']).reset_index(drop=True)

    tmp['aa1'], tmp['res_num1'], tmp['chain1'] = get_res_info(tmp['res1'])
    tmp['aa2'], tmp['res_num2'], tmp['chain2'] = get_res_info(tmp['res2'])

    tmp = tmp.merge(ca_coordinates, left_on=['res_num1', 'aa1', 'chain1'], right_on=['residue_number', 'amino_acid', 'chain'], how='left')
    tmp.rename(columns={'x': 'x1', 'y': 'y1', 'z': 'z1', 'foo_name': 'foo_name1', 'atom_name': 'atom_name1'}, inplace=True)

    tmp = tmp.merge(ca_coordinates, left_on=['res_num2', 'aa2', 'chain2'], right_on=['residue_number', 'amino_acid', 'chain'], how='left')
    tmp.rename(columns={'x': 'x2', 'y': 'y2', 'z': 'z2', 'foo_name': 'foo_name2', 'atom_name': 'atom_name2'}, inplace=True)

    count = 1

    for _, row in tmp.iterrows():
        bond = "b"+str(count)
        out += f"create {bond}, {row['foo_name1']} or {row['foo_name2']}\n"
        out += f"bond {bond}////{row['atom_name1']}, {bond}////{row['atom_name2']}\n"
        out += f"set grid_slot, 1, {bond}\n"
        count += 1

    out += "remove resn HOH\n"
    out += 'delete '
    for i in ca_coordinates['foo_name']:
        out += i + ' or '
    out = out[:-3]
    out += '\n'

    out += "set grid_mode, 1\n"
    out += f"fetch {pdb_id}\n"
    out += f"set grid_slot, 2, {pdb_id}\n"

    out += f"copy {pdb_id}_copy, {pdb_id}\n"
    out += f"set grid_slot, 1, {pdb_id}_copy\n"

    out += f"hide everything, {pdb_id}_copy\n"
```

```
93      out += f"show spheres, name ca and {pdb_id}_copy\nset sphere_scale,
        0.25, (all)\n"
94
95      out += f"show cartoon, {pdb_id}_copy\n"
96      out += f"set cartoon_transparency, 0.85, {pdb_id}_copy\n"
97
98      cmd = f"python pdb_color_generic_v3.py -c {str(score_col)} -d , -i {
        score_file_path} -l {palette} --site-column {str(site_col)} "
99      if chain_col:
100         cmd += f"--chain-column {str(chain_col)} "
101     if not meaningful_zero:
102         cmd += "-z "
103     if reverse_color:
104         cmd += "-r "
105
106     script_args = cmd.strip().split(' ')
107     out += subprocess.run(script_args, capture_output=True, text=True).
        stdout
108
109     # recolour bonds, since they are overwritten by the pdb_color_generic
         script
110     count = 1
111     for _, row in tmp.iterrows():
112         bond = "b"+str(count)
113         out += f"color grey50, {bond}\n"
114         count += 1
115
116     return out
117
118 # EXAMPLE USE:
119 pdb_id = "2OCJ"
120 score_name = "network_score"
121
122 ns_data = pd.read_csv("../lung_cancer/lung_genes_sbna.csv")
123 ns_data = ns_data[ns_data['pdb_id']==pdb_id]
124 score_filepath = f"data/{pdb_id}_{score_name}.csv"
125 ns_data = ns_data[ns_data['uniprot_num']!="?"]
126 ns_data['uniprot_num'] = ns_data['uniprot_num'].astype(int)
127 ns_data = ns_data[['chain', 'uniprot_num', score_name]].sort_values(['
    chain', 'uniprot_num'])
128 ns_data.to_csv(score_filepath, index=False) # save a temporary score file
129 script = create_pymol_script(score_filepath, pdb_id, score_name,
    reverse_color=True, meaningful_zero=False)
130
131 # write script to file
132 with open(f"scripts/{pdb_id}_{score_name}.pml", "w") as f:
133     f.write(script)
```

Listing C.1: Code to generate a script for the network representation in PyMOL

C.2 Compare multiple structures

```
# create a dynamic script that takes in multiple pdbs
score_name = "network_score"

pdbs = ["2OCJ", "4IBQ", "4IBS"]
aligned_chain = ["A","C","C"] # how chains are aligned

score_filepath = f"data/{score_name}.csv"
ns_data = pd.read_csv("../lung_cancer/lung_genes_sbna.csv")
ns_data = ns_data[ns_data['pdb_id'].isin(pdbs)]
ns_data = ns_data[ns_data['uniprot_num']!="?"]
ns_data['uniprot_num'] = ns_data['uniprot_num'].astype(int)
ns_data = ns_data[['pdb_id', 'chain', 'uniprot_num', score_name]].
    sort_values(['pdb_id', 'chain', 'uniprot_num'])
ns_data.to_csv(score_filepath, index=False) # save a temporary score file

out = f"""bg_color grey30
set grid_mode, 1
"""

for pdb in pdbs:
    out += f"fetch {pdb}\n"

for pdb, chain in zip(pdbs, aligned_chain):
    out += f"align (chain {chain} & {pdb}), (chain {aligned_chain[0]} & {
    pdbs[0]})\n"
out += f"zoom (chain {chain} & {pdb})\n"

out += "remove resn HOH\n"

script_args = f"python pdb_color_generic_v4_comparison.py -c 3 -d , -i {
    score_filepath} -l RdBu -r -z --pdb-column 0 --chain-column 1 --site-
    column 2".split(' ')
out += subprocess.run(script_args, capture_output=True, text=True).stdout

