ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Stanford University

Multidimensional Insights into Hypertrophic Cardiomyopathy

An Integrative Analysis using Structural, Clinical, and AlphaMissense Data

Master's Thesis

Christian Michael Cadisch
cadischc@student.ethz.ch

Supervisors: Prof. Dr. Euan Ashley
Prof. Dr. Valentina Boeva
Advisors: Prof. Dr. Victoria Parikh
Dr. Yuta Yamamoto
Dr. Laurens van de Wiel

September 2024

Abstract

Hypertrophic cardiomyopathy (HCM), the most common heritable heart disease, is characterized by thickening of the left ventricular wall, reducing the ventricular cavity and impairing cardiac function. Affecting about 1 in 500 individuals, HCM presents a wide range of symptoms and is the leading cause of sudden cardiac arrests in young people. Despite recent advancements, genetic testing for HCM has only a 30-50% success rate, leaving many genetic causes of the disease still unknown.

Advances in genome sequencing have enriched genetic datasets. In this work, we leveraged the largest dataset ever used in HCM research, aggregating data from two large-scale population datasets and contrasting them with exome sequencing data from patients with HCM. Additionally, we developed and evaluated two strategies to include novel in silico predictions from the transformer-based model AlphaMissense, further expanding the underlying dataset used for the analysis. We generated structural models of disease-related proteins and applied a spatial scan statistic to identify spherical regions enriched with disease-associated mutations. This approach enabled the detection of mutational hotspots on 3D protein structures, offering valuable insights into potential disease mechanisms.

We identified 16 regions enriched for disease-associated variants across six genes involved in the pathogenesis of HCM. By simulating protein interactions and incorporating experimental structures, we proposed disease pathways for the identified hotspot regions. Finally, we validated our findings using ClinVar as an independent, high-quality dataset, corroborating the robustness of our results.

Integrating clinical data, structural data, and AlphaMissense predictions has proven valuable and may be applicable to other diseases, underscoring the potential of integrative, multi-dimensional approaches to enhance our understanding of genetic variation and its impact on disease. Future research could extend this approach by incorporating experimental data from multiplexed assays in human-induced pluripotent stem cells (hiPSCs), further enhancing disease hotspot detection and providing deeper insights into disease pathways.

Acknowledgements

I would like to express my deepest gratitude to Dr. Ashley for his invaluable supervision, overall guidance, and mentorship throughout this project. I am equally thankful to Dr. Boeva for enabling inter-institutional research and providing the supervision of my thesis. Special thanks go to Dr. Parikh, whose expert insights into HCM significantly enriched the medical relevance of this work, and to Dr. Yamamoto, whose assistance in biological and gene-specific analysis was essential to its success. Finally, I am sincerely grateful to Dr. van de Wiel, whose expertise in computational biology and structure prediction greatly enhanced the technical rigor of this research.

Contents

Chapter 1

Introduction

This chapter begins with a background in molecular biology needed for understanding the scope and objectives of this work. Next, it introduces the clinical manifestations of hypertrophic cardiomyopathy (HCM) and provides an overview of the current understanding of the genetic causes of the disease. Finally, we summarize the contributions of this research and draw an outline of the structure of the thesis.

1.1 Background on Molecular Biology

1.1.1 Flow of Genetic Information: From DNA to Functional Proteins

The central dogma of molecular biology, first stated by Francis Crick, describes how genetic information flows from DNA to RNA and from RNA to proteins [7]. DNA encodes for the genetic information via sequences of the four different nucleotides C (Cytosine), T (Thymine), A (Adenine), and G (Guanine). During transcription, RNA polymerase synthesizes a complementary single-stranded pre-mRNA. In humans and other eukaryotic organisms, this pre-mRNA undergoes processing, which includes splicing, to remove the non-coding intronic sequences and join the coding exonic regions. The mature mRNA is then exported to the cytoplasm. In translation, ribosomes read the mRNA codons (triplets of nucleotides) and, with the help of transfer RNA (tRNA), assemble the corresponding chain of amino acids. This polypeptide chain then folds into its three-dimensional structure, driven by the chemical interactions between the amino acids and their environment.

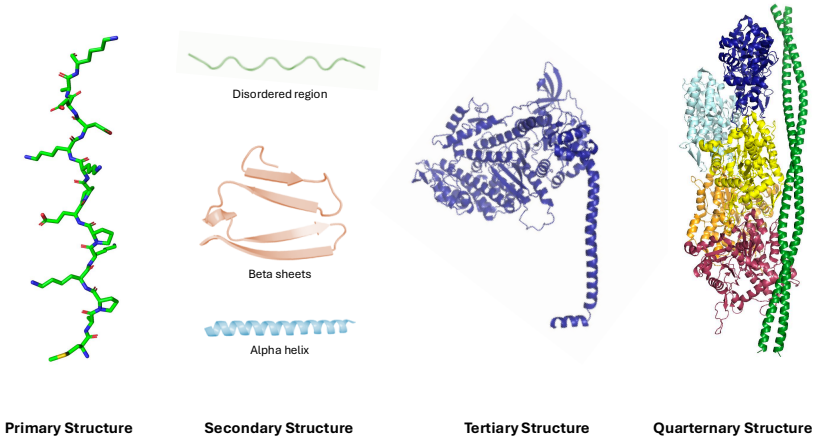

Figure 1.1: Hierarchical Organization of Protein Structure: This illustration depicts the four levels of protein structure: primary (amino acid sequence), secondary (disordered regions, alpha-helices, and beta-sheets), tertiary (three-dimensional folding), and quaternary (assembly of multiple subunits). The tropomyosin complex, shown here with PDB ID 5JLF [12], exemplifies how these structural levels collectively build the protein structure.

1.1.2 Protein Structure, Function, and Folding Dynamics

Proteins have four levels of structure that determine their function (see Figure 1.1). The primary structure is the linear sequence of amino acids linked by peptide bonds. The secondary structure refers to local sub-structures like alpha-helices and beta-sheets, formed by hydrogen bonds along the polypeptide backbone. The tertiary structure is the three-dimensional shape of a single protein molecule, in which alpha-helices and beta-sheets fold into a compact, globular form. The quaternary structure arises when multiple protein subunits assemble to form a functional multimer. The specific arrangement of these structures dictates the protein's function, as its shape and chemical properties enable interactions with other molecules, driving biological processes.

1.1.3 Genetic Variation and Mutations

Genetic variation refers to differences in DNA sequences among individuals, which can be identified through genome sequencing. While many variants have little to no effect on the phenotype, some can lead to pathogenic outcomes. The ultimate source of genetic variation is mutations, which can occur in both exonic and intronic regions

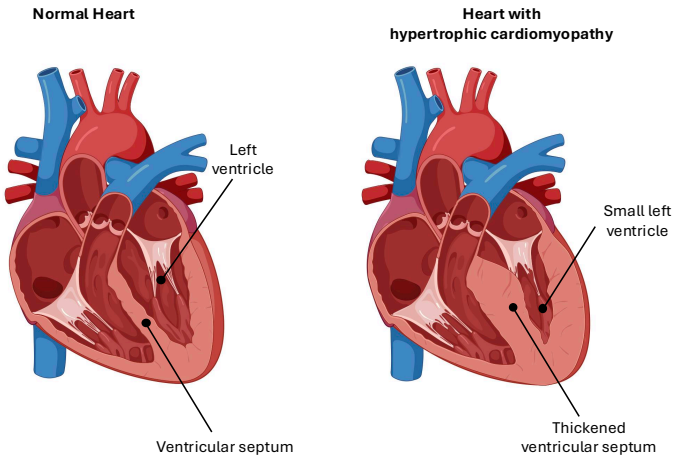

Figure 1.2: Comparison of a healthy heart with one affected by hypertrophic cardiomyopathy (HCM), highlighting the thickening of the left ventricular wall, which reduces the ventricular cavity and impairs normal cardiac function. Image created using Biorender.

of the genome. Intronic mutations occur within the non-coding regions of a gene and may affect gene expression or splicing. In contrast, exonic mutations occur within the coding regions and directly impact the protein sequence. Among exonic variants, there are several types: missense mutations, which change an amino acid; nonsense mutations, that introduce a stop codon; frameshift mutations, which disrupt the reading frame; and silent mutations, which do not alter the amino acid sequence.

1.2 Background on Hypertrophic Cardiomyopathy

1.2.1 Clinical Manifestation of Hypertrophic Cardiomyopathy

HCM is a genetic disease of the heart muscle. HCM is characterized by a thickened left ventricular wall causing a reduction of the left ventricular cavity (see Figure 1.2). It is the most common heritable heart disease, with an estimated prevalence of one in 500 individuals [39].

The clinical course of HCM is highly variable. Individuals with HCM may experience a variety of symptoms such as shortness of breath during physical activity,

fatigue, a fast or irregular heartbeat, dizziness, fainting, and unusual chest pain. In severe cases, these symptoms can escalate, leading to the risk of sudden cardiac death (SCD) [59]. Notably, HCM is the primary cause of sudden cardiac death in adolescents and young adults, especially in athletes [40]. Tragically, SCD can sometimes be the initial indication of the disease. Despite the occurrence of SCD, the condition is generally considered relatively mild and about two-thirds of individuals with HCM have a normal life span without major health issues [13].

The diagnosis of HCM is determined by the presence of left ventricular hypertrophy, typically defined in adults as an end-diastolic ventricular septal thickness of 13 mm or more. This thickening occurs without any underlying conditions that could explain it, such as high blood pressure, aortic stenosis, or the normal hypertrophy seen in athletes [38]. The current approach to HCM treatment centers on two primary goals: first, managing symptoms, and second, evaluating risks to prevent sudden cardiac death (SCD) [59]. Pharmacologic therapy, primarily using beta-blockers and calcium channel blockers is used to control the excessive systolic function and arrhythmias observed [28]. Despite optimal pharmacological therapy, some patients may still require mechanical interventions. These interventions include surgically removing the obstructive myocardial tissue (myectomy) or injecting alcohol into the coronary artery supplying the septum, to induce a controlled heart attack, which helps reduce the extra tissue [28].

1.2.2 Genetic Basis of Hypertrophic Cardiomyopathy

Genetic studies have linked mutations in over a dozen genes encoding sarcomere-associated proteins to patients with HCM. The sarcomere, the basic contractile unit of muscle, is understood to play a central role in the development of HCM, making it a disease of the sarcomere [32]. However, there remain significant challenges in the interpretation of genetic data. The success rate of finding causative mutations through genetic testing remains variable, ranging from 30% to 50% [19]. The sections below provide detailed insights into the structure and function of the myosin proteins MYH7 and MYBPC3; the troponin proteins TNNT2 and TNNI3; and the regulatory proteins MYL2 and MYL3; as well as their association with HCM.

MYH7

The MYH7 gene encodes the beta-myosin heavy chain, a vital protein within the thick filaments of the sarcomere. These thick filaments, essential for generating muscle contraction, are found in cardiac myocytes — the specialized muscle cells of the heart

— and are expressed to a lesser extent in skeletal muscle fibers. Structurally, MYH7 includes a globular head where ATP binding, hydrolysis, and actin interaction occur, followed by a lever arm and a long alpha-helical coiled-coil rod domain that makes up most of the protein. This protein is vital for muscle contraction, forming a hexameric complex with MYL2 and MYL3, which is essential for generating the force of contraction through ATP hydrolysis. MYH7's significance in cardiac function makes it a central contributor to the pathogenesis of HCM. Over 200 mutations, primarily missense, have been identified, with a significant concentration in the motor domain of the head region, which the American College of Medical Genetics (ACMG) recognizes as a mutation hotspot [19]. In this study, for the identification of disease-hotspots within MYH7, we focus on the head domain. However, even within this region, significant variations in severity and prognosis exist. Past research has identified the converter domain and the myosin mesa as a hotspot for variants causing HCM which are also associated with a more severe prognosis and earlier disease onset [18][25].

MYBPC3

Mutations in the MYBPC3 gene, encoding myosin-binding protein C, are the most frequent cause of HCM, yet they are often considered less severe compared to mutations in other sarcomeric genes [19]. The variants typically result in a later onset and a more favorable prognosis for the patient. MYBPC3 is crucial for cardiac muscle function, as it binds to MYH7 and cardiac actin, playing a key role in regulating actomyosin interactions. A distinctive feature of MYBPC3 mutations is the high number of truncating variants, which cause premature termination codons and thereby lead to an incomplete expression of the protein. The truncated proteins are often targeted by nonsense-mediated decay, resulting in protein haploinsufficiency [37]. While missense mutations are less common, they can disrupt essential residues that are vital for the protein's structural integrity or function. Despite the recognized significance of MYBPC3 in HCM, the precise molecular mechanisms through which these mutations contribute to the disease are still not fully understood. Conflicting findings from previous studies underscore the necessity for further investigation to elucidate these mechanisms.

TNNT2

The TNNT2 gene encodes cardiac troponin T, which works together with cardiac troponin C (encoded by TNNC1) and cardiac troponin I (encoded by TNNI3) to form the troponin complex. TNNT2 plays a key role by anchoring this complex to tropomyosin 1 (TPM1) [37]. Mutations in TNNT2 account for approximately 15% of HCM cases.

Structurally, TNNT2 consists of three regions: a hypervariable region (residues 1-79), TNT1 (residues 80-180), which interacts with the overlapping heads of tropomyosin dimers and contains over 65% of cardiomyopathy-linked substitutions, and the C-terminal TNT2 (residues 181-289), which includes a 25-residue N-terminal tropomyosin binding region, with its C terminus integrated into the globular end of the troponin complex [17].

TNNI3

TNNI3 encodes cardiac troponin I, the inhibitory subunit of the thin filament, which facilitates cardiac relaxation by restricting actin-myosin interactions during low calcium levels [29]. Within the troponin complex, TNNI3 is the second most frequently implicated gene for causing HCM, accounting for about 5% of all cases [20]. TNNI3 can be divided into five functional domains, with the majority of pathogenic mutations occurring in the inhibitory domain (residues 128–147), which binds to cardiac troponin C and actin-tropomyosin, or in the mobile domain (residues 164–210).

MYL2

The regulatory light chain MYL2 is involved in structurally supporting the lever-arm region of MYH7 and also has a role in modulating the myosin motor activity and thus influence the cardiac muscle contraction. In MYL2, past research has failed to identify a clear correlation between the location of variants and the resulting phenotype [64].

MYL3

MYL3, encoding the essential light chain, is similar to MYL2 in being relevant for supporting the structure of MYH7 and regulating the myosin motor activity. Compared to MYH7 and MYBPC3, variants in both light chains are quite rare causes for HCM, together accounting for less than 1% of HCM cases [37]. Yet, there are multiple variants that have been associated with severe outcomes [64]. Yadav et al. have reported that most known pathogenic mutations in MYL3 are located in the non-functional EF-hand Ca^{2+} binding motifs [64]. The relatively limited number of observed mutations in the myosin light chains has contributed to this area being largely understudied in scientific research.

1.3 Contributions and Thesis Organization

Advances in genome-sequencing technologies have made it possible to compile extensive datasets of human genetic variation in both general and HCM-affected populations. Additionally, advances in transformer-based computational models like AlphaMissense, which can predict the pathogenicity of missense variants, have significantly enriched the available data [5]. Furthermore, the introduction of AlphaFold has marked significant progress in solving the 'protein folding problem,' greatly enhancing the atomic-level accuracy of predicted protein structures [26].

We hypothesized that by aggregating the exomic data from gnomAD with 730,947 exomes and with data from the UK BioBank with 500,000 participants and contrasting it to the disease-specific dataset SHaRe with over 17,000 patients we could identify novel mutational hotspots and increase the understanding of the structure-function relationships within key sarcomeric genes. We created structural models of MYH7, MYBPC3, MYL2, MYL3, TNNT2 and TNNI3; mapped variants observed in the datasets to the 3D structures and we adopted a statistical approach based on the spatial scan statistic to identify regions of disease-enrichment. Moreover, we developed two statistical frameworks to incorporate in-silico pathogenicity predictions from AlphaMissense into our analysis, enhancing dataset size and increasing statistical power.

This thesis is organized into six chapters. The Introduction establishes the biological and medical context relevant to understanding the motivation and scope of this research. We then provide an overview of the related work. The Methods chapter details the computational approaches used for protein structure modeling, introduces the datasets utilized, and outlines the statistical frameworks applied. The Results section begins by evaluating the accuracy of the protein structure predictions, followed by the presentation of the identified disease hotspots and an ablation study of the mathematical frameworks used, and assess the impact of different data sources. In the Discussion section, these hotspots are contextualized against prior findings, and the performance of different statistical models is critically evaluated. Finally, the Conclusion summarizes the key contributions and findings of this research and suggests potential directions for future work.