# write script to file
with open(f"scripts/{"_".join(pdbs)}_{score_name}.pml", "w") as f:
    f.write(out)
```

Listing C.2: Code to generate the side by side visualisations for multiple structures coloured by network score.

Reference List

[1] "Amino Acid". *Wikipedia*. May 2019. URL: https://en.wikipedia.org/wiki/Amino_acid (visited on 06/13/2024).

[2] "Locant". *Wikipedia*. June 2021. URL: https://en.wikipedia.org/wiki/Locant (visited on 06/13/2024).

[3] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630 (8016 June 2024), pp. 493–500. ISSN: 0028-0836. DOI: 10.1038/s41586-024-07487-w.

[4] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. "The Shape and Structure of Proteins". In: 4th. Garland Science, 2002. URL: https://www.ncbi.nlm.nih.gov/books/NBK26830/.

[5] Ludmil B Alexandrov and Michael R Stratton. "Mutational signatures: the patterns of somatic mutations hidden in cancer genomes". In: *Current Opinion in Genetics Development* 24 (Feb. 2014), pp. 52–60. ISSN: 0959437X. DOI: 10.1016/j.gde.2013.11.014.

[6] Emil Alexov and Michael Sternberg. "Understanding Molecular Effects of Naturally Occurring Genetic Differences". In: *Journal of Molecular Biology* 425 (21 Nov. 2013), pp. 3911–3913. ISSN: 00222836. DOI: 10.1016/j.jmb.2013.08.013.

[7] Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanely, Ilya Venger, and Shmuel Pietrokovski. "Network Analysis of Protein Structures Identifies Functional Residues". In: *Journal of Molecular Biology* 344 (4 Dec. 2004), pp. 1135–1146. ISSN: 00222836. DOI: 10.1016/j.jmb.2004.10.055.

[8] Christian Arlt, Christian H. Ihling, and Andrea Sinz. "Structure of full-length p53 tumor suppressor probed by chemical cross-linking and mass spectrometry". In: *PROTEOMICS* 15 (16 Aug. 2015), pp. 2746–2755. ISSN: 1615-9853. DOI: 10.1002/pmic.201400549.

[9] Ali Rana Atilgan, Deniz Turgut, and Canan Atilgan. "Screened Nonbonded Interactions in Native Proteins Manipulate Optimal Paths for Robust Residue Communication". In: *Biophysical Journal* 92 (9 May 2007), pp. 3052–3062. ISSN: 00063495. DOI: 10.1529/biophysj.106.099440.

[10] Australian Institute of Health and Welfare. *Deaths in Australia*. Australian Institute of Health and Welfare, July 2023. URL: https://www.aihw.gov.au/reports/life-expectancy-deaths/deaths-in-australia/contents/leading-causes-of-death (visited on 06/09/2024).

[11] Ganesh Bagler and Somdatta Sinha. "Assortative mixing in Protein Contact Networks and protein folding kinetics". In: *Bioinformatics* 23 (14 July 2007), pp. 1760–1767. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm257.

[12] Sailen Barik. "The Uniqueness of Tryptophan in Biology: Properties, Metabolism, Interactions and Localization in Proteins". In: *International Journal of Molecular Sciences* 21 (22 Nov. 2020), p. 8776. ISSN: 1422-0067. DOI: 10.3390/ijms21228776.

[13] Niko Beerenwinkel, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz. "Cancer Evolution: Mathematical Models and Computational Inference". In: *Systematic Biology* 64 (1 Jan. 2015), e1–e25. ISSN: 1076-836X. DOI: 10.1093/sysbio/syu081.

[14] H. M. Berman. "The Protein Data Bank". In: *Nucleic Acids Research* 28 (1 Jan. 2000), pp. 235–242. ISSN: 13624962. DOI: 10.1093/nar/28.1.235.

[15] Liacine Bouaoun, Dmitriy Sonkin, Maude Ardin, Monica Hollstein, Graham Byrnes, Jiri Zavadil, and Magali Olivier. "¡i¿TP53¡/i¿ Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data". In: *Human Mutation* 37 (9 Sept. 2016), pp. 865–876. ISSN: 10597794. DOI: 10.1002/humu.23035.

[16] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W. Kinzler, Bert Vogelstein, and Martin A. Nowak. "Accumulation of driver and passenger mutations during tumor progression". In: *Proceedings of the National Academy of Sciences* 107 (43 Oct. 2010), pp. 18545–18550. ISSN: 0027-8424. DOI: 10.1073/pnas.1010978107.