Chapter 2

Related Work

In this chapter, we first introduce past research efforts aimed at detecting mutational hotspots on sarcomeric genes. Next, we present the architectures of the three foundational models utilized in this work, AlphaFold 2, AlphaFold 3, and ColabFold. Finally, we introduce the variant effect prediction tool AlphaMissense.

2.1 Identification of Disease Hotspots

2.1.1 Spatial Scan Statistic

Homburger et al. combined spatial models of MYH7 with variant data to identify regions enriched for variants associated with HCM [25]. The authors utilized homology modelling to create the structure of beta myosin heavy chain and mapped data from 100,000 genomes sourced from the Exome Aggregation Consortium (ExAC) and DiscovEHR to the structure. This data was contrasted with the genomes of 2,913 patients with HCM, as documented in the Sarcomeric Human Cardiomyopathy Registry (SHaRe). A spatial scan statistic was then used to identify hotspots for disease-associated variants in three-dimensional space and on the protein surface. This analysis revealed significant disease enrichment in the converter domain and the myosin mesa. Furthermore, patients with HCM harboring variants in these regions were found to have worse clinical outcomes.

2.1.2 Generalized Additive Modelling

Waring et al. introduced a generalized additive modeling (GAM) framework to identify mutational hotspots in six sarcomeric genes [63]. Their approach analyzed the linear sequence of amino acids and leveraged genetic data from a cohort of 5,338 HCM

patients and over 100,000 control samples obtained from gnomAD. Their approach significantly enhanced the detection of pathogenic variants by integrating positional information with burden testing, resulting in a statistical model for predicting variant pathogenicity.

2.1.3 Signal-to-Noise Ratio

Kurzlechner et al. used a signal-to-noise analysis comparing the frequency of exome sequencing and HCM variants with allele frequencies from the gnomAD dataset to detect hotspots of pathogenic mutations at a residue-level [31]. In the MYH7 gene, they observed that nearly all pathogenic amino acid positions were concentrated in the head domain, with two significant areas of high signal coinciding with actin binding sites at amino acids 657–671 and 762–768. Conversely, in the MYBPC3 gene, the pathogenic regions did not align with any specific domain type.

2.1.4 Domain-based Analysis

Helms et al. used data from SHaRe to conduct a comparative analysis with variants observed in gnomAD, focusing on the MYBPC3 gene [22]. The authors identified a cluster of non-truncating pathogenic variants in the C3, C6, and C10 domains. In contrast, the truncating variants were evenly dispersed throughout the gene.

2.2 Foundation Models for Protein Structure Prediction

2.2.1 AlphaFold 2

The introduction of AlphaFold 2 represented an important advance in computational biology, making considerable progress in addressing the challenge of predicting a protein's three-dimensional structure from its amino acid sequence, referred to as the "protein folding problem" [26]. Below, we describe the key concepts used in AlphaFold 2's architecture, which can also be seen in Figure 2.1.

AlphaFold 2 begins by processing the input sequence to generate two 2D matrices, which are then fed into the evoformer, a transformer-based network that further refines them. These refined matrices are subsequently used by the structure model to predict the protein's 3D conformation.

The first matrix captures multiple sequence alignments (MSAs), which are designed to identify regions of similarity across different sequences, indicating shared

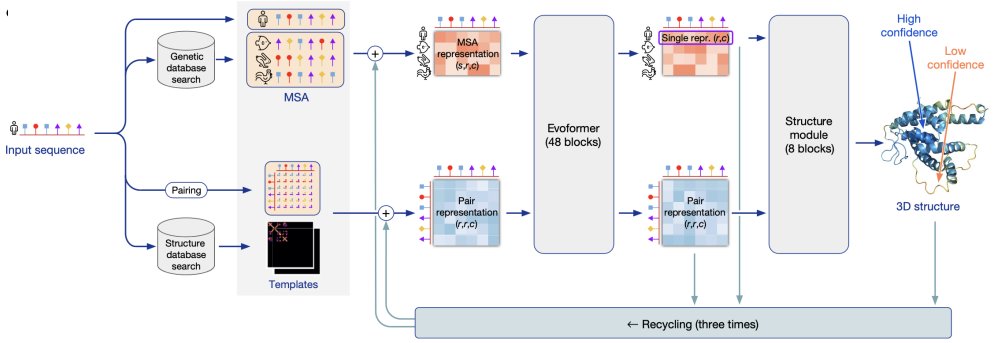

Figure 2.1: Overview of the AlphaFold 2 architecture, showcasing its workflow from input sequence processing, through multiple sequence alignments (MSAs) and pairwise residue relationships, to the final 3D protein structure prediction.

functional, structural, or evolutionary traits. In this matrix, each row represents a sequence, and each column corresponds to a specific residue position, allowing for the detection of conserved regions critical to the protein's structure and function.

The second matrix, called the pairwise representation, encodes the pairwise evolutionary relationships between individual residues. The matrix values provide insights into the interactions between these residues.

Both matrices are input into the evoformer, which consists of 48 blocks. Within each block, various forms of self-attention are applied to the MSAs and pairwise representations. The refined outputs from these matrices are then used by the structure module to accurately predict the protein's three-dimensional conformation.

2.2.2 AlphaFold 3

AlphaFold 3 builds upon the architecture of AlphaFold 2 and introduces two new modules: the pairformer module replaces the evoformer, and a diffusion module replaces the structure module (see Figure 2.2) [1]. This updated model is capable of predicting the joint structure of complexes that include proteins, nucleic acids, small molecules, ions, and modified residues.

The two main architectural innovations are the replacement of the evoformer with the pairformer and the introduction of the diffusion module in place of the previous structure module. Unlike the evoformer, the pairformer does not retain the MSA representation and instead transmits the information using the pair representation. The output of the pairformer is the updated pair and single representation, which, together with

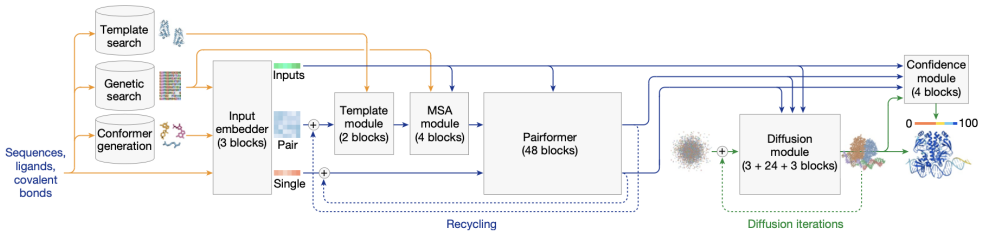

Figure 2.2: Diagram of the AlphaFold 3 architecture. The evoformer of Alphafold 2 is replaced with the pairformer and the structure module is replaced by the diffusion module.

the input representation, is passed to the new diffusion module. The diffusion module operates directly on raw atom coordinates, differing from AlphaFold 2's method based on translation and rotation properties.

2.2.3 ColabFold v2.3

ColabFold provides an accelerated approach to protein structure prediction by combining the rapid homology search capabilities of MMseqs2 with the advanced prediction architectures of AlphaFold 2 or RoseTTAFold [56] [43]. The MMseqs2 search server efficiently aligns input sequences against databases like UniRef100, PDB70, and an environmental sequence set. By substituting AlphaFold 2's homology search with the 40–60-fold faster MMseqs2, ColabFold significantly reduces the time for single predictions. Additionally, batch predictions are accelerated by approximately 90 times through the avoidance of recompilation and the introduction of an early stop criterion. Despite these optimizations, ColabFold maintains prediction quality, matching AlphaFold 2 on CASP14 targets and performing comparably to AlphaFold-multimer on the ClusPro dataset.

2.3 Variant Pathogenicity Prediction

2.3.1 AlphaMissense

The authors of AlphaMissense tackled the challenge of predicting the clinical effect of missense variants [5]. To date, only about 2% of observed missense variants have been clinically determined to be either pathogenic or benign, leaving the vast majority

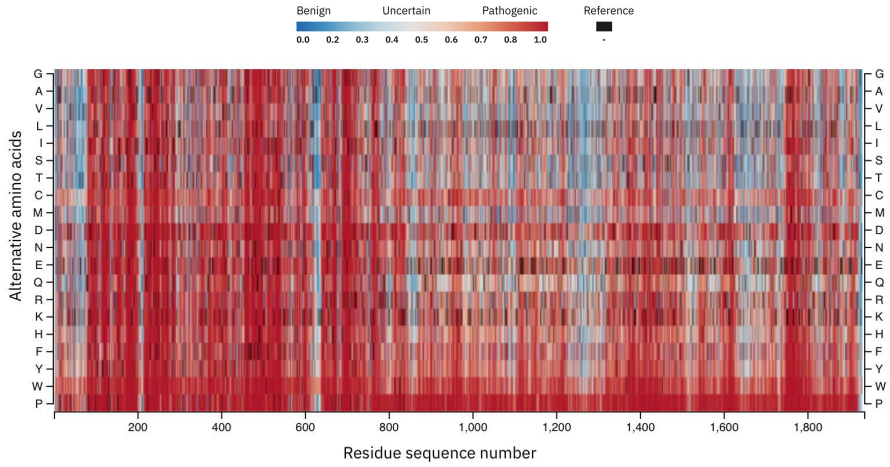

Figure 2.3: AlphaMissense pathogenicity predictions for MYH7, where more pathogenic variants are represented by red tones, while more benign variants are shown in blue tones. The wild-type amino acid is colored black. A higher concentration of predicted pathogenic variants is visible within the S1 domain (residues 1-840) compared to the S2 domain (residues 841-1,280) and LMM domain (residues 1,281-1935).

with uncertain clinical significance. This highlights the potential value of prediction models in improving the interpretation of these variants.

AlphaMissense builds on the progress made in protein language modeling to achieve state-of-the-art missense pathogenicity predictions. The architecture is largely based on the implementation of AlphaFold 2, where the structural output is replaced with a pathogenicity score. In the pre-training stage, the network is trained similarly to AlphaFold 2, focusing on structure prediction while also undergoing a masked training approach where the goal is to predict a masked amino acid based on its context. In the fine-tuning stage, the model is trained with variant frequency data to ultimately learn to predict variant pathogenicity.

The output of AlphaMissense is a continuous pathogenicity score between 0 and 1, where 0 indicates the most benign score and 1 the most pathogenic score. The authors note that the scores can be understood as an estimate of the probability that a variant is clinically pathogenic. An example heatmap of predicted scores is displayed in Figure 2.3. Here, the pathogenicity predictions for all possible missense variants are shown, where the red tones indicate the model predicts pathogenic consequences and the blue tones predict likely benign consequences. The observation that variants in the head domain of MYH7 tend to be more pathogenic (see Section 1.2.2) can be recognized by the elevated content of red tones between residues 1-840.

Chapter 3

Materials and Methods

This chapter begins by providing an overview of the modeling approaches we used to predict protein structures, followed by a presentation of the datasets included in this study. We then discuss the strategies for classifying genetic variants into categories as well as continuous scores and describe the statistical frameworks applied to identify disease-associated hotspots within the 3D protein structures. Finally, we provide insights into the validation of our findings through ClinVar and age of onset analyses.

3.1 Development of Protein Models

3.1.1 Approach for Protein Structure Modelling

To obtain optimal structural models for each protein under study, we evaluated four computational modeling approaches and selected the most accurate models based on a set of evaluation criteria.

Firstly, we obtained pre-computed models from the AlphaFold Protein Structure Database, a collaborative effort by Google DeepMind and European Bioinformatics Institute (EMBL-EBI) [61]. These models were generated using the AlphaFold Monomer v2.0 pipeline and last updated in November 1, 2022 [26]. Secondly, we utilized ColabFold v1.5.5 to create structure predictions based on the AlphaFold2 v2.2 model weights and using the faster homology search of MMseqs2 [43][56]. Thirdly, we attempted to enhance the prediction accuracy of the ColabFold model by providing experimentally obtained template structures of the respective proteins. The templates can act as a reference to the model, guiding the structure prediction process. We utilized Stanford's shared computing cluster, Sherlock, for GPU access. ColabFold was set to use a custom template mode with MMseqs2 browsing UniRef and environmen-

tal datasets. The number of recycles was set to 3 with early stopping tolerance on auto, and relax was limited to 200 iterations. We employed a greedy pairing strategy, and dropout was disabled. The templates used for each protein are described in Section 3.1.2. Lastly, we used the AlphaFold Server to create structure predictions with AlphaFold 3 [1]. Additionally, for all protein-ion and protein-ligand interactions we used AlphaFold 3. For the analysis of MYH7, we focused exclusively on the subfragment 1 (S1), encompassing residues 1-840, as this region possesses distinct functional significance and is highly enriched with pathogenic variants.

3.1.2 Template Selection for Structural Modeling with ColabFold

For modelling the beta-myosin heavy chain (encoded by MYH7), we provided the model with the template structures with PDB IDs 8ACT [21], 8EFH [**?**], 8EFI [8], 8ENC [8], and 8G4L [11]. For modelling the myosin-binding protein C (encoded by MYBPC3), we provided the model with the template structures with PDB IDs 6CXJ [49], 6G2T [49], 7LRG [50], 7TIJ [51] and 8G4L [11]. For modelling the essential light chain (encoded by MYL3) and regulatory light chain (encoded by MYL2), we provided the model with the template structures with PDB IDs 5TBY [2], 8ACT [21] and 8G4L [11]. For modelling the cardiac troponin T (encoded by TNNT2) and troponin I (encoded by TNNI3), we provided the model with the template structures with PDB IDs 6KN7 [65], 6KN8 [65], 7UTI [48] and 7UTL [48].

3.1.3 Assessment and Comparison of Predicted Protein Structures

To assess and compare the quality of all protein models, we analyzed predictive and experimental metrics. The structure prediction models we used are equipped with two predictive scoring methods: predicted local distance difference test (pLDDT), giving a per-residue measure of local confidence, and the predicted aligned error (PAE), giving an estimate of the expected error between each pair of amino acids in the model and allowing us to better understand the factors contributing to model error. We also calculated the mean pLDDT and mean PAE for each protein model to gain an understanding of overall accuracy and local confidence levels. Furthermore, we aligned the predicted structures with partial experimentally determined protein structures available in PDB and measured the minimal root mean square error (RMSE) between the predicted and experimental structures. Visualizations and the alignments of structures were made using PyMol 2.5.8.

3.2 Datasets

Gene Name	UKBB+gnomAD Data		SHaRe Data		AlphaMissense
	Patients	Unique Variants	Patients	Unique Variants	Predicted Scores
MYH7	61,866	1,893	1,332	363	38,700
MYBPC3	640,171	1,852	2,523	522	20,480
TNNT2	76,958	392	218	63	5,960
TNNI3	148,175	336	184	56	4,200
MYL2	7,025	274	103	22	3,320
MYL3	2,904	231	58	30	3,900

Table 3.1: Comparison of datasets sizes included in our study: UKBB and gnomAD for the population data, SHaRe for the disease-enriched data and the pathogenicity predictions created with AlphaMissense.

3.2.1 Sarcomeric Human Cardiomyopathy Registry

The Sarcomeric Human Cardiomyopathy Registry (SHaRe) is a collaborative initiative that aggregates anonymized patient data from multiple research institutions to advance the study of cardiomyopathies, consolidating and standardizing an extensive dataset. It includes information from patients diagnosed with HCM, Dilated Cardiomyopathy (DCM), and Arrhythmogenic Cardiomyopathy. The registry contains genetic, clinical, and demographic information from over 17,000 of patients, of which 12,326 were diagnosed with HCM. Of these, 4,427 patients have genetic variants within the six genes analyzed in this research, totaling 1,056 unique variants.

3.2.2 Genome Aggregation Database

The Genome Aggregation Database (gnomAD), originally established as the Exome Aggregation Consortium (ExAC), represents a global collaboration among researchers committed to sharing data from exome and genome sequencing projects. Unlike disease-focused databases such as SHaRe, gnomAD compiles data from the general population, providing a valuable resource for studying genetic variation amongst the whole population. In its latest release (v4), gnomAD has expanded its database to include 730,947 exomes. Among these, 586,403 patients have variants in the sarcomeric genes we focused on, amounting to 4,620 unique observed variants.