[17] Ino de Bruijn, Ritika Kundra, Brooke Mastrogiacomo, Thinh Ngoc Tran, Luke Sikina, Tali Mazor, Xiang Li, Angelica Ochoa, Gaofei Zhao, Bryan Lai, Adam Abeshouse, Diana Baiceanu, Ersin Ciftci, Ugur Dogrusoz, Andrew Dufilie, Ziya Erkoc, Elena Garcia Lara, Zhaoyuan Fu, Benjamin Gross, Charles Haynes, Allison Heath, David Higgins, Prasanna Jagannathan, Karthik Kalletla, Priti Kumari, James Lindsay, Aaron Lisman, Bas Leenknegt, Pieter Lukasse, Divya Madela, Ramyasree Madupuri, Pim van Nierop, Oleguer Plantalech, Joyce Quach, Adam C. Resnick, Sander Y.A. Rodenburg, Baby A. Satravada, Fedde Schaeffer, Robert Sheridan, Jessica Singh, Rajat Sirohi, Selcuk Onur Sumer, Sjoerd van Hagen, Avery Wang, Manda Wilson, Hongxin Zhang, Kelsey Zhu, Nicole Rusk, Samantha Brown, Jessica A. Lavery, Katherine S. Panageas, Julia E. Rudolph, Michele L. LeNoue-Newton, Jeremy L. Warner, Xindi Guo, Haley Hunter-Zinck, Thomas V. Yu, Shirin Pilai, Chelsea Nichols, Stuart M. Gardos, John Philip, Kenneth L. Kehl, Gregory J. Riely, Deborah Schrag, Jocelyn Lee, Michael V. Fiandalo, Shawn M. Sweeney, Trevor J. Pugh, Chris Sander, Ethan Cerami, Jianjiong Gao, and Nikolaus Schultz. "Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal". In: *Cancer Research* 83 (23 Dec. 2023), pp. 3861–3867. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-23-0816.

[18] Guillaume Brysbaert and Marc F. Lensink. "Centrality Measures in Residue Interaction Networks to Highlight Amino Acids in Protein–Protein Binding". In: *Frontiers in Bioinformatics* 1 (June 2021). ISSN: 2673-7647. DOI: 10.3389/fbinf.2021.684970.

[19] Alex N Bullock, Julia Henckel, and Alan R Fersht. "Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy". In: *Oncogene* 19 (10 Mar. 2000), pp. 1245–1256. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1203434.

[20] Alex N. Bullock and Alan R. Fersht. "Rescuing the function of mutant p53". In: *Nature Reviews Cancer* 1 (1 Oct. 2001), pp. 68–76. ISSN: 1474-175X. DOI: 10.1038/35094077.

[21] Cancer Australia. *Cancer in Australia Statistics*. Cancer Australia, 2022. URL: https://www.canceraustralia.gov.au/impacted-cancer/what-cancer/cancer-australia-statistics (visited on 06/09/2024).

[22] Matias Casás-Selves and James DeGregori. "How Cancer Shapes Evolution and How Evolution Shapes Cancer". In: *Evolution: Education and Outreach* 4 (4 Dec. 2011), pp. 624–634. ISSN: 1936-6426. DOI: 10.1007/s12052-011-0373-y.

[23] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data". In: *Cancer Discovery* 2 (5 May 2012), pp. 401–404. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-12-0095.

[24] Broto Chakrabarty, Varun Naganathan, Kanak Garg, Yash Agarwal, and Nita Parekh. "NAPS update: network analysis of molecular dynamics data and protein–nucleic acid complexes". In: *Nucleic Acids Research* 47 (W1 July 2019), W462–W470. ISSN: 0305-1048. DOI: 10.1093/nar/gkz399.

[25] Broto Chakrabarty and Nita Parekh. "NAPS: Network Analysis of Protein Structures". In: *Nucleic Acids Research* 44 (W1 July 2016), W375–W382. ISSN: 0305-1048. DOI: 10.1093/nar/gkw383.

[26] Broto Chakrabarty and Nita Parekh. *NAPS: Network Analysis of Protein Structures*. URL: https://bioinf.iiit.ac.in/NAPS/ (visited on 06/13/2024).

[27] Shuo Chen, Jia-Le Wu, Ying Liang, Yi-Gang Tang, Hua-Xin Song, Li-Li Wu, Yang-Fei Xing, Ni Yan, Yun-Tong Li, Zheng-Yuan Wang, Shu-Jun Xiao, Xin Lu, Sai-Juan Chen, and Min Lu. "Arsenic Trioxide Rescues Structural p53 Mutations through a Cryptic Allosteric Site". In: *Cancer Cell* 39 (2 Feb. 2021), 225–239.e8. ISSN: 15356108. DOI: 10.1016/j.ccell.2020.11.013.

[28] Yunje Cho, Svetlana Gorina, Philip D. Jeffrey, and Nikola P. Pavletich. "Crystal Structure of a p53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations". In: *Science* 265 (5170 July 1994), pp. 346–355. ISSN: 0036-8075. DOI: 10.1126/science.8023157.

[29] Damiano Clementel, Alessio Del Conte, Alexander Miguel Monzon, Giorgia F Camagni, Giovanni Minervini, Damiano Piovesan, and Silvio C E Tosatto. "RING 3.0: fast generation of probabilistic residue interaction networks from structural ensembles". In: *Nucleic Acids Research* 50 (W1 July 2022), W651–W656. ISSN: 0305-1048. DOI: 10.1093/nar/gkac365.

[30] Paul A. Craig, Lea Vacca Michel, and Robert C. Bateman. "A survey of educational uses of molecular visualization freeware". In: *Biochemistry and Molecular Biology Education* 41 (3 May 2013), pp. 193–205. ISSN: 1470-8175. DOI: 10.1002/bmb.20693.

[31] Peter Csermely, Tamás Korcsmáros, Huba J.M. Kiss, Gábor London, and Ruth Nussinov. "Structure and dynamics of molecular networks: A novel paradigm of drug discovery". In: *Pharmacology & Therapeutics* 138 (3 June 2013), pp. 333–408. ISSN: 01637258. DOI: 10.1016/j.pharmthera.2013.01.016.

[32] A. Eldar, H. Rozenberg, Y. Diskin-Posner, R. Rohs, and Z. Shakked. "Structural studies of p53 inactivation by DNA-contact mutations and its rescue by suppressor mutations via alternative protein-DNA interactions". In: *Nucleic Acids Research* 41 (18 Oct. 2013), pp. 8748–8759. ISSN: 0305-1048. DOI: 10.1093/nar/gkt630.

[33] Gregory M. Findlay, Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. "Accurate classification of BRCA1 variants with saturation genome editing". In: *Nature* 562 (7726 Oct. 2018), pp. 217–222. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0461-z.

[34] Warren Francis. *pdbcolor*. https://github.com/wrf/pdbcolor. 2023.

[35] P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. "A census of human cancer genes". In: *Nature Reviews Cancer* 4 (3 Mar. 2004), pp. 177–183. ISSN: 1474-175X. DOI: 10.1038/nrc1299.

[36] Gaurav D. Gaiha, Elizabeth J. Rossin, Jonathan Urbach, Christian Landeros, David R. Collins, Chioma Nwonu, Itai Muzhingi, Melis N. Anahtar, Olivia M. Waring, Alicja Piechocka-Trocha, Michael Waring, Daniel P. Worrall, Musie S. Ghebremichael, Ruchi M. Newman, Karen A. Power, Todd M. Allen, James Chodosh, and Bruce D. Walker. "Structural topology defines protective CD8 ¡sup¿+¡/sup¿ T cell epitopes in the HIV proteome". In: *Science* 364 (6439 May 2019), pp. 480–484. ISSN: 0036-8075. DOI: 10.1126/science.aav5095.

[37] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal". In: *Science Signaling* 6 (269 Apr. 2013). ISSN: 1945-0877. DOI: 10.1126/scisignal.2004088.

[38] A. Garcia-Robledo, A. Diaz-Perez, and G. Morales-Luna. "Characterization and Traversal of Large Real-World Networks". In: Elsevier, 2016, pp. 119–136. DOI: 10.1016/B978-0-12-805394-2.00005-2.

[39] Ankush Garg, Jagadish Prasad Hazra, Malay Kumar Sannigrahi, Sabyasachi Rakshit, and Sharmistha Sinha. "Variable Mutations at the p53-R273 Oncogenic Hotspot Position Leads to Altered Properties". In: *Biophysical Journal* 118 (3 Feb. 2020), pp. 720–728. ISSN: 00063495. DOI: 10.1016/j.bpj.2019.12.015.

[40] Lukas Gerasimavicius, Benjamin J. Livesey, and Joseph A. Marsh. "Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure". In: *Nature Communications* 13 (1 July 2022), p. 3895. ISSN: 2041-1723. DOI: 10.1038/s41467-022-31686-6.

[41] Nina M Goodey and Stephen J Benkovic. "Allosteric regulation and catalysis emerge via a common route". In: *Nature Chemical Biology* 4 (8 Aug. 2008), pp. 474–482. ISSN: 1552-4450. DOI: 10.1038/nchembio.98.

[42] L. H. Greene. "Protein structure networks". In: *Briefings in Functional Genomics* 11 (6 Nov. 2012), pp. 469–478. ISSN: 2041-2649. DOI: 10.1093/bfgp/els039.

[43] Rajdeep Grewal and Soumen Roy. "Modeling proteins as residue interaction networks". In: *Protein & Peptide Letters* 22 (10 Aug. 2015), pp. 923–933. ISSN: 09298665. DOI: 10.2174/0929866522666150728115552.

[44] M. Michael Gromiha. *Protein Bioinformatics: From Sequence to Function*. Academic Press, 2010. ISBN: 9788131222973. DOI: 10.1016/C2009-0-63223-2.

[45] Peter William Gunning. "Protein Isoforms and Isozymes". In: Wiley, Jan. 2006. DOI: 10.1038/npg.els.0005717.