3.2.3 UK Biobank

The UK Biobank (UKBB) is a comprehensive repository of biomedical data and research materials. It contains anonymized information on genetics, lifestyle habits, and health outcomes, along with biological samples, collected from 500,000 individuals across the United Kingdom. Like the gnomAD dataset, it is curated from the general population and not enriched for people with specific conditions. The effort results in 346,977 patients with variants in the genes of interest for our study, yielding a total of 1,479 unique variants.

3.2.4 AlphaMissense

AlphaMissense is a transformer-based model developed to predict the pathogenicity of missense variants in the human genome (see Section 2.3.1). For each gene, the model creates pathogenicity predictions for the 19 possible missense mutations at each amino acid position. Within the six genes we studied, this results in a total of 77,482 pathogenicity scores. This number significantly exceeds the total number of variants observed in both the general population cohorts and the disease-enriched cohort combined (see Table 3.1).

For the MYH7 gene, which has a length of 1,935 amino acids, we obtained 38,700 predictions. For MYBPC3, with 1,274 amino acids, there were 20,480 predictions. TNNT2, with 298 amino acids, resulted in 5,960 predictions, while TNNI3, with 210 amino acids, yielded 4,200 predictions. Additionally, MYL2, which has 166 amino acids, generated 3,320 predictions, and for MYL3, with 195 amino acids, there were 3,900 predictions.

3.2.5 ClinVar

ClinVar is a public archive that collects and shares information about human genetic variations and their relationships to diseases and drug responses. Researchers, clinicians, and laboratories can submit and access data on genetic variants, including their clinical significance and supporting evidence. We used the ClinVar Miner tool to access data on our genes of interest with data current up to February 5, 2024 [23]. In total, this resulted in 5,252 annotations, of which 2,336 are not categorized as variants of uncertain significance. We used data from ClinVar as an independent, high-quality validation set to further substantiate the validity of the identified hotspot regions.

3.3 Classification of Variants

3.3.1 Binary Classification of Variants

Our initial analysis focuses on identifying disease-enriched regions using a discrete, binomial statistical framework, as outlined in Section 3.4.1. To accomplish this, variants were classified as either disease-associated or from the general population. This section details the categorization strategies employed.

Clinical Data: SHaRe vs. Population Datasets

For the clinical data, we adopted the same categorization strategy as proposed by Homburger et al. and labelled all variants found in HCM patients from the SHaRe cohort as disease-associated [25]. Variants exclusively appearing in one of the two population datasets were considered reference variants. While the gnomAD and UKBB population datasets may include individuals with HCM, the incidence is not expected to exceed that of the general population. Thus, most rare variants observed in these cohorts are likely unrelated to HCM. Conversely, rare variants observed in the SHaRe cohort are more likely to be causative for HCM. Since our analysis aimed to assess disease enrichment within the protein structure, we restricted our focus to missense variants, as these directly alter amino acid sequences and thereby impact the protein's structure and function.

AlphaMissense

AlphaMissense provides pathogenicity predictions as a continuous score ranging from 0 to 1, rather than a binary classification. To add these continuous scores into our discrete statistical framework, we developed a classification strategy that categorizes variants into three groups:

- Disease-associated variants
- Population variants
- Excluded variants

This classification approach is validated in the AlphaMissense paper, which supports flexibility in choosing cutoffs to align with specific use cases [5].

To address the data imbalance caused by the limited number of observed variants in clinical datasets compared to the extensive in silico predictions available for all

possible missense variants in all genes, we implemented a selective inclusion strategy. This approach increased the number of variants included in our analysis by 50%, while keeping the ratio of disease-associated to reference variants constant. Maintaining this ratio is necessary to preserve the validity of the hypothesis tested using the binomial likelihood assessment, as described in Section 3.4.1.

Before integrating AlphaMissense data into our model, we performed a deduplication step to remove any variants already present in the SHaRe cohort or population datasets. This filtering process was essential to prevent the overrepresentation of single variants. If a single variant was both present in the SHaRe cohort and predicted as highly pathogenic by AlphaMissense, it would essentially be double counted and the permutation test would artificially inflate their significance.

We sought to increase the size of our dataset by 50% by employing the following strategy: Let N_{Disease} represent the number of disease-associated variants and $N_{\text{Population}}$ the number of population variants observed in the clinical data. After the deduplication step, we ranked all variants based on their predicted pathogenicity score. From this ranked list, we expanded our set of disease-associated variants by 50% by adding the highest-scoring variants:

$$AddedDiseaseVariants = sortedPredictions[0 : N_{Disease} * 0.5]$$

Similarly, for the list of population variants, we added the lowest-scoring variants to increase the total size of population variants by 50%:

$$AddedPopulationVariants = sortedPredictions[-N_{Population} * 0.5 :]$$

This strategy allowed us to maintain balance while significantly expanding the dataset, thereby enhancing the statistical power of our analysis.

3.3.2 Continuous Classification of Variants

To address the potential limitations inherent in the binary classification of AlphaMissense pathogenicity scores as well as the loss of information due to the selective inclusion criteria which only increases the dataset size by 50%, we developed a complementary analytical approach. This method aims to validate our findings by instead of discretizing the continuous AlphaMissense predictions transforming the clinical data into continuous scores. The discrete statistic assuming a binomial distribution of variants was replaced with a continuous statistic assuming a Gaussian distribution of the

pathogenicity scores, as detailed in Section 3.4.2.

Clinical Data: SHaRe vs. Population Datasets

For each variant at a given amino acid, we calculated a clinical HCM-score by first counting the number of variants within the spherical region observed in the SHaRe cohort, denoted as $n_{HCM_{in}}$, and the number of variants only observed in the population cohorts, denoted $n_{population_{in}}$. We then applied a softmax function to obtain a score between 0 and 1, where a higher score indicates more disease-associated variants present on the respective amino acid.

$$\text{HCM-score} = \frac{e^{n_{HCM_{in}}}}{e^{n_{HCM_{in}}} + e^{n_{population_{in}}}}$$

This approach of calculating a clinical HCM-score by contrasting the number of disease-associated variants from the SHaRe cohort with those found in population datasets results in a scoring pattern that bears similarity to that resulting from AlphaMissense predictions. Before inputting the score into the statistical model, we normalized it to prevent a potential higher variance from disproportionately influencing the subsequent Gaussian likelihood calculation.

AlphaMissense

AlphaMissense already offers pathogenicity predictions as a continuous score that ranges from 0 to 1. To derive a pathogenicity score for each residue, we averaged the predictions across all 19 possible missense variants for each amino acid. In this case as well, we applied normalization after averaging to ensure that potentially higher variance did not disproportionately affect the Gaussian likelihood calculation.

3.4 Statistical Frameworks

3.4.1 Binomial Spatial Scan Statistic

In our binomial spatial scan analysis, we examined the locations of unique disease-associated variants in comparison to variants found in the population datasets, using the structural models of the proteins. The aim was to identify regions where disease-associated variants occur significantly more often than can be expected by chance.

The statistical framework utilizes a sliding-window approach, using spheres with radii of 10Å, 15Å, and 20Å, which iterate across all amino acids of the protein.

For each step of the iteration, the number of disease-associated variants within (y_{in}) and outside (y_{out}) the sphere was calculated, along with the number of population variants within (z_{in}) and outside (z_{out}) the sphere. We then compared the proportions of HCM-associated variants within and outside the sphere using the following definitions:

- $p_{\text{in}} = \frac{y_{\text{in}}}{y_{\text{in}}+z_{\text{in}}}$: The proportion of HCM-associated variants within a given window.
- $p_{\text{out}} = \frac{y_{\text{out}}}{y_{\text{out}}+z_{\text{out}}}$: The proportion of HCM-associated variants outside the window.
- $r = \frac{y_{\text{in}}+y_{\text{out}}}{y_{\text{in}}+z_{\text{in}}+y_{\text{out}}+z_{\text{out}}}$: The overall rate of HCM-associated variants across all windows.

For each window, we compute the binomial log-likelihood ratio statistic. This statistic compares the null model, where p_{in} and p_{out} are both equal to the overall rate r, against an alternative model where p_{in} differs from p_{out}. The log-likelihood ratio statistic for each window is calculated as follows:

$$\log(\Lambda_w) = \log\left[p_{\text{in}}^{y_{\text{in}}}(1-p_{\text{in}})^{z_{\text{in}}}p_{\text{out}}^{y_{\text{out}}}(1-p_{\text{out}})^{z_{\text{out}}}\right] - \log\left[r^{y_{\text{in}}+y_{\text{out}}}(1-r)^{z_{\text{in}}+z_{\text{out}}}\right]$$

For each residue of the protein model, the likelihood ratio was calculated. We then assessed the statistical significance of the identified hotspot regions using a permutation test with 1,000 permutations. During each permutation, the variant labels were shuffled randomly, and the number of variants per residue was held constant. This ensured that the significance assessment remained unaffected by any inherent non-uniform distribution of variants across the full protein.

3.4.2 Gaussian Spatial Scan Statistic

As in the binomial spatial scan analysis, we used a sliding-window approach with spheres of sizes 10Å, 15Å, and 20Å, iterating over all residues of the given protein. However, the binomial setting was replaced by assuming a gaussian distribution. For each location, we created two vectors: HCM_{in}, containing the HCM-scores of all amino acids within the sphere, and HCM_{out}, containing the HCM-scores of all amino acids outside the sphere.

We then defined μ_{in} as the mean of the HCM_{in} vector, and μ_{out} as the mean of the HCM_{out} vector. For each window, we compute the log-likelihood ratio statistic. This statistic compares the null model, where μ_{in} and μ_{out} are both equal to the overall mean

μ, against an alternative model where μ_{in} differs from μ_{out}. The log-likelihood ratio statistic ($\log(\Lambda_w)$) for each window is calculated as follows:

$$\log(\Lambda_w) = (\log(\text{LL}_{\text{in}}) + \log(\text{LL}_{\text{out}})) - \log(\text{LL}_{\text{all}})$$

where LL_{in} is the log-likelihood of the Gaussian distribution for scores within the window, LL_{out} is the log-likelihood for scores outside the window, and LL_{all} is the log-likelihood assuming all scores come from the same Gaussian distribution:

$$\text{LL} = -\frac{1}{2}\left(\sum_{i=1}^{n}\log(2\pi\sigma^2) + \frac{(x_i-\mu)^2}{\sigma^2}\right)$$

For each residue of the protein models, this log-likelihood ratio was calculated. We then evaluated their statistical significance in descending order using a permutation test with 1,000 permutations, where HCM-scores were randomly shuffled across all amino acids.

This Gaussian spatial scan statistic allows us to identify regions of the protein where the HCM-associated scores are significantly elevated, potentially indicating functionally important areas more prone to disease-causing mutations. By incorporating multiple sources of evidence into a composite score and using a Gaussian distribution, we aimed to capture more nuanced patterns of variant effects compared to the discrete binomial approach.

3.5 Validation Strategies

3.5.1 Validation with ClinVar

Integrating cohort-based comparisons from the SHaRe dataset and two population datasets with in silico data from AlphaMissense significantly expanded our dataset, enhancing the statistical power of our analysis. However, this gain in statistical power came at the cost of interpretability, as the statistical analysis arises from the combination of two data sources.

To address this, we performed a qualitative assessment of ClinVar classifications within and outside the identified hotspots. By examining histograms, we analyzed the distribution of pathogenic and benign variants across these regions, enabling us to validate our findings using a robust, high-quality holdout dataset. The histograms supporting this assessment are provided in the supplementary chapter of this thesis (see Section A.1).

For a more quantitative analysis, we compared the proportions of pathogenic and

likely pathogenic variants against benign and likely benign variants, both within and outside the hotspot regions. To evaluate the statistical significance of these differences, we conducted a Fisher's exact test. This method helped assess the strength of our findings and provided additional evidence to support or challenge the patterns detected by the scan statistic framework in relation to established medical literature.

3.5.2 Validation with Disease Onset Age Analysis

The clinical manifestation of HCM exhibits considerable variability, with some patients experiencing minimal symptoms and having a normal lifespan, while others may face sudden death or require cardiac transplantation at a young age. The age at which HCM presents also varies significantly among patients, with earlier onset generally associated with a more severe phenotype [41]. Consequently, we aimed to compare the age of disease onset in patients with variants located within the identified hotspot regions to those with variants outside the sphere. To determine the statistical significance of any observed differences, we employed a Wilcoxon rank-sum test and analyzed Kaplan-Meier curves for disease onset. Both the Kaplan-Meier curves and scatter plots of disease onset per category, are provided in the supplementary material (see Section A.1).

Chapter 4

Results

In this chapter, we first evaluate the protein structures generated using the four structure prediction models. Next, we present the identified hotspots, followed by an ablation study that compares the two mathematical frameworks and examines the impact of both data sources, AlphaMissense and clinical data, on the detection of disease-enriched regions.

4.1 Evaluation and Selection of Protein Models

In this section, we present the results of our predicted protein structures and provide the rationale behind the selection of the structure used for subsequent statistical analysis. The evaluation was based on an assessment of predicted metrics, including the averaged pLDDT and PAE scores, alongside a detailed examination of the PAE plot to find patterns of high or low confidence. A major emphasis was placed on the alignment errors when the predicted models were aligned with partial protein structures obtained through cryo-EM.

4.1.1 MYH7

The AlphaFold 2 structure was selected as the best structural model for MYH7, with a superior performance across multiple evaluation metrics (see Figure 4.1). With a mean pLDDT score of 73.75 and a mean PAE of 23.43 Å, the AlphaFold 2 model demonstrated the lowest overall PAE and the second-highest pLDDT among the evaluated models.

All models consistently showed higher accuracy in predicting the structure of subfragment 1 (S1) of the heavy meromyosin (HMM) domain, which encompasses the

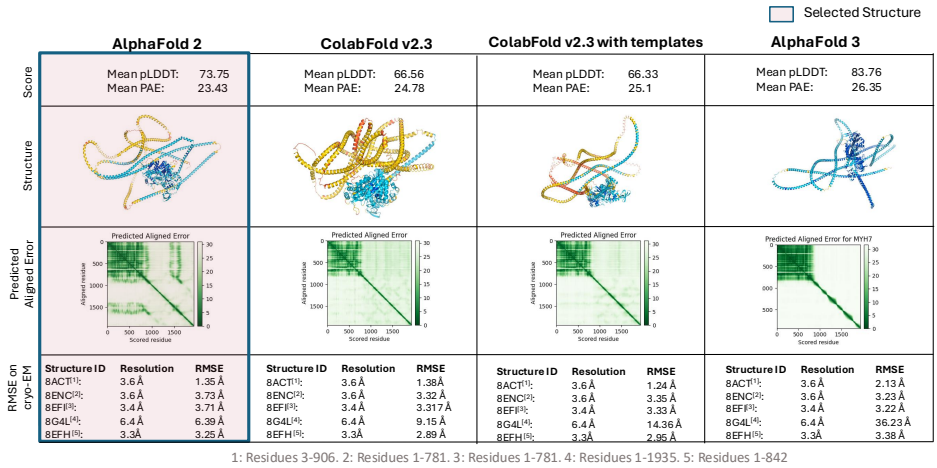

Figure 4.1: Comparison of MYH7 models using AlphaFold 2, ColabFold, and AlphaFold 3. AlphaFold 2, with the second-highest pLDDT and lowest PAE, performed best overall. All models show higher accuracy for the S1 domain and lower accuracy for the S2 and LMM domains.

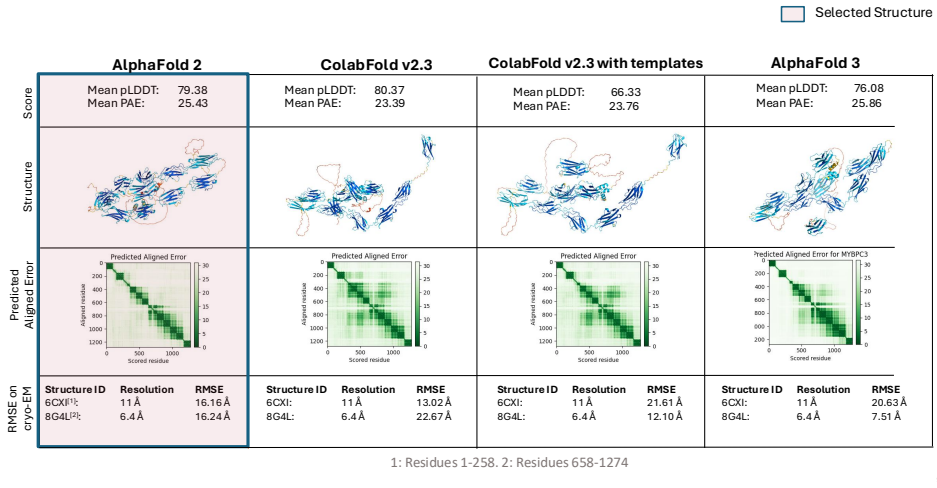

Figure 4.2: Comparison of MYBPC3 structure predictions. AlphaFold 2, with a pLDDT of 79.38 and PAE of 25.43, was selected as the best model, achieving consistently low RMSE. All models demonstrate high local confidence, but show lower confidence in domain-domain distances, as indicated by the PAE plots.

myosin head (residues 1-840). In contrast, the subfragment 2 (S2) (residues 841-1280) and the light meromyosin (LMM) domain (residues 1281-1935) exhibited lower confidence, with higher predicted errors in these regions. This result likely arises because the S2 and LMM domains are more influenced by interactions with other proteins in the thick filament, leading to greater variability and less structural definition in these areas.

When aligned with experimentally determined cryo-EM structures, the AlphaFold 2 model consistently produced the lowest root mean square error (RMSE) across three out of the four tested structures. It achieved an RMSE of 1.35 Å for the structure with PDB ID 8ACT [21], an RMSE of 3.73 Å for the structure with PDB ID 8ENC [8], an RMSE of 3.71 Å for the structure with PDB ID 8EFI [8], an RMSE of 6.39 Å for the structure with PDB ID 8G4L [11] and an RMSE of 3.25 Å for the structure with PDB ID 8EFH [**?**]. With an average RMSE of 3.67 Å, this model achieved the lowest overall error across the evaluated structures.

4.1.2 MYBPC3

In modeling the structure of MYBPC3, all models exhibited similar mean PAE scores, indicating comparable overall accuracy in predicting residue-residue distances across the protein (see Figure 4.2. Additionally, all models, except for ColabFold v2.3 with templates, achieved similar pLDDT scores close to 80, demonstrating high confidence in local structural predictions. The AlphaFold 2 structure, with a mean pLDDT of 79.38 and a mean PAE of 25.43, was ultimately chosen as the best model due to its consistently good performance in terms of RMSE when aligned with experimentally determined cryo-EM structures. All structures except for the AlphaFold 2 structure have an RMSE of over 20 Å for one of the experimental structures.

A detailed inspection of the PAE plots across all models reveals that while the confidence in long-distance relationships between residues far apart in the sequence is low, the local accuracy for individual domains of MYBPC3 is high. This pattern suggests that, although the models may struggle with accurately predicting the overall topology and domain-domain interactions, they reliably capture the secondary structures and the shape of the individual domains.

4.1.3 TNNT2

In the evaluation of structural models for TNNT2, all models showed comparable PAE scores around 22 Å (see Figure 4.3). The pLDDT scores for the models predicted using the AlphaFold 2 architecture (79.75, 78.24, and 78.22) were higher compared to the AlphaFold 3 model, which achieved a pLDDT of 73.32. Yet, the AlphaFold 3 structure was ultimately selected as the best model due to its superior performance in terms of RMSE when aligned with the experimentally determined cryo-EM structure. It achieved an RMSE of 2.49 Å for the structure with PDB ID 7UTI [48], beating the second-closest structure created with ColabFold v2.3 by 0.5 Å. Also, the PAE plots across all models highlight that the models reliably capture the folding and structural integrity of the three coil segments (residues 1-50, residue 100-180 and residues 220-280) while displaying a lower confidence in predicting the relationships between these domains.

4.1.4 TNNI3

The ColabFold v2.3 model with templates achieves the best evaluation criteria in the structural modeling of TNNI3 (see Figure 4.4). The model achieved a pLDDT score of 81.19 and a PAE of 17.6, both of which are superior to those of all other predicted structures. The outstanding performance is also reflected in achieving the lowest

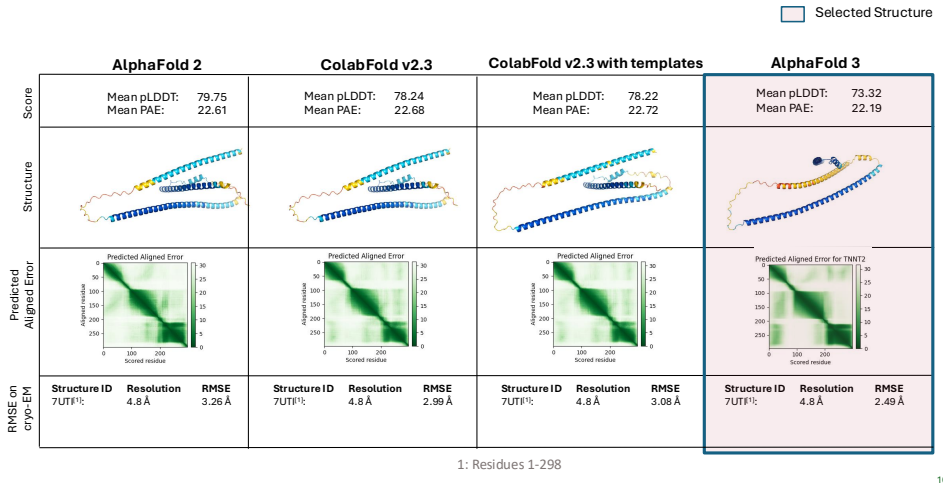

Figure 4.3: Comparison of TNNT2 structure predictions. Despite higher pLDDT scores for AlphaFold 2 models, AlphaFold 3 was selected as the best model, achieving the lowest RMSE (2.49 Å) when aligned to a cryo-EM structure and achieving the lowest PAE.

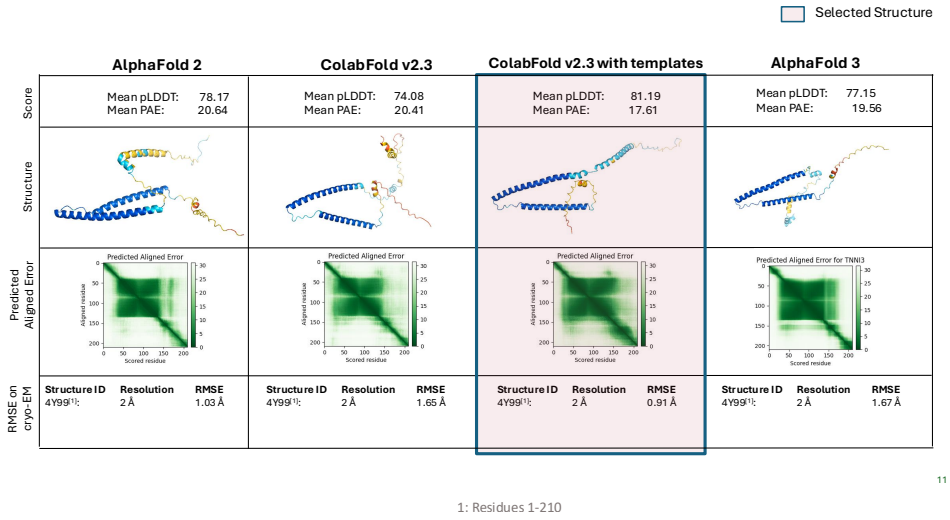

Figure 4.4: Comparison of TNNI3 structure predictions. ColabFold v2.3 with templates was selected as best, achieving the highest pLDDT (81.19), lowest PAE (17.6), and the lowest RMSE (0.91 Å) in cryo-EM alignments, with high confidence in residue-residue distances between amino acids 20 and 150.

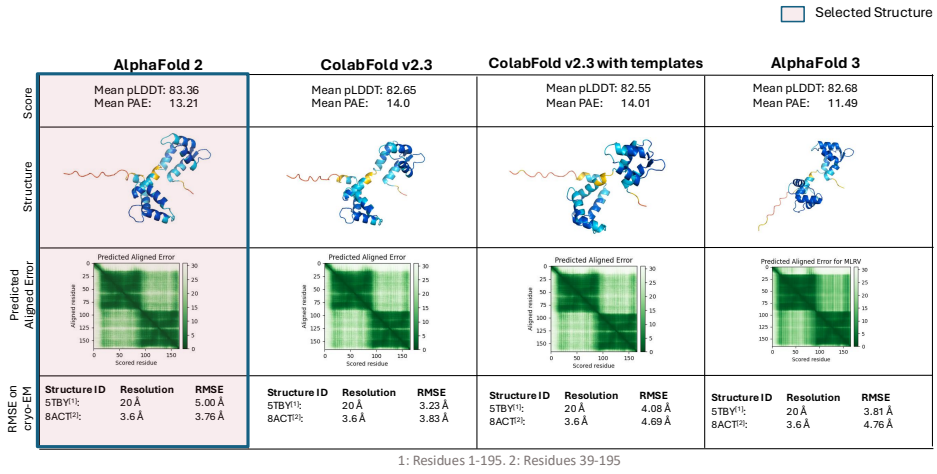

Figure 4.5: Comparison of MYL2 structure predictions. AlphaFold 2 was selected as most accurate with the highest pLDDT (83.36) and the second-best PAE (13.21 Å), and with the lowest RMSE in cryo-EM alignments. All models show low confidence in modeling the first 20 amino acids and high confidence for the rest of the structure.

RMSE when aligned with the experimentally determined cryo-EM structure, achieving an RMSE of 0.91 when aligned to the structure with PDB ID 4Y99 [52]. An inspection of the PAE plot reveals an especially high confidence of residue-residue distances between amino acids 20 and 150, where both the TNC-binding and actin-binding sites are.

4.1.5 MYL2

For the gene MYL2, we selected the model created with AlphaFold 2 as the best structure (see Figure 4.5). It achieved the best mean pLDDT of 83.36 and the second-best mean PAE with 13.21 Å, only surpassed by the structure created with AlphaFold 3 which achieved a mean PAE of 11.49 Å. The decision to use the AlphaFold 2 model was driven by the excellent RMSE when the structure was aligned to the cryo-EM structure with PDB ID 8ACT [21], which itself has a high resolution of 3.6Å. An inspection of the PAE plots reveals that all structure prediction models are confident in the aligned distances in the domain of the EF-hands: the first EF-hand between residue 24 to 59, and the second and third EF-hands 3 between residues 94-129 and 130-165.

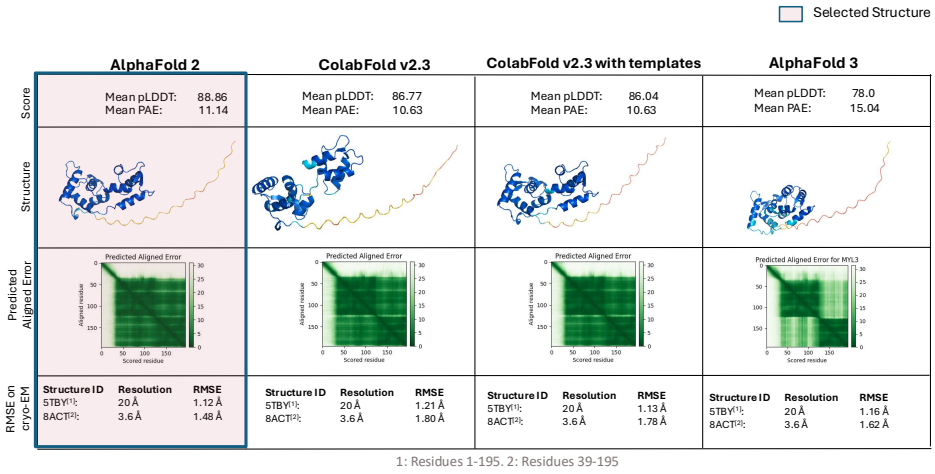

Figure 4.6: Comparison of MYL3 models, with AlphaFold 2 selected as the best. It achieved a high pLDDT (88.86) and strong cryo-EM alignment, with RMSE values of 1.48 Å (PDB ID 8ACT [21]) and 1.12 Å (PDB ID 5TBY [2]). High confidence was observed in folding predictions, except for the disordered N-terminal region.

4.1.6 MYL3

Upon evaluating the structural predictions of MYL3, the AlphaFold 2 model was identified as the most reliable, achieving a high mean pLDDT of 88.86 and a mean PAE of 11.14 Å, which, while third lowest, is comparable to the lowest PAE of 10.63 (see Figure 4.6).

This model exhibited exceptional alignment with the cryo-EM structures used for comparison. When aligned with the structure having PDB ID 8ACT [21], the model achieved an RMSE of 1.48 Å, and with the structure having PDB ID 5TBY [2], it achieved an RMSE of 1.12 Å. The PAE plots indicate that, apart from the disordered N-terminal region (amino acids 1-37), the predicted folding confidence is very high.

4.2 Hotspot Identification for Disease-Associated Variants

In this section, the identified hotspot regions enriched in disease-associated variants are presented, together with the respective validation using ClinVar data and an analysis of the observed age of disease onset.

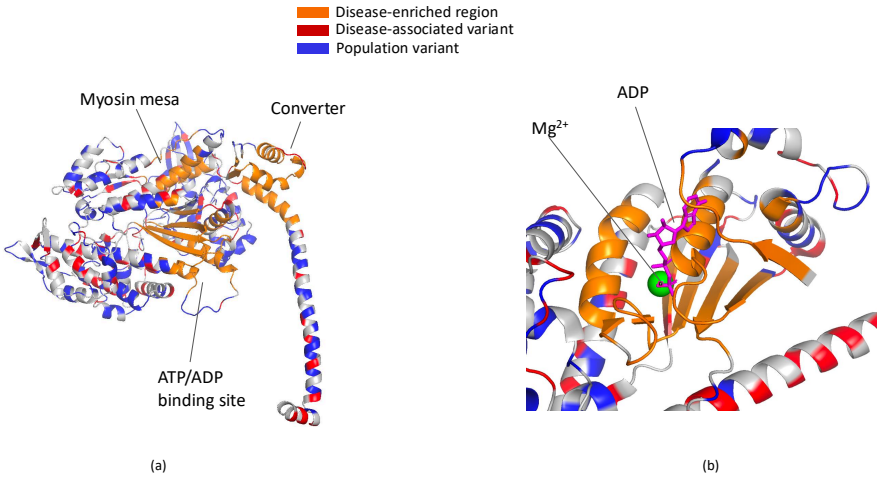

Figure 4.7: Structural model of MYH7 S1 domain highlighting regions enriched with disease-associated variants. (a) The myosin mesa, converter, and ATP/ADP binding site are shown in orange, marking disease-enriched regions, with red indicating disease-associated variants and blue representing population variants. (b) A closer view of the ATP/ADP binding site, with ADP and Mg^{2+} bound, illustrates the hotspot near the nucleotide pocket [9].

4.2.1 MYH7

Our statistical analysis of the S1 domain of the beta myosin heavy chain identified three domains enriched for disease-associated variants: the converter domain, the myosin mesa, and the ATP/ADP binding site (see Figure 4.7).

We first identified a region for disease-enrichment on the converter domain, centered on residue 724 with a 15 Å sphere ($p = 0.004$). Patients with variants in this area have a 10.5 years earlier disease onset, indicating a more severe phenotype ($p = 2.8 \times 10^{-7}$). This domain also holds significantly more variants annotated as pathogenic or likely pathogenic (P/LP) in ClinVar ($p = 0.024$). The converter is essential for the mechanical transduction of the energy released during ATP hydrolysis into the movement of the myosin lever arm. Pathogenic variants in the Converter domain may alter the efficiency of this process, leading to an abnormal contractile function.

Second, two adjacent hotspots located on the myosin mesa were identified, the first with a 10 Å sphere centered on residue 493 ($p = 0.001$) and the second one with a 10 Å sphere centered on residue 696 ($p = 0.001$). Notably, people with variants in the

second sphere seem to have a later disease onset (8.4 years later, $p = 0.016$). Both locations contain significantly more P/LP annotations in ClinVar than can be expected by chance ($p = 0.042$ and 0.041, respectively). The myosin mesa is a relatively flat surface on MYH7, and variants in this region might disturb interactions with the proximal S2 domain.

Third, we identified two spherical regions around the ATP/ADP binding site. The first is directly on the site centered on residue 187 with a 10 Å sphere ($p = 0.007$), as shown in Figure 4.7 (b), which highlights the hotspot's proximity to the binding site (visualization created using cryo-EM model with PDB ID 8EFE [9]). Notably, the hotspot area is not reflected when inspecting the annotations in ClinVar (see Appendix A.1). The region identified near the binding site is located on residue 257, encapsulating the region with a 15 Å sphere ($p = 0.001$). Although variants at the ATP/ADP binding site itself do not significantly alter the age of disease onset, patients with variants in the nearby hotspot experience a significantly earlier disease onset by 8.1 years earlier ($p = 0.010$). Furthermore, the region harbors significantly more P/LP annotations in ClinVar ($p = 0.004$).