[46] Blake M. Hauser, Yuyang Luo, Anusha Nathan, Ahmad Al-Moujahed, Demetrios G. Vavvas, Jason Comander, Eric A. Pierce, Emily M. Place, Kinga M. Bujakowska, Gaurav D. Gaiha, and Elizabeth J. Rossin. "Structure-based network analysis predicts pathogenic variants in human proteins associated with inherited retinal disease". In: *npj Genomic Medicine* 9 (1 May 2024), p. 31. ISSN: 2056-7944. DOI: 10.1038/s41525-024-00416-w.

[47] Scott A. Hollingsworth and Ron O. Dror. "Molecular Dynamics Simulation for All". In: *Neuron* 99 (6 Sept. 2018), pp. 1129–1143. ISSN: 08966273. DOI: 10.1016/j.neuron.2018.08.011.

[48] Lamei Huang, Zhixing Guo, Fang Wang, and Liwu Fu. "KRAS mutation: from undruggable to druggable in cancer". In: *Signal Transduction and Targeted Therapy* 6 (1 Nov. 2021), p. 386. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00780-4.

[49] A. C. Joerger, H. C. Ang, and A. R. Fersht. "Structural basis for understanding oncogenic p53 mutations and designing rescue drugs". In: *Proceedings of the National Academy of Sciences* 103 (Oct. 2006), pp. 15056–15061. DOI: 10.1073/pnas.0607286103. (Visited on 09/23/2021).

[50] Andreas C. Joerger, Hwee Ching Ang, Dmitry B. Veprintsev, Caroline M. Blair, and Alan R. Fersht. "Structures of p53 Cancer Mutants and Mechanism of Rescue by Second-site Suppressor Mutations". In: *Journal of Biological Chemistry* 280 (16 Apr. 2005), pp. 16030–16037. ISSN: 00219258. DOI: 10.1074/jbc.M500179200.

[51] N. Kannan and S. Vishveshwara. "Identification of side-chain clusters in protein structures by a graph spectral method 1 1Edited by J. M. Thornton". In: *Journal of Molecular Biology* 292 (2 Sept. 1999), pp. 441–464. ISSN: 00222836. DOI: 10.1006/jmbi.1999.3058.

[52] Eran Kotler, Odem Shani, Guy Goldfeld, Maya Lotan-Pompan, Ohad Tarcic, Anat Gershoni, Thomas A. Hopf, Debora S. Marks, Moshe Oren, and Eran Segal. "A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation". In: *Molecular Cell* 71 (1 July 2018), 178–190.e8. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.06.012.

[53] Michael Krone, Katrin Bidmon, and Thomas Ertl. *GPU-based Visualisation of Protein Secondary Structure.* Jan. 2008, pp. 115–122. DOI: 10.2312/LocalChapterEvents/TPCG/TPCG08/115-122.

[54] D. P. Lane. "p53, guardian of the genome". In: *Nature* 358 (6381 July 1992), pp. 15–16. ISSN: 0028-0836. DOI: 10.1038/358015a0.

[55] Sebastien Lelong, Xinghua Zhou, Cyrus Afrasiabi, Zhongchao Qian, Marco Alvarado Cano, Ginger Tsueng, Jiwen Xin, Julia Mullen, Yao Yao, Ricardo Avila, Greg Taylor, Andrew I Su, and Chunlei Wu. "BioThings SDK: a toolkit for building high-performance data APIs in biomedical research". In: *Bioinformatics* 38 (7 Mar. 2022), pp. 2077–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac017.

[56] Yizhou Li, Zhining Wen, Jiamin Xiao, Hui Yin, Lezheng Yu, Li Yang, and Menglong Li. "Predicting disease-associated substitution of a single amino acid by analyzing residue interactions". In: *BMC Bioinformatics* 12 (1 Dec. 2011), p. 14. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-14.

[57] Zhongjie Liang, Gennady M Verkhivker, and Guang Hu. "Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications". In: *Briefings in Bioinformatics* 21 (3 May 2020), pp. 815–835. ISSN: 1467-5463. DOI: 10.1093/bib/bbz029.

[58] Kamil A. Lipinski, Louise J. Barber, Matthew N. Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. "Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine". In: *Trends in Cancer* 2 (1 Jan. 2016), pp. 49–63. ISSN: 24058033. DOI: 10.1016/j.trecan.2015.11.003.

[59] Michael J Lopez and Shamin S Mohiuddin. *Biochemistry, Essential Amino Acids.* Jan. 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK557845/.

[60] Xavier Martinez, Matthieu Chavent, and Marc Baaden. "Visualizing protein structures — tools and trends". In: *Biochemical Society Transactions* 48 (2 Apr. 2020), pp. 499–506. ISSN: 0300-5127. DOI: 10.1042/BST20190621.

[61] K. G. McLure. "How p53 binds DNA as a tetramer". In: *The EMBO Journal* 17 (12 June 1998), pp. 3342–3350. ISSN: 14602075. DOI: 10.1093/emboj/17.12.3342.