4.2.2 MYBPC3

On MYBPC3, we identified four hotspots, one located in the C6 domain with a 15 Å sphere centered on residue 857 ($p = 0.007$), one in the C8 domain with a 10 Å sphere centered on residue 1030 ($p = 0.001$), one in the C9 domain with a 10 Å sphere centered on residue 1102 ($p = 0.001$) and finally one in the C10 domain with a 10 Å sphere centered on residue 1253 ($p = 0.008$) (see Figure 4.8). All hotspots do not show a statistically significant change in disease onset. Although all identified hotspots display a higher-than-expected proportion of P/LP annotations, only the hotspot within the C6 domain demonstrates a statistically significant enrichment of P/LP annotations ($p = 0.003$).

Upon examining the disease hotspots illustrated in Figure 4.8, it is apparent that the hotspots are localized predominantly within regions where the secondary protein structure consists of beta-sheets. A t-test supports the visual observation, indicating that the clustering of disease hotspots on beta-sheets is highly unlikely to occur by chance ($p = 1 \times 10^{-6}$). This result is further validated by comparing the P/LP and benign/likely benign (B/LB) annotations from ClinVar ($p = 0.049$), corroborating the finding that pathogenic variants cluster on beta-sheets.

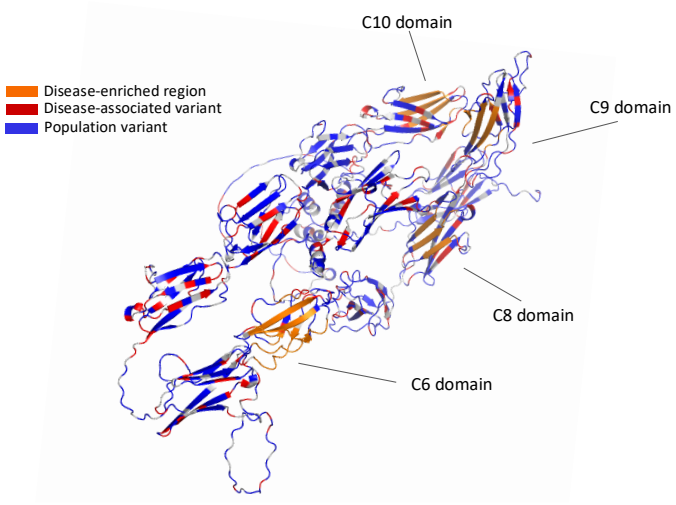

Figure 4.8: Structural model of MYBPC3 highlighting disease-enriched regions in the C6, C8, C9, and C10 domains. Hotspots are shown in orange, with red indicating disease-associated variants and blue representing population variants. Statistical analysis confirmed significant clustering of disease-causing variants on beta-sheets, indicating a structural preference for variant accumulation in these regions.

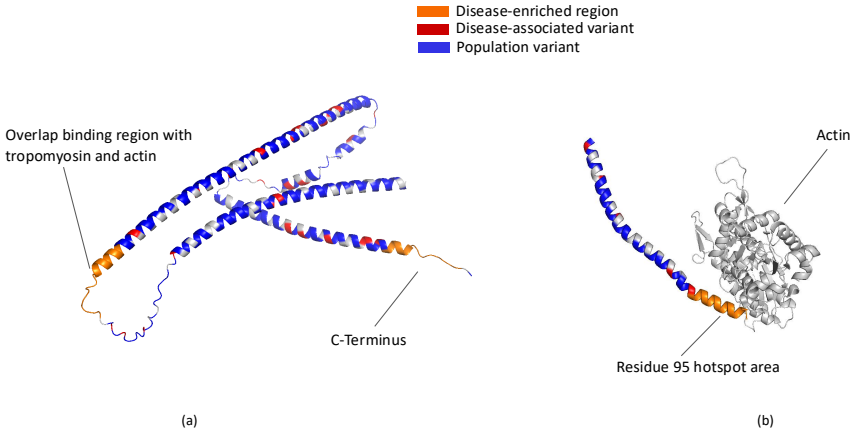

Figure 4.9: Structural model of TNNT2 highlighting disease-enriched regions. (a) The C-terminus and overlap binding region with tropomyosin and actin show hotspots of disease-associated variants. (b) Close-up of the residue 95 hotspot, where TNNT2 interacts with actin [48].

4.2.3 TNNT2

For cardiac troponin T, we identified two regions enriched with disease-associated variants (see Figure 4.9).

The first region is located at the edge of the domain where TNNT2 interacts with tropomyosin and actin, defined by a 15 Å sphere centered on residue 95 ($p = 0.001$). The interaction of TNNT2 with actin based on the cryo-EM structure from PDB ID 7UTL [48] is shown in Figure 4.9 (b). The proximity of the hotspot region, colored in orange, suggests variants in this location might disturb the interaction. Variants in this region are associated with an earlier disease onset, averaging 10.2 years earlier ($p = 0.003$). This finding is further supported by a higher proportion of P/LP annotations in ClinVar ($p = 0.001$).

The second region is situated at the C-terminus of the gene, with a 15 Å sphere centered on residue 290 ($p = 0.002$). Variants in this region are linked to a later disease onset, averaging 8.5 years later ($p = 0.005$), with no significant increase in the proportion of P/LP annotations in ClinVar.

4.2.4 TNNI3

Upon examining cardiac troponin I, we identified three hotspots across adjacent domains (see Figure 4.10).

The first hotspot is located within the switch domain, defined by a 10 Å sphere centered on residue 141 ($p = 0.001$). This hotspot is situated at the TNNI3-actin interaction site, as shown in Figure 4.10 (b), based on the cryo-EM structure from PDB ID 7UTI [48]. Additionally, there is a significantly higher-than-expected number of P/LP annotations in ClinVar ($p = 0.026$).

The second hotspot is defined by a 10 Å sphere centered on residue 174, where the switch domain connects to the mobile domain ($p = 0.001$). Also, in ClinVar the proportion of P/LP variants vs. B/LB is elevated, but not statistically significant ($p = 0.06$).

The third hotspot is located within the mobile domain, characterized by a 15 Å sphere centered on residue 194 ($p = 0.001$). Variants in this region are associated with an earlier disease onset, averaging 9.6 years earlier ($p = 0.006$), and the region also shows a significantly higher proportion of P/LP annotations in ClinVar ($p = 0.001$).

4.2.5 MYL2

The analysis of MYL2 identified a disease-enriched region characterized by a 20 Å sphere centered on residue 98 ($p = 0.005$; see Figure 4.11 (a)). Variants in this re-

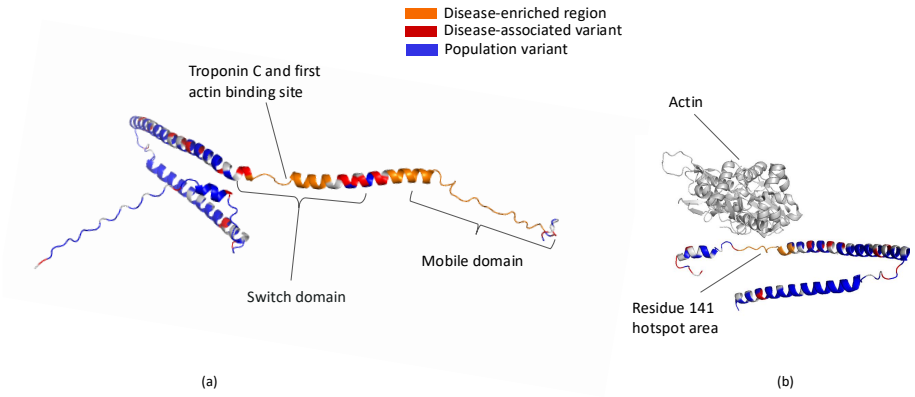

Figure 4.10: Structural model of TNNI3 highlighting disease-enriched regions. (a) Three hotspots are identified across adjacent domains: the switch domain, the mobile domain, and the Troponin C/first actin binding site. (b) Close-up of the residue 141 hotspot region, where TNNI3 interacts with actin [48].

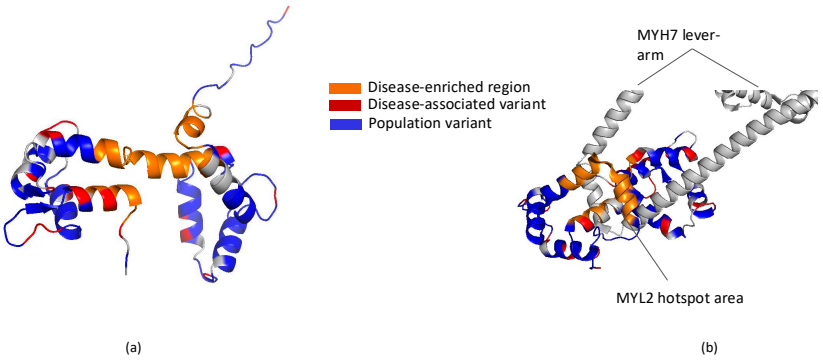

Figure 4.11: Structural model of MYL2 highlighting the disease-enriched region. (a) A hotspot centered on residue 98 is associated with significantly later disease onset. (b) Close-up of the MYL2 hotspot region, showing its proximity to the MYH7 lever arm, suggesting that variants in this region may affect the interaction between MYL2 and MYH7 [11].

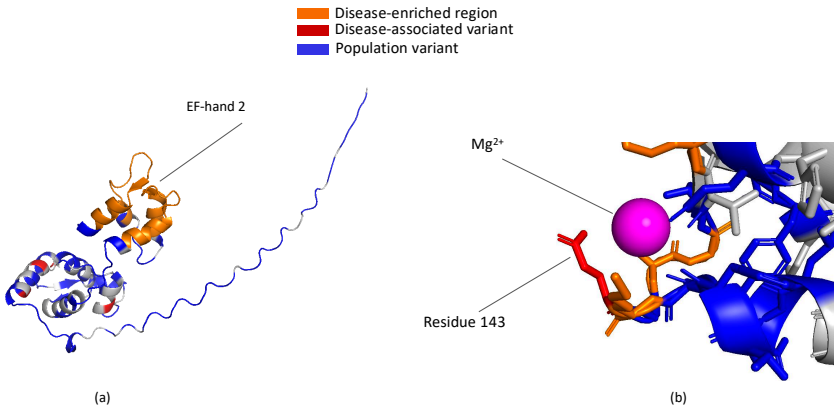

Figure 4.12: Structural model of MYL3 highlighting the disease-enriched region. (a) The hotspot is identified on the second EF-hand domain, centered on residues 182. (b) Close-up of the Ca^{2+} interaction site, showing the proximity of residue 143 to the Calcium ion. This variant might have its pathogenic effect by disrupting MYL3's interaction with Ca^{2+}.

gion are associated with a significantly later disease onset, averaging 12.8 years later ($p = 0.005$). Furthermore, this enrichment is not reflected in the P/LP annotations in ClinVar.

Inspection of the cryo-EM structure (PDB ID 8G4L [11]) reveals that this region surrounds the MYH7 lever arm and may exert its pathogenic effects by disrupting the interaction (see Figure 4.11 (b)).

4.2.6 MYL3

The examination of MYL3 revealed one region enriched with disease-associated variants on the second EF-hand of the protein defined by a 15 Å sphere centered on residue 182 ($p = 0.001$; see Figure 4.12 (a)). This region also shows a significantly higher number of P/LP annotations in ClinVar than would be expected by chance ($p = 0.006$). This region corresponds to the functional Ca^{2+} binding domain of MYL3. However, past research indicates that despite having the functional Ca^{2+} binding motif, MYL3 evolutionarily lost its ability to bind calcium [46][58].

Within the enriched sphere, 39% of patients carry the p.Glu143Lys variant. A

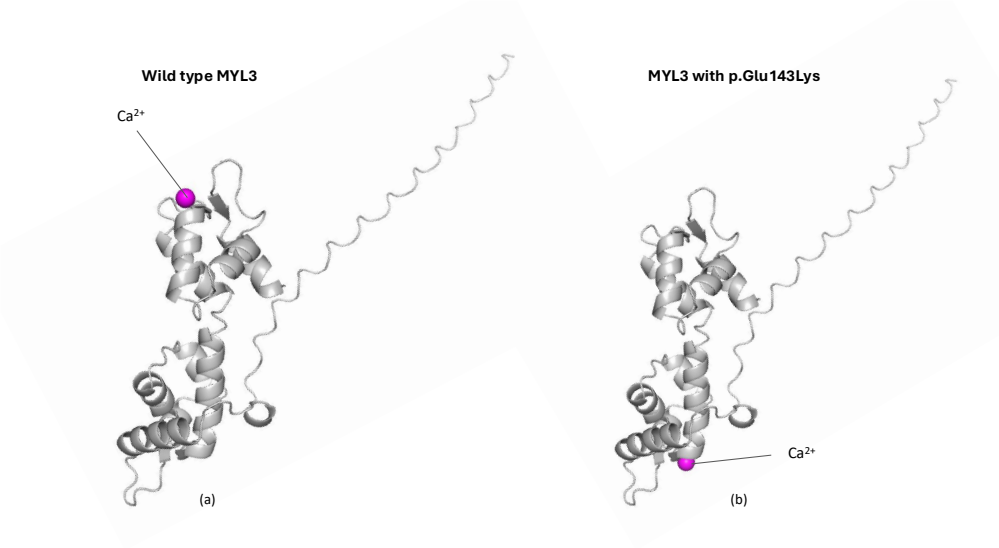

Figure 4.13: Comparison of the interaction of MYL3 with Ca^{2+}. For the wild-type structure, AlphaFold 3 simulates an interaction near residue 143 (see (a)). For the structure with variant p.Glu143Lys, the interaction is moved (see (b)).

protein-ion interaction simulation using AlphaFold 3 demonstrated that residue 143 is in close proximity to the defunct ion binding site (see Figure 4.12 (b)). Introducing the p.Glu143Lys variant into the amino acid sequence and re-running the simulation showed a displacement of the binding from its original position (see Figure 4.13). This evidence supports the hypothesis that substituting the negatively charged glutamic acid with the positively charged lysine influences the affinity of MYL3 to Ca^{2+} and thereby causing pathogenic effects.

4.3 Ablation of Mathematical Frameworks and Data Modalities

4.3.1 Binomial vs. Gaussian Statistical Framework

The binomial framework identified 16 hotspots with $p < 0.05$ and 15 with significance levels of $p < 0.01$ (see Table 4.1). In contrast, the Gaussian setting only detected twelve hotspots with $p < 0.05$ and 9 with $p < 0.01$. Among the common hotspots, the discrete statistic showed greater significance in six regions, whereas the Gaussian method was more significant in two areas. Three hotspots exhibited an equivalent level of significance in both frameworks. These results indicate that the discrete statistic tends to identify clinically relevant hotspots more reliably, with stronger associations evident in a greater number of regions compared to the Gaussian approach.

4.3.2 Effects of AlphaMissense Scores and Clinical Data on Binomial Statistic

The impact of integrating both data sources for the binomial statistic was highly notable. Using only AlphaMissense data, twelve out of 16 hotspots achieved a significance level of $p < 0.01$, and the same number achieved $p < 0.05$. In contrast, clinical data alone produced fewer significant results, with 10 out of 16 achieving $p < 0.05$ and only six out of 16 hotspots achieving $p < 0.01$. The combination of both data sources led to improved p-values in six locations compared to AlphaMissense alone, while it produced less significant p-values in two locations, including MYL2 and the TNNT2 C-terminus. When compared to clinical data alone, the combination showed superior p-values in 13 out of 16 locations, equal p-values in two, and only one instance of an inferior p-value in the converter domain of MYH7.

4.3.3 Effects of AlphaMissense Scores and Clinical Data on Gaussian Statistic

The Gaussian statistical framework identified twelve out of 16 hotspots with $p < 0.05$, and nine of these with $p < 0.01$. When using AlphaMissense predictions alone, all twelve hotspots maintained significance at $p < 0.05$, with one hotspot on MYBPC3 exhibiting a strongly increased significance, resulting in ten out of twelve achieving $p < 0.01$.

In contrast, clinical data alone yielded fewer significant hotspots. Only seven out of twelve reached $p < 0.05$ and four out of twelve reached $p < 0.01$. The combined analysis with both data sources resulted in a trade-off, with AlphaMissense alone being more significant than the combination in six hotspots, less significant in two, and equally significant in four. The clinical data alone did not surpass the combined ap-

proach in any hotspot and was less significant in ten regions, with two hotspots showing equal significance.