[62] Taylor L. Mighell, Sara Evans-Dutson, and Brian J. O'Roak. "A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships". In: *The American Journal of Human Genetics* 102 (5 May 2018), pp. 943–955. ISSN: 00029297. DOI: 10.1016/j.ajhg.2018.03.018.

[63] Cameron Mura, Colin M. McCrimmon, Jason Vertrees, and Michael R. Sawaya. "An Introduction to Biomolecular Graphics". In: *PLoS Computational Biology* 6 (8 Aug. 2010), e1000918. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000918.

[64] National Cancer Institute. *Common Cancer Sites - Cancer Stat Facts.* SEER, 2024. URL: https://seer.cancer.gov/statfacts/html/common.html (visited on 06/14/2024).

[65] Simona Negrini, Vassilis G. Gorgoulis, and Thanos D. Halazonetis. "Genomic instability — an evolving hallmark of cancer". In: *Nature Reviews Molecular Cell Biology* 11 (3 Mar. 2010), pp. 220–228. ISSN: 1471-0072. DOI: 10.1038/nrm2858.

[66] Ryo Nitta, Tsuyoshi Imasaki, and Eriko Nitta. "Recent progress in structural biology: lessons from our research history". In: *Microscopy* 67 (4 Aug. 2018), pp. 187–195. ISSN: 2050-5698. DOI: 10.1093/jmicro/dfy022. URL: https://doi.org/10.1093/jmicro/dfy022.

[67] M. Olivier, M. Hollstein, and P. Hainaut. "TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use". In: *Cold Spring Harbor Perspectives in Biology* 2 (1 Jan. 2010), a001008–a001008. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a001008.

[68] Arthur J. Olson. "Perspectives on Structural Molecular Biology Visualization: From Past to Present". In: *Journal of Molecular Biology* 430 (21 Oct. 2018), pp. 3997–4012. ISSN: 00222836. DOI: 10.1016/j.jmb.2018.07.009.

[69] OpenAI. *ChatGPT*. 2024. URL: https://chat.openai.com/chat.

[70] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani. "Protein Contact Networks: An Emerging Paradigm in Chemistry". In: *Chemical Reviews* 113 (3 Mar. 2013), pp. 1598–1613. ISSN: 0009-2665. DOI: 10.1021/cr3002356.

[71] Marharyta Petukh, Tugba G. Kucukkal, and Emil Alexov. "On Human Disease-Causing Amino Acid Variants: Statistical Study of Sequence and Structural Patterns". In: *Human Mutation* 36 (5 May 2015), pp. 524–534. ISSN: 10597794. DOI: 10.1002/humu.22770.

[72] George N Phillips. "Describing protein conformational ensembles: beyond static snapshots". In: *F1000 Biology Reports* 1 (May 2009). ISSN: 1757594X. DOI: 10.3410/B1-38.

[73] Damiano Piovesan, Giovanni Minervini, and Silvio C.E. Tosatto. "The RING 2.0 web server for high quality residue interaction networks". In: *Nucleic Acids Research* 44 (W1 July 2016), W367–W374. ISSN: 0305-1048. DOI: 10.1093/nar/gkw315.

[74] Kevin W Plaxco, Stefan Larson, Ingo Ruczinski, David S Riddle, Edward C Thayer, Brian Buchwitz, Alan R Davidson, and David Baker. "Evolutionary conservation in protein folding kinetics". In: *Journal of Molecular Biology* 298 (2 Apr. 2000), pp. 303–312. ISSN: 00222836. DOI: 10.1006/jmbi.1999.3663.

[75] Julia R. Pon and Marco A. Marra. "Driver and Passenger Mutations in Cancer". In: *Annual Review of Pathology: Mechanisms of Disease* 10 (1 Jan. 2015), pp. 25–50. ISSN: 1553-4006. DOI: 10.1146/annurev-pathol-012414-040312.

[76] RCSB Protein Data Bank. *PDB Statistics: Overall Growth of Released Structures Per Year*. www.rcsb.org, 2024. URL: https://www.rcsb.org/stats/growth/growth-released-structures (visited on 06/14/2024).

[77] N. Rivlin, R. Brosh, M. Oren, and V. Rotter. "Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis". In: *Genes & Cancer* 2 (4 Apr. 2011), pp. 466–474. ISSN: 1947-6019. DOI: 10.1177/1947601911408889.

[78] Yana Rose, Jose M. Duarte, Robert Lowe, Joan Segura, Chunxiao Bi, Charmi Bhikadiya, Li Chen, Alexander S. Rose, Sebastian Bittrich, Stephen K. Burley, and John D. Westbrook. "RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive". In: *Journal of Molecular Biology* 433 (11 May 2021), p. 166704. ISSN: 00222836. DOI: 10.1016/j.jmb.2020.11.003.

[79] Serena Rosignoli and Alessandro Paiardini. “Boosting the Full Potential of PyMOL with Structural Biology Plugins”. In: *Biomolecules* 12 (12 Nov. 2022), p. 1764. ISSN: 2218-273X. DOI: 10.3390/biom12121764.

[80] Serena Rosignoli, Luisa di Paola, and Alessandro Paiardini. “PyPCN: protein contact networks in PyMOL”. In: *Bioinformatics* 39 (11 Nov. 2023). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad675.

[81] Chris Sander and Reinhard Schneider. “Database of homology-derived protein structures and the structural meaning of sequence alignment”. In: *Proteins: Structure, Function, and Bioinformatics* 9 (1 Jan. 1991), pp. 56–68. ISSN: 0887-3585. DOI: 10.1002/prot.340090107.

[82] Dmitrii Shcherbinin and Alexander Veselovsky. “Analysis of Protein Structures Using Residue Interaction Networks”. In: 2019, pp. 55–69. DOI: 10.1007/978-3-030-05282-9_3.

[83] Carlos H. da Silveira, Douglas E. V. Pires, Raquel C. Minardi, Cristina Ribeiro, Caio J. M. Veloso, Julio C. D. Lopes, Wagner Meira, Goran Neshich, Carlos H. I. Ramos, Raul Habesch, and Marcelo M. Santoro. “Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 74 (3 Feb. 2009), pp. 727–743. ISSN: 0887-3585. DOI: 10.1002/prot.22187.

[84] Vladimir Sladek, Hiroaki Tokiwa, Hitoshi Shimano, and Yasuteru Shigeta. “Protein Residue Networks from Energetic and Geometric Data: Are They Identical?” In: *Journal of Chemical Theory and Computation* 14 (12 Dec. 2018), pp. 6623–6631. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.8b00733.

[85] Vladimir Sladek, Yuta Yamamoto, Ryuhei Harada, Mitsuo Shoji, Yasuteru Shigeta, and Vladimir Sladek. “pyProGA—A PyMOL plugin for protein residue network analysis”. In: *PLOS ONE* 16 (7 July 2021), e0255167. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0255167.

[86] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. “Residues crucial for maintaining short paths in network communication mediate signaling in proteins”. In: *Molecular Systems Biology* 2 (1 Jan. 2006). ISSN: 1744-4292. DOI: 10.1038/msb4100063.

[87] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers”. In: *Nature Reviews Cancer* 18 (11 Nov. 2018), pp. 696–705. ISSN: 1474-175X. DOI: 10.1038/s41568-018-0060-1.

[88] Huaxin Song, Jiale Wu, Yigang Tang, Yuting Dai, Xinrong Xiang, Ya Li, Lili Wu, Jiaqi Wu, Ying Liang, Yangfei Xing, Ni Yan, Yuntong Li, Zhengyuan Wang, Shujun Xiao, Jiabing Li, Derun Zheng, Xinjie Chen, Hai Fang, Chenjing Ye, Yuting Ma, Yu Wu, Wen Wu, Junming Li, Sujiang Zhang, and Min Lu. "Diverse rescue potencies of p53 mutations to ATO are predetermined by intrinsic mutational properties". In: *Science Translational Medicine* 15 (690 Apr. 2023). ISSN: 1946-6234. DOI: 10.1126/scitranslmed.abn9155.

[89] Valentina Sora, Matteo Tiberti, Ludovica Beltrame, Deniz Dogan, Shahriyar Mahdi Robbani, Joshua Rubin, and Elena Papaleo. "PyInteraph2 and PyInKnife2 to Analyze Networks in Protein Structural Ensembles". In: *Journal of Chemical Information and Modeling* 63 (14 July 2023), pp. 4237–4245. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.3c00574.

[90] Cristina Sotomayor-Vivas, Enrique Hernández-Lemus, and Rodrigo Dorantes-Gilardi. "Linking protein structural and functional change to mutation using amino acid networks". In: *PLOS ONE* 17 (Jan. 2022). Ed. by Sriparna Saha, e0261829. DOI: 10.1371/journal.pone.0261829.

[91] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. "COSMIC: the Catalogue Of Somatic Mutations In Cancer". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D941–D947. ISSN: 0305-1048. DOI: 10.1093/nar/gky1015.

[92] MIT The Walker Lab @ The Ragon Institute of Harvard, MGH, and Olivia Waring. *WalkerLabRagon/NetworkAnalysis: Network Analysis Pipeline*. Version v1.0. Mar. 2019. DOI: 10.5281/zenodo.2597484. URL: https://doi.org/10.5281/zenodo.2597484.