Chapter 5

Discussion

This chapter begins by discussing and comparing the performance of the protein structure prediction models used in this study. We then examine the disease hotspots identified, contextualizing them within existing research. Lastly, we discuss the choice of mathematical frameworks and the influence of different data modalities.

5.1 Performance of Structure Prediction Models

The overall performance of the predicted models was highly comparable, with all predictions achieving similar results across the metrics used. Although local prediction confidence was generally very high, as indicated by the PAE scores, the models often struggled to accurately predict distances between different domains, particularly in more flexible or interaction-dependent regions. This indicates that, while current methods excel at capturing local structural accuracy in the sarcomeric genes modeled in this study, predicting long-distance residue interactions remains a challenge.

Among the models evaluated, AlphaFold 2 was selected as the best option for four out of six proteins, while ColabFold with templates and AlphaFold 3 were each chosen for one protein. ColabFold without templates was not selected in any case.

5.2 Discussion of identified Hotspot Regions

5.2.1 MYH7

Our analysis confirmed the previously reported converter domain hotspot [44][25][6]. We found that patients with variants in this region were diagnosed with HCM earlier

than those with variants in other regions, consistent with the findings of previous studies [25][18]. Homburger et al. demonstrated that this region plays a crucial role in force transduction, as the lever arm undergoes an approximately 70° swing from its pre-stroke position during systolic contraction. It has been hypothesized that mutations in the converter domain and proximal S2 tail might cause hypercontractility by disrupting inter- or intra-molecular interactions for maintaining the folded-back state, thereby shifting the balance towards more myosin heads in an "active state", increasing the number of myosin heads available for interaction with actin [28].

Consistent with previous reports, we observed disease enrichment within the myosin mesa region [55][25]. However, in contrast to the findings by Homburger et al., we not only identified a surface area but a spherical region enriched for disease-associated variants [25]. Variants on the mesa potentially disrupt the binding electrostatic interaction between the mesa and the proximal S2. The reduced binding affinity to the proximal S2 could increase the number of accessible myosin heads and thereby cause the hypercontractility [45]. Furthermore, the surface area is at a steep angle to the actin-binding domain and variants might alter the interaction of both proteins [60].

We identified two previously unnoticed hotspots near or exactly on the ATP/ADP-binding site. Interestingly, only variants in the region near the binding site were associated with an earlier onset of HCM but not the variants located on the ATP/ADP-binding site. In the region adjacent to the binding site, the two variants p.Arg453Cys and p.Arg249Gln, together responsible for 46% of observed patients with variants in the sphere, have been studied in detail. Bloemink et al. found that the p.Arg453Cys mutation causes a 35% decrease in ATP binding rate and a threefold slowdown in the ATP hydrolysis step/recovery stroke [3]. Sata et al. suggested that charge changes caused by the p.Arg249Gln mutation may result in a negative influence on the overall myosin ATPase, given the residue's proximity to the ATP-binding pocket [54]. Furthermore, in-vitro studies have shown that both the p.Arg249Gln and the p.Arg453Cys mutation result in decreased actin translocating activity with decreased intrinsic ATPase activity.

5.2.2 MYBPC3

The investigation of myosin-binding protein C identified four regions enriched for disease-associated variants within the C6, C8, C9, and C10 domains. Among these, the C10 domain is consistently highlighted as a potential hotspot for pathogenic variants, while the significance of the other regions is less certain [22][62][63]. Notably, some studies have found no evidence of variant clustering in any domain [27][31].

Waring et al. have reported clusters of pathogenic variants in the C1, C3, C7,

and C10 domains [63], while Helms et al. observed a concentration of non-truncating variants in the C3, C6, and C10 domains [22]. Similarly, Walsh et al. have found variant clustering restricted to the C3 and C10 domains [62]. However, other studies have failed to identify any domains with significant pathogenic variant enrichment [31][27].

Our discovery that both the identified disease hotspots and individual variants classified as pathogenic/likely pathogenic (P/LP) in ClinVar are disproportionately located on beta-sheets represents a novel insight into the structural basis of variant pathogenicity. One possible explanation for this clustering on beta-sheets could be the structural role they play in protein stability. Beta-sheets often form the core of protein structures, and disruptions in these regions may more readily lead to misfolding or destabilization, thereby triggering pathogenic effects [30][35]. Another hypothesis is that beta-sheets may be involved in important protein-protein interactions, and variants in these regions could disrupt essential functional interfaces. Given the potential implications of this finding, further research is needed to investigate whether the clustering of pathogenic variants in beta-sheets MYBPC3 could be linked to disruptions in its structural integrity or functional binding interfaces.

5.2.3 TNNT2

Upon examining TNNT2, we identified two hotspots: one at the interaction site with actin and tropomyosin, and another at the gene's C-terminus.

The primary factors underlying the first identified hotspot at the interaction site are likely related to alterations in the binding affinity of TNNT2 to tropomyosin, as well as disruptions in calcium sensitivity. These changes could compromise the functional interaction between TNNT2 and tropomyosin, leading to aberrant calcium regulation, which is critical for proper muscle contraction [17][24]. Madan et al. hypothesized that mutations that reduce the binding affinity of TNNT2 to tropomyosin might cause HCM by disrupting tropomyosin's inhibitory position along the actin filament, effectively increasing the exposed myosin binding sites at rest [36].

The biological significance of the disease hotspot at the C-terminus lies in its interaction with other troponin subunits and its binding to tropomyosin. This region has been shown to be crucial for maintaining the structural stability of the inactive state [15].

5.2.4 TNNI3

The spatial scan analysis of cardiac troponin I revealed three hotspots adjacent to each other on the switch domain and the mobile domain at the C-terminus.

The first region, located on TNNI3's C-terminus region, encapsulates a disordered region of the protein. In vitro motility assay studies have shown that the deletion of the last 17 residues was associated with myocardial stunning and increased Ca^{2+} sensitivity [14].

The second enriched region was found on the alpha helix connecting the switch domain and the mobile domain. This region has been shown to be essential for the interaction with actin and stabilizing the off-state of regulatory units at low levels of Ca^{2+}[42].

The third region of disease-enrichment was located on the switch domain. Molecular dynamics simulations found that the p.Arg145Trp variant, responsible for 48% of patients within the enriched area, led to a reduction in interaction between this residue and cardiac troponin-C [33]. The electron microscopy image derived from the PDB with ID 7UTI [48] also shows the proximity of this hotspot to the actin filament. Variants in this region might therefore also disturb the interaction of TNNI3 with actin.

5.2.5 MYL2

The inspection of variants on the regulatory light chain MYL2 revealed a previously unnoticed disease hotspot on the interaction site with the MYH7 lever arm. This region encapsulates the phosphorylation site on residue 15 and is in close proximity to the Ca^{2+} binding site located between residues 37-48 [10]. Previous studies focused on the p.Glu22Lys variant, responsible for 59% of observed patients in this sphere, have shown that this mutation decreases the Ca^{2+} binding affinity 17-fold, underscoring the functional importance of this region [57]. Two other variants contained in the sphere, p.Phe18Leu and p.Pro95Ala, have also been shown to have a 3-fold decrease in calcium binding affinity [57]. These findings emphasize the critical role of calcium binding in the regulatory light chain's function and highlight the potential for further exploration of this region in understanding HCM pathogenesis.

5.2.6 MYL3

Lastly, the investigation of variants found in MYL3 resulted in the detection of one hotspot located on the second EF-hand, centered on residue 182.

EF-hand motifs act as calcium sensors, undergoing conformational changes upon calcium binding, which are essential for proper muscle contraction, force generation,

and cardiac function [47]. Several mutations in this region have been shown to affect protein function and stability. The variant p.Glu143Lys, responsible for 39% of observed patients in this region, has been shown to significantly weaken the binding of MYL3 to the myosin lever arm and an increased ability to generate force [34][66]. Additionally, the p.Arg154His mutation significantly weakens the binding to the myosin lever arm, exhibiting a 3-fold higher KD in protein-binding experiments [34]. In experiments conducted with mice, p.Glu143Lys has been associated with increased Ca^{2+} sensitivity and significant deficits in relaxation [53].

Despite human MYL3 lost its ability to bind Ca^{2+}, the mutational hotspot identified on the protein structure is exactly on the evolutionarily lost binding site. Using AlphaFold 3 to model the interaction between MYL3 and Ca^{2+}, we found that the p.Glu143Lys variant, which replaces the negatively charged glutamate with lysine, disrupts this binding site and moves the calcium ion to a different location. This discovery supports the hypothesis that while MYL3 does not directly bind Ca^{2+}, the evolutionarily conserved second EF-hand still has a regulatory function, possibly by interactions through ions such as Mg^{2+}.

5.3 Mathematical Frameworks and Data Modalities

Our results suggest that the discrete statistical framework is more effective than the Gaussian framework in identifying clinically relevant hotspots. The discrete approach appears better suited for modeling the statistical characteristics of genetic variation within populations. One possible explanation is that the continuous scoring used in the Gaussian framework may not accurately reflect the categorical nature of variants often having either a clearly pathogenic or benign effect, and thereby leading to reduced sensitivity in identifying critical regions.

Integrating AlphaMissense predictions markedly increased the significance of the identified hotspots compared to analysis using only clinical data. These prediction scores provide valuable insights that complement clinical observations. However, the integration comes with a trade-off, as incorporating prediction scores reduces interpretability due to the lack of direct biological evidence. To address this limitation, we validated our findings using ClinVar annotations, ensuring their robustness. Still, future studies should explore alternative validation strategies to refine the evaluation of in-silico predictions. Despite the reduced interpretability, it is noteworthy that a significant proportion of disease hotspots were identified using AlphaMissense scores alone. Given that AlphaMissense is a recent development, this result is encouraging and highlights its potential for future genetic research.

An additional research direction that could provide more complementary and valuable data involves variant effect mapping using multiplexed assays [16]. These assays introduce genetic variants into human-induced pluripotent stem cells (hiPSCs) and assess pathogenicity by measuring relevant biomarkers, such as B-type natriuretic peptide (BNP) levels [4]. This approach has the potential to not only improve the detection of disease hotspots but also provide deeper insights into the underlying disease pathways in HCM.

Chapter 6

Conclusion

In this study, we conducted the largest-scale analysis of genetic data on HCM to date, encompassing over 1.2 million population-based genome sequences alongside more than 10,000 exome sequences from HCM patients. For the first time, we integrated pathogenicity predictions from the transformer-based AlphaMissense model into the analysis, expanding on traditional approaches that use clinical data. We built our statistical frameworks on well-established binomial frameworks previously used for analyzing clinical data, which we then extended to also include pathogenicity predictions created using AlphaMissense and further tested with a novel Gaussian framework.

Our analysis identified several new disease hotspots and confirmed previously reported ones. Additionally, by examining cryo-EM structures and performing simulations with AlphaFold 3, we obtained further structural evidence supporting the proposed hypotheses on the functional impacts and disease mechanisms of these hotspot regions. Moreover, leveraging ClinVar as a high-quality, independent dataset, we were able to strengthen the evidence supporting our findings. Our approach not only highlights the utility of in silico models in genetic variant analysis but also underscores the importance of integrating the data with structural models of the proteins for more comprehensive disease association studies.

Bibliography

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[2] Lorenzo Alamo, James S Ware, Antonio Pinto, Richard E Gillilan, Jonathan G Seidman, Christine E Seidman, and Raúl Padrón. Effects of myosin variants on interacting-heads motif explain distinct hypertrophic and dilated cardiomyopathy phenotypes. *elife*, 6:e24634, 2017.

[3] Marieke Bloemink, John Deacon, Stephen Langer, Carlos Vera, Ariana Combs, Leslie Leinwand, and Michael A Geeves. The hypertrophic cardiomyopathy myosin mutation r453c alters atp binding and hydrolysis of human cardiac β-myosin. *Journal of Biological Chemistry*, 289(8):5158–5167, 2014.

[4] Zhipeng Cao, Yuqing Jia, and Baoli Zhu. Bnp and nt-probnp as diagnostic biomarkers for cardiac dysfunction in both clinical and forensic medicine. *International journal of molecular sciences*, 20(8):1820, 2019.

[5] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

[6] Melanie Colegrave and Michelle Peckham. Structural implications of β-cardiac myosin heavy chain mutations in human disease. *The Anatomical Record*, 297(9):1670–1680, 2014.

[7] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[8] Matthew H Doran, Michael J Rynkiewicz, Elumalai Pavadai, Skylar ML Bodt, David Rasicci, Jeffrey R Moore, Christopher M Yengo, Esther Bullitt, and William Lehman. Myosin loop-4 is critical for optimal tropomyosin repositioning on actin during muscle activation and relaxation. *Journal of General Physiology*, 155(2):e202213274, 2022.

[9] Matthew H Doran, Michael J Rynkiewicz, David Rasicci, Skylar ML Bodt, Meaghan E Barry, Esther Bullitt, Christopher M Yengo, Jeffrey R Moore, and William Lehman. Conformational changes linked to adp release from human cardiac myosin bound to actin-tropomyosin. *Journal of General Physiology*, 155(3):e202213267, 2023.

[10] Disha Dumka, John Talent, Irina Akopova, Georginna Guzman, Danuta Szczesna-Cordary, and Julian Borejdo. E22k mutation of rlc that causes familial hypertrophic cardiomyopathy in heterozygous mouse myocardium: effect on cross-bridge kinetics. *American Journal of Physiology-Heart and Circulatory Physiology*, 291(5):H2098–H2106, 2006.

[11] Debabrata Dutta, Vu Nguyen, Kenneth S Campbell, Raúl Padrón, and Roger Craig. Cryo-em structure of the human cardiac myosin filament. *Nature*, 623(7988):853–862, 2023.

[12] Julian von der Ecken, Sarah M Heissler, Salma Pathan-Chhatbar, Dietmar J Manstein, and Stefan Raunser. Cryo-em structure of a human cytoplasmic actomyosin complex at near-atomic resolution. *Nature*, 534(7609):724–728, 2016.

[13] Maria J Eriksson, Brian Sonnenberg, Anna Woo, Paul Rakowski, Thomas G Parker, E Douglas Wigle, and Harry Rakowski. Long-term outcome in patients with apical hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, 39(4):638–45, Feb 2002.

[14] D Brian Foster, Teruo Noguchi, Peter VanBuren, Anne M Murphy, and Jennifer E Van Eyk. C-terminal truncation of cardiac troponin i causes divergent effects on atpase and force: implications for the pathophysiology of myocardial stunning. *Circulation research*, 93(10):917–924, 2003.

[15] Andrew J Franklin, Tamatha Baxley, Tomoyoshi Kobayashi, and Joseph M Chalovich. The c-terminus of troponin t is essential for maintaining the inactive state of regulated actin. *Biophysical journal*, 102(11):2536–2544, 2012.

[16] Clayton E Friedman, Shawn Fayer, Sriram Pendyala, Wei-Ming Chien, Alexander Loiben, Linda Tran, Leslie S Chao, Ashley McKinstry, Dania Ahmed, Stephen D Farris, et al. Multiplexed functional assessments of myh7 variants in human cardiomyocytes. *Circulation: Genomic and Precision Medicine*, 17(2):e004377, 2024.

[17] Binnu Gangadharan, Margaret S Sunitha, Souhrid Mukherjee, Ritu Roy Chowdhury, Farah Haque, Narendrakumar Sekar, Ramanathan Sowdhamini, James A Spudich, and John A Mercer. Molecular mechanisms and structural features of cardiomyopathy-causing troponin t mutants in the tropomyosin overlap region. *Proceedings of the National Academy of Sciences*, 114(42):11115–11120, 2017.

[18] Ortíz-Genga M Barriales-Villa R Fernández X Rodríguez-García I Mazzanti A Veira E Maneiro E Rebolo P Lesende I Cazón L Freimark D Gimeno-Blanes JR Seidman C Seidman J McKenna W Monserrat L García-Giustiniani D, Arad M. Phenotype and prognostic correlations of the converter region mutations affecting the beta myosin heavy chain. *Heart*, 2015.

[19] Ochoa JP McKenna WJ García-Hernández S, de la Higuera Romero L. Emerging themes in genetics of hypertrophic cardiomyopathy: Current status and clinical application. *Canadian Journal of Cardiology*, 2024.

[20] Aldrin V Gomes and James D Potter. Cellular and molecular aspects of familial hypertrophic cardiomyopathy caused by mutations in the cardiac troponin i gene. *Molecular and cellular biochemistry*, 263:99–114, 2004.