[93] Philip J. Thomas, Bao-He Qu, and Peter L. Pedersen. "Defective protein folding as a basis of human disease". In: *Trends in Biochemical Sciences* 20 (11 Nov. 1995), pp. 456–459. ISSN: 09680004. DOI: 10.1016/S0968-0004(00)89100-8.

[94] Matteo Tiberti, Gaetano Invernizzi, Matteo Lambrughi, Yuval Inbar, Gideon Schreiber, and Elena Papaleo. "PyInteraph: A Framework for the Analysis of Interaction Networks in Structural Ensembles of Proteins". In: *Journal of Chemical Information and Modeling* 54 (5 May 2014), pp. 1537–1551. ISSN: 1549-9596. DOI: 10.1021/ci400639r.

[95] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus. "Small-world view of the amino acids that play a key role in protein folding". In: *Physical Review E* 65 (6 June 2002), p. 061910. ISSN: 1063-651X. DOI: 10.1103/PhysRevE.65.061910.

[96] Gennady M. Verkhivker. "Biophysical simulations and structure-based modeling of residue interaction networks in the tumor suppressor proteins reveal functional role of cancer mutation hotspots in molecular communication". In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1863 (1 Jan. 2019), pp. 210–225. ISSN: 03044165. DOI: 10.1016/j.bbagen.2018.10.009.

[97] Juan Salamanca Viloria, Maria Francesca Allega, Matteo Lambrughi, and Elena Papaleo. "An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass". In: *Scientific Reports* 7 (1 June 2017), p. 2838. ISSN: 2045-2322. DOI: 10.1038/s41598-017-01498-6.

[98] Saraswathi Vishveshwara, Amit Ghosh, and Priti Hansia. "Intra and Inter-Molecular Communications Through Protein Structure Network". In: *Current Protein & Peptide Science* 10 (2 Apr. 2009), pp. 146–160. ISSN: 13892037. DOI: 10.2174/138920309787847590.

[99] Haolan Wang, Ming Guo, Hudie Wei, and Yongheng Chen. "Targeting p53 pathways: mechanisms, structures, and advances in therapy". In: *Signal Transduction and Targeted Therapy* 8 (1 Mar. 2023), p. 92. ISSN: 2059-3635. DOI: 10.1038/s41392-023-01347-1.

[100] Zhen Wang and John Moult. "SNPs, protein structure, and disease". In: *Human Mutation* 17 (4 Apr. 2001), pp. 263–270. ISSN: 1059-7794. DOI: 10.1002/humu.22.

[101] Kam-Bo Wong, Brian S. DeDecker, Stefan M. V. Freund, Mark R. Proctor, Mark Bycroft, and Alan R. Fersht. "Hot-spot mutants of p53 core domain evince characteristic local structural changes". In: *Proceedings of the National Academy of Sciences* 96 (15 July 1999), pp. 8438–8442. ISSN: 0027-8424. DOI: 10.1073/pnas.96.15.8438.

[102] World Health Organization. *The Top 10 Causes of Death*. World Health Organization, Dec. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[103] Chunlei Wu, Ian MacLeod, and Andrew I. Su. "BioGPS and MyGene.info: organizing online, gene-centric information". In: *Nucleic Acids Research* 41 (D1 Jan. 2013), pp. D561–D565. ISSN: 0305-1048. DOI: 10.1093/nar/gks1114.

[104] Jiajing Wu, Jieli Liu, Yijing Zhao, and Zibin Zheng. "Analysis of cryptocurrency transactions from a network perspective: An overview". In: *Journal of Network and Computer Applications* 190 (Sept. 2021), p. 103139. ISSN: 10848045. DOI: 10.1016/j.jnca.2021.103139.

[105] Jiwen Xin, Adam Mark, Cyrus Afrasiabi, Ginger Tsueng, Moritz Juchler, Nikhil Gopal, Gregory S. Stupp, Timothy E. Putman, Benjamin J. Ainscough, Obi L. Griffith, Ali Torkamani, Patricia L. Whetzel, Christopher J. Mungall, Sean D. Mooney, Andrew I. Su, and Chunlei Wu. "High-performance web services for querying gene and variant annotation". In: *Genome Biology* 17 (1 Dec. 2016), p. 91. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0953-9.

[106] Christopher M. Yates and Michael J.E. Sternberg. "The Effects of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs) on Protein–Protein Interactions". In: *Journal of Molecular Biology* 425 (21 Nov. 2013), pp. 3949–3963. ISSN: 00222836. DOI: 10.1016/j.jmb.2013.07.012.

[107] Shuguang Yuan, H.C. Stephen Chan, and Zhenquan Hu. "Using PyMOL as a platform for computational drug design". In: *WIREs Computational Molecular Science* 7 (2 Mar. 2017). ISSN: 1759-0876. DOI: 10.1002/wcms.1298.