[21] Alessandro Grinzato, Daniel Auguin, Carlos Kikuti, Neha Nandwani, Dihia Moussaoui, Divya Pathak, Eaazhisai Kandiah, Kathleen Ruppel, James A Spudich, Anne M Houdusse, et al. Cryo-em structure of the folded-back state of human β-cardiac myosin. *Biophysical Journal*, 122(3):258a–259a, 2023.

[22] Adam S Helms, Andrea D Thompson, Amelia A Glazier, Neha Hafeez, Samat Kabani, Juliani Rodriguez, Jaime M Yob, Helen Woolcock, Francesco Mazzarotto, Neal K Lakdawala, et al. Spatial and functional distribution of mybpc3 pathogenic variants and clinical outcomes in patients with hypertrophic cardiomyopathy. *Circulation: Genomic and Precision Medicine*, 13(5):396–405, 2020.

[23] Alex Henrie, Sarah E Hemphill, Nicole Ruiz-Schultz, Brandon Cushman, Marina T DiStefano, Danielle Azzariti, Steven M Harrison, Heidi L Rehm, and

Karen Eilbeck. Clinvar miner: demonstrating utility of a web-based tool for viewing and filtering clinvar data. *Human mutation*, 39(8):1051–1060, 2018.

[24] Olga M Hernandez, Philippe R Housmans, and James D Potter. Invited review: pathophysiology of cardiac muscle contraction and relaxation as a result of alterations in thin filament regulation. *Journal of applied physiology*, 90(3):1125–1136, 2001.

[25] Julian R Homburger, Eric M Green, Colleen Caleshu, Margaret S Sunitha, Rebecca E Taylor, Kathleen M Ruppel, Raghu Prasad Rao Metpally, Steven D Colan, Michelle Michels, Sharlene M Day, et al. Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proceedings of the National Academy of Sciences*, 113(24):6701–6706, 2016.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[27] Jamie D Kapplinger, Andrew P Landstrom, J Martijn Bos, Benjamin A Salisbury, Thomas E Callis, and Michael J Ackerman. Distinguishing hypertrophic cardiomyopathy-associated mutations from background genetic noise. *Journal of cardiovascular translational research*, 7:347–361, 2014.

[28] Ruppel KM Kawana M, Spudich JA. Hypertrophic cardiomyopathy: Mutations to mechanisms to therapies. *Frontiers in Physiology*, 2022.

[29] Lucas K Keyt, Jason M Duran, Quan M Bui, Chao Chen, Michael I Miyamoto, Jorge Silva Enciso, Jil C Tardiff, and Eric D Adler. Thin filament cardiomyopathies: A review of genetics, disease mechanisms, and emerging therapeutics. *Frontiers in cardiovascular medicine*, 9:972301, 2022.

[30] Doo Nam Kim, Timothy M Jacobs, and Brian Kuhlman. Boosting protein stability with the computational design of β-sheet surfaces. *Protein Science*, 25(3):702–710, 2016.

[31] Leonie M Kurzlechner, Edward G Jones, Amy M Berkman, Hanna J Tadros, Jill A Rosenfeld, Yaping Yang, Hari Tunuguntla, Hugh D Allen, Jeffrey J Kim, and Andrew P Landstrom. Signal-to-noise analysis can inform the likelihood that

incidentally identified variants in sarcomeric genes are associated with pediatric cardiomyopathy. *Journal of Personalized Medicine*, 12(5):733, 2022.

[32] Crocini C. Leinwand L.A. Lehman, S.J. Targeting the sarcomere in inherited cardiomyopathies. *Nature Reviews Cardiology*, 2022.

[33] Steffen Lindert, Yuanhua Cheng, Peter Kekenes-Huskey, Michael Regnier, and J Andrew McCammon. Effects of hcm ctni mutation r145g on troponin structure and modulation by pka phosphorylation elucidated by molecular dynamics simulations. *Biophysical Journal*, 108(2):395–407, 2015.

[34] Janine Lossie, Dmitry S Ushakov, Michael A Ferenczi, Sascha Werner, Sandro Keller, Hannelore Haase, and Ingo Morano. Mutations of ventricular essential myosin light chain disturb myosin binding and sarcomeric sorting. *Cardiovascular research*, 93(3):390–396, 2012.

[35] Nikolaos Louros, Joost Schymkowitz, and Frederic Rousseau. Mechanisms and pathology of protein misfolding and aggregation. *Nature Reviews Molecular Cell Biology*, 24(12):912–933, 2023.

[36] Aditi Madan, Meera C Viswanathan, Kathleen C Woulfe, William Schmidt, Agnes Sidor, Ting Liu, Tran H Nguyen, Bosco Trinh, Cortney Wilson, Sineej Madathil, et al. Tnnt2 mutations in the tropomyosin binding region of tnt1 disrupt its role in contractile inhibition and stimulate cardiac dysfunction. *Proceedings of the National Academy of Sciences*, 117(31):18822–18831, 2020.

[37] AJ Marian and Robert Roberts. The molecular genetic basis for hypertrophic cardiomyopathy. *Journal of molecular and cellular cardiology*, 33(4):655–670, 2001.

[38] Braunwald E Marian AJ. Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circulation Research*, 2017.

[39] B J Maron, J M Gardin, J M Flack, S S Gidding, T T Kurosaki, and D E Bild. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. echocardiographic analysis of 4111 subjects in the cardia study. coronary artery risk development in (young) adults. *Circulation*, 92(4):785–9, Aug 1995.

[40] B J Maron, J Shirani, L C Poliac, R Mathenge, W C Roberts, and F O Mueller. Sudden death in young competitive athletes. clinical, demographic, and pathological profiles. *JAMA*, 276(3):199–204, Jul 1996.

[41] Barry J Maron, Ethan J Rowin, Susan A Casey, John R Lesser, Ross F Garberich, Deepa M McGriff, and Martin S Maron. Hypertrophic cardiomyopathy in children, adolescents, and young adults associated with low cardiovascular mortality with contemporary management strategies. *Circulation*, 133(1):62–73, 2016.

[42] Nancy L Meyer and P Bryant Chase. Role of cardiac troponin i carboxy terminal mobile domain and linker sequence in regulating cardiac contraction. *Archives of biochemistry and biophysics*, 601:80–87, 2016.

[43] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

[44] Jeffrey R Moore, Leslie Leinwand, and David M Warshaw. Understanding cardiomyopathy phenotypes based on the functional impact of mutations in the myosin motor. *Circulation research*, 111(3):375–385, 2012.

[45] Suman Nag, Darshan V Trivedi, Saswata S Sarkar, Arjun S Adhikari, Margaret S Sunitha, Shirley Sutton, Kathleen M Ruppel, and James A Spudich. The myosin mesa and the basis of hypercontractility caused by hypertrophic cardiomyopathy mutations. *Nature structural & molecular biology*, 24(6):525–533, 2017.

[46] Krzysztof Nieznanski, Hanna Nieznanska, Krzysztof Skowronek, Andrzej A Kasprzak, and Dariusz Stepkowski. Ca2+ binding to myosin regulatory light chain affects the conformation of the n-terminus of essential light chain and its binding to actin. *Archives of biochemistry and biophysics*, 417(2):153–158, 2003.

[47] Daniel Peter Sayer Osborn, Leila Emrahi, Joshua Clayton, Mehrnoush Toufan Tabrizi, Alex Yui Bong Wan, Reza Maroofian, Mohammad Yazdchi, Michael Leon Enrique Garcia, Hamid Galehdari, Camila Hesse, et al. Autosomal recessive cardiomyopathy and sudden cardiac death associated with variants in myl3. *Genetics in Medicine*, 23(4):787–792, 2021.

[48] Elumalai Pavadai, William Lehman, and Michael J Rynkiewicz. Protein-protein docking reveals dynamic interactions of tropomyosin on actin filaments. *Biophysical journal*, 119(1):75–86, 2020.

[49] Cristina Risi, Betty Belknap, Eva Forgacs-Lonart, Samantha P Harris, Gunnar F Schröder, Howard D White, and Vitold E Galkin. N-terminal domains of cardiac myosin binding protein c cooperatively activate the thin filament. *Structure*, 26(12):1604–1611, 2018.

[50] Cristina M Risi, Malay Patra, Betty Belknap, Samantha P Harris, Howard D White, and Vitold E Galkin. Interaction of the c2 ig-like domain of cardiac myosin binding protein-c with f-actin. *Journal of molecular biology*, 433(19):167178, 2021.

[51] Cristina M Risi, Edwin Villanueva, Betty Belknap, Rachel L Sadler, Samantha P Harris, Howard D White, and Vitold E Galkin. Cryo-electron microscopy reveals cardiac myosin binding protein-c m-domain interactions with the thin filament. *Journal of molecular biology*, 434(24):167879, 2022.

[52] Takeda S. Core domain of human cardiac troponin. 2016.

[53] Atsushi Sanbe, David Nelson, James Gulick, Elizabeth Setser, Hanna Osinska, Xuejun Wang, Timothy E Hewett, Raisa Klevitsky, Eric Hayes, David M Warshaw, et al. In vivo analysis of an essential myosin light chain mutation linked to familial hypertrophic cardiomyopathy. *Circulation research*, 87(4):296–302, 2000.

[54] Masataka Sata, Mitsuo Ikebe, et al. Functional analysis of the mutations in the human cardiac beta-myosin that are responsible for familial hypertrophic cardiomyopathy. implication for the clinical outcome. *The Journal of clinical investigation*, 98(12):2866–2873, 1996.

[55] James A Spudich. The myosin mesa and a possible unifying hypothesis for the molecular basis of human hypertrophic cardiomyopathy, 2015.

[56] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[57] Danuta Szczesna, Debalina Ghosh, Qi Li, Aldrin V Gomes, Georgianna Guzman, Carlos Arana, Gang Zhi, James T Stull, and James D Potter. Familial hypertrophic cardiomyopathy mutations in the regulatory light chains of myosin affect their structure, ca2+ binding, and phosphorylation. *Journal of Biological Chemistry*, 276(10):7086–7092, 2001.

[58] Danuta Szczesna-Cordary and Pieter P de Tombe. Myosin light chain phosphorylation, novel targets to repair a broken heart?, 2016.

[59] Huang HC Fung E. Teekakirikul P, Zhu W. Hypertrophic cardiomyopathy: An overview of genetics and management. *Biomolecules*, 2019.

[60] Darshan V Trivedi, Arjun S Adhikari, Saswata S Sarkar, Kathleen M Ruppel, and James A Spudich. Hypertrophic cardiomyopathy and the myosin mesa: viewing an old disease in a new light. *Biophysical reviews*, 10(1):27–48, 2018.

[61] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

[62] Roddy Walsh, Francesco Mazzarotto, Nicola Whiffin, Rachel Buchan, William Midwinter, Alicja Wilk, Nicholas Li, Leanne Felkin, Nathan Ingold, Risha Govind, et al. Quantitative approaches to variant classification increase the yield and precision of genetic testing in mendelian diseases: the case of hypertrophic cardiomyopathy. *Genome medicine*, 11:1–18, 2019.

[63] Adam Waring, Andrew Harper, Silvia Salatino, Christopher Kramer, Stefan Neubauer, Kate Thomson, Hugh Watkins, and Martin Farrall. Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathy. *Journal of Medical Genetics*, 58(8):556–564, 2021.

[64] Sunil Yadav, Yoel H Sitbon, Katarzyna Kazmierczak, and Danuta Szczesna-Cordary. Hereditary heart disease: pathophysiology, clinical presentation, and animal models of hcm, rcm, and dcm associated with mutations in cardiac myosin light chains. *Pflügers Archiv-European Journal of Physiology*, 471:683–699, 2019.

[65] Yurika Yamada, Keiichi Namba, and Takashi Fujii. Cardiac muscle thin filament structures reveal calcium regulatory mechanism. *Nature communications*, 11(1):153, 2020.

[66] Chen-Ching Yuan, Katarzyna Kazmierczak, Jingsheng Liang, Rosemeire Kanashiro-Takeuchi, Thomas C Irving, Aldrin V Gomes, Yihua Wang, Thomas P Burghardt, and Danuta Szczesna-Cordary. Hypercontractile mutant of ventricular myosin essential light chain leads to disruption of sarcomeric structure and function and results in restrictive cardiomyopathy in mice. *Cardiovascular research*, 113(10):1124–1136, 2017.

List of Figures

List of Tables

Appendix A

Appendix

A.1 Graphical Validation of Hotspot Regions

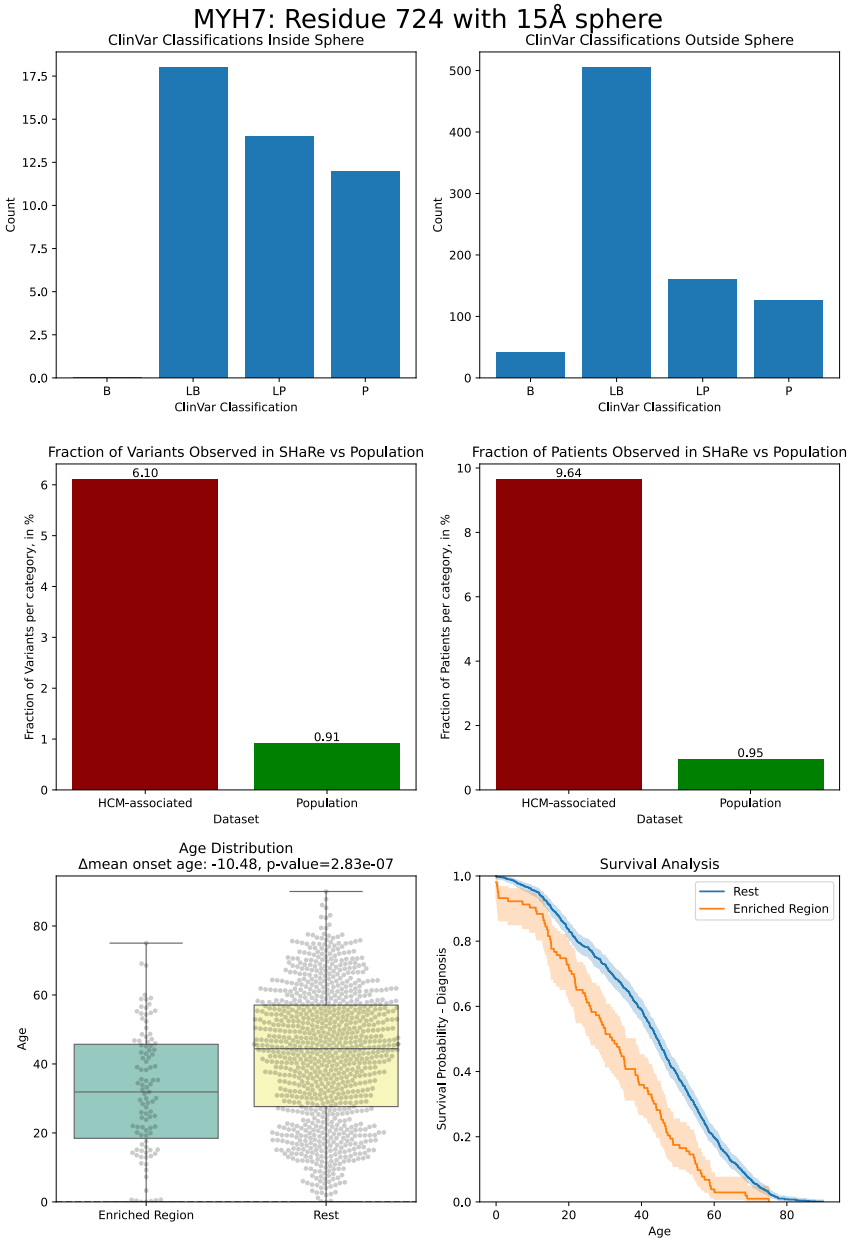

Figure A.1: Validation plot for MYH7 hotspot located on residue 724.

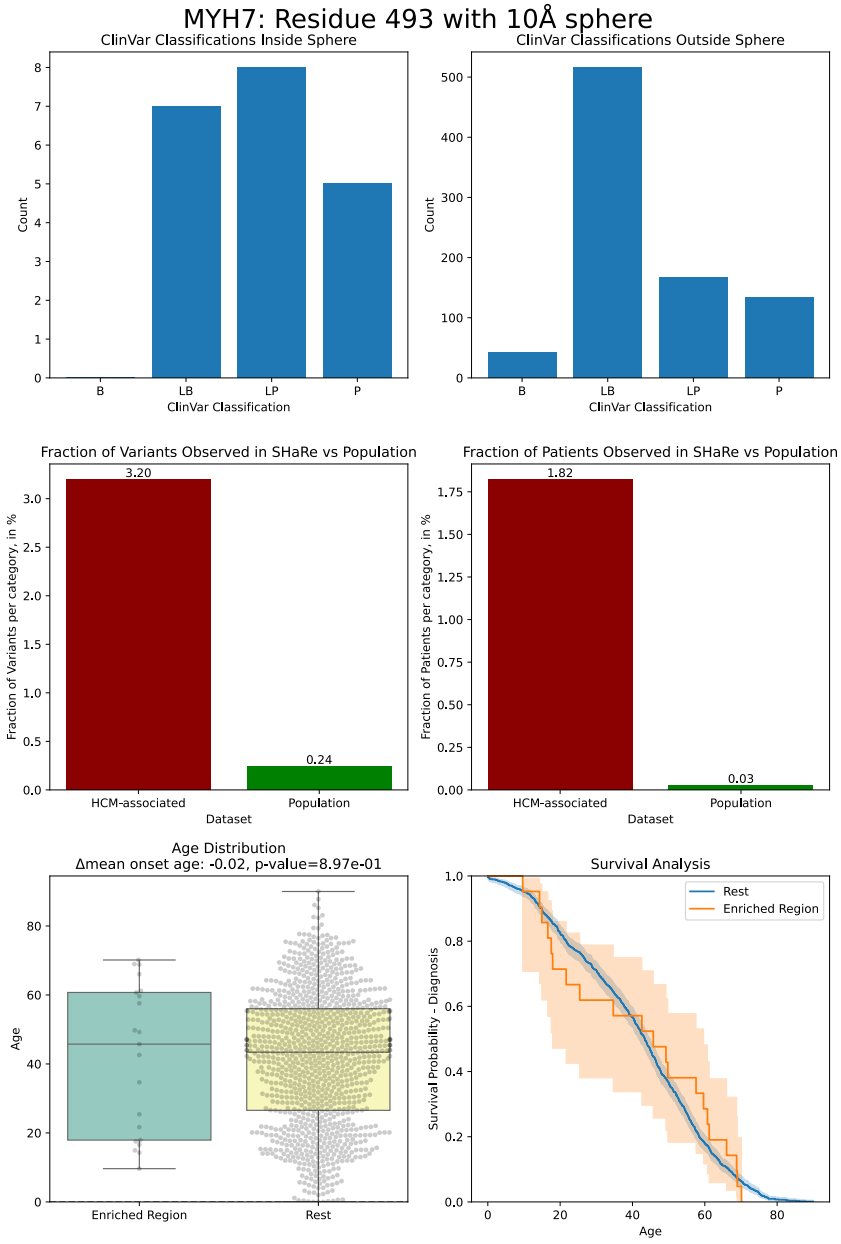

Figure A.2: Validation plot for MYH7 hotspot located on residue 493.

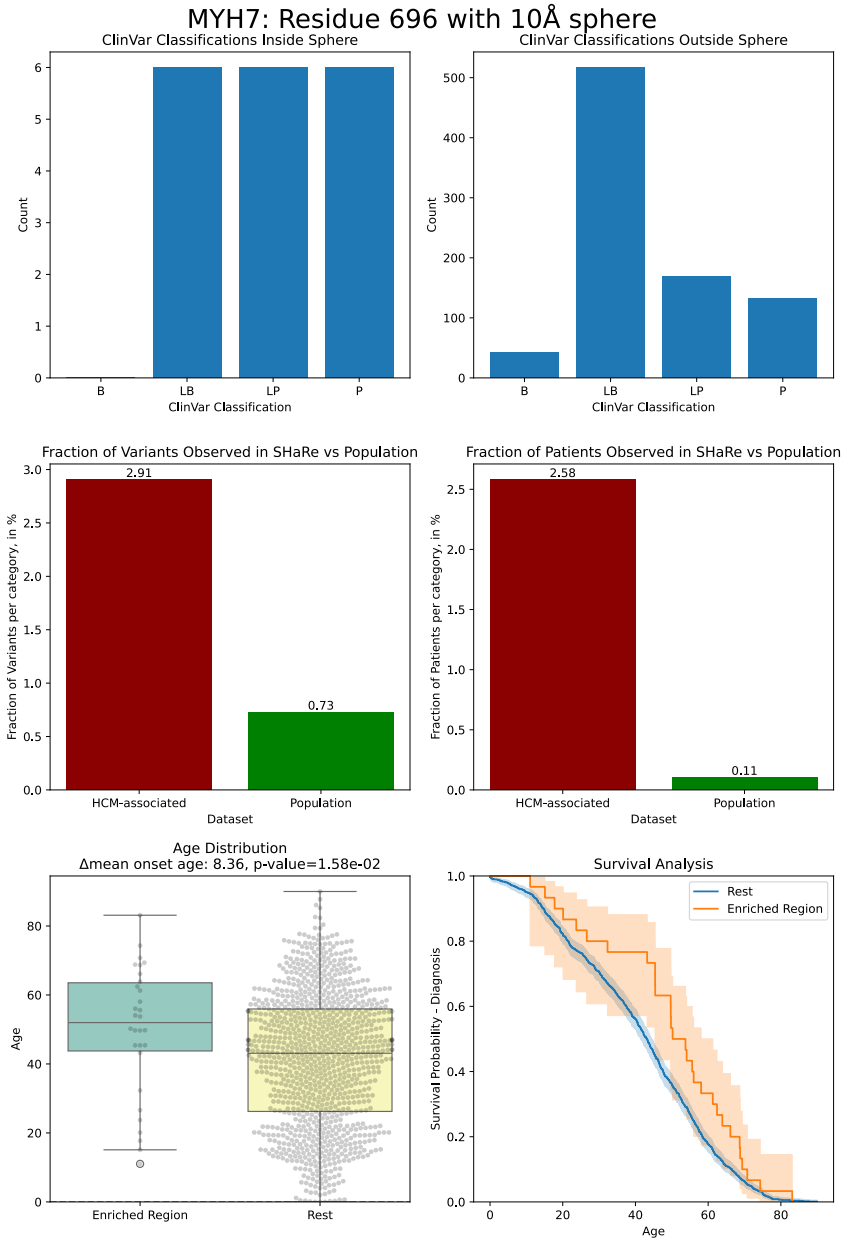

Figure A.3: Validation plot for MYH7 hotspot located on residue 696.

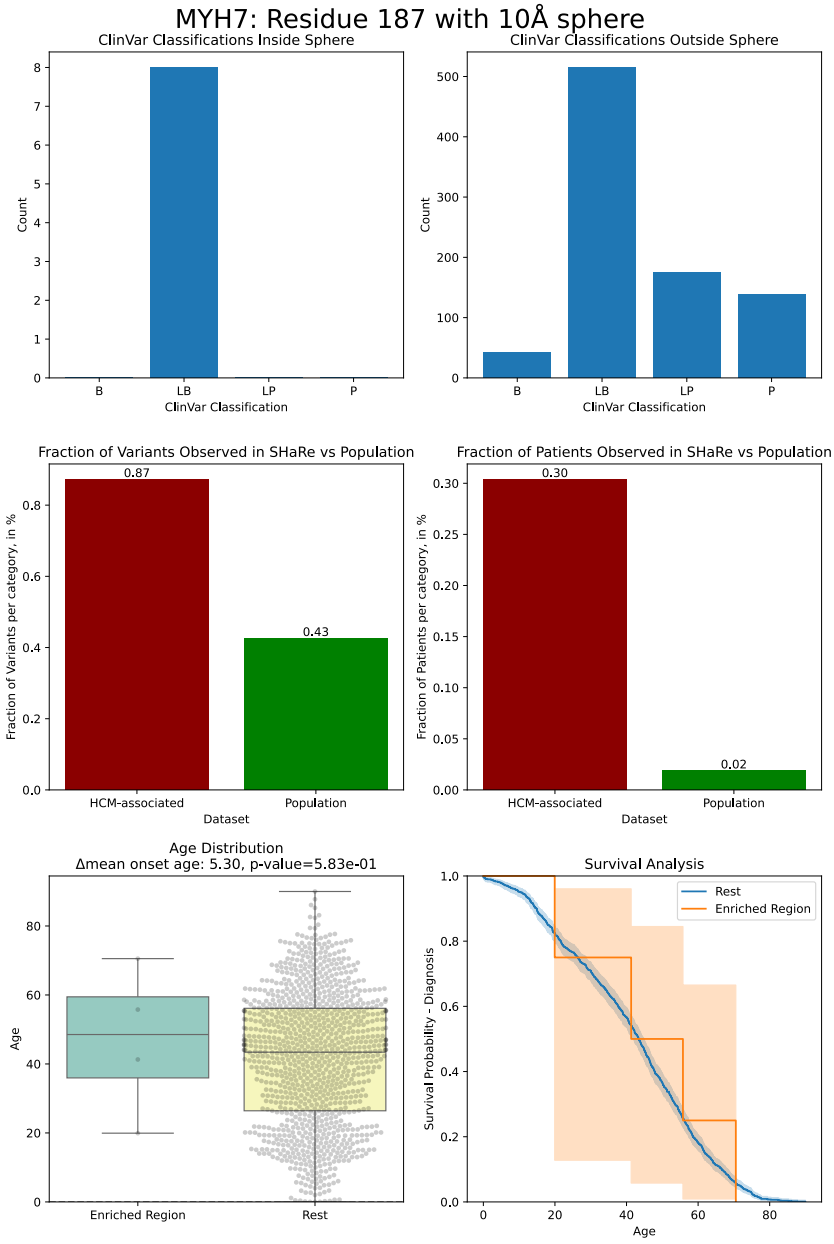

Figure A.4: Validation plot for MYH7 hotspot located on residue 187.

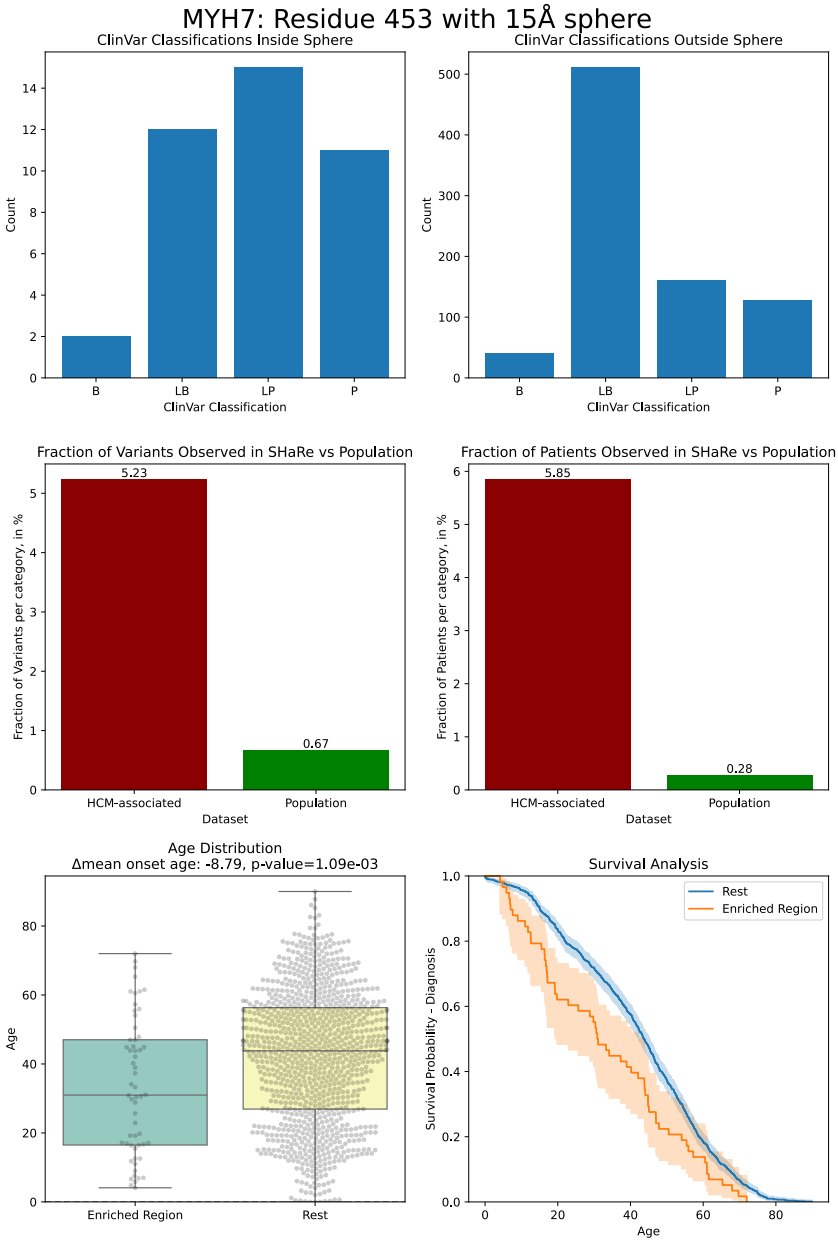

Figure A.5: Validation plot for MYH7 hotspot located on residue 453.

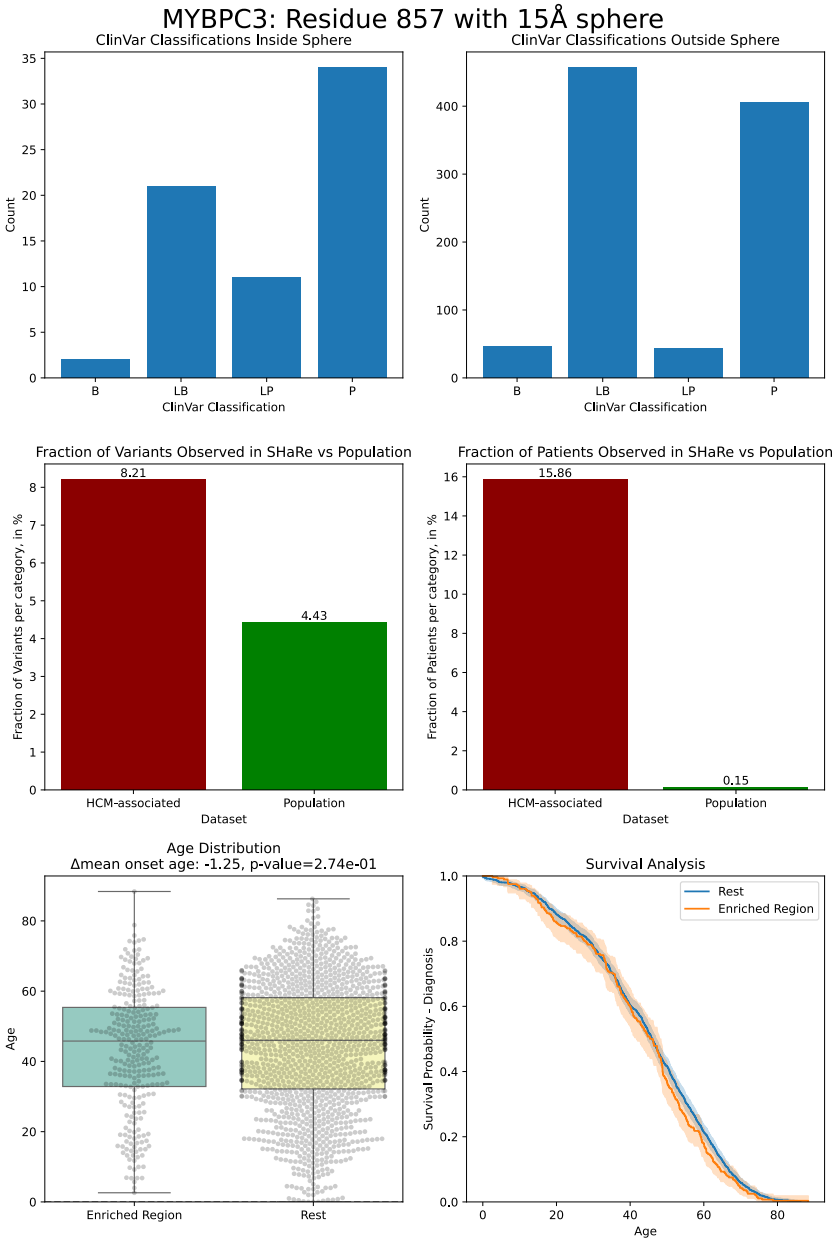

Figure A.6: Validation plot for MYBPC3 hotspot located on residue 857.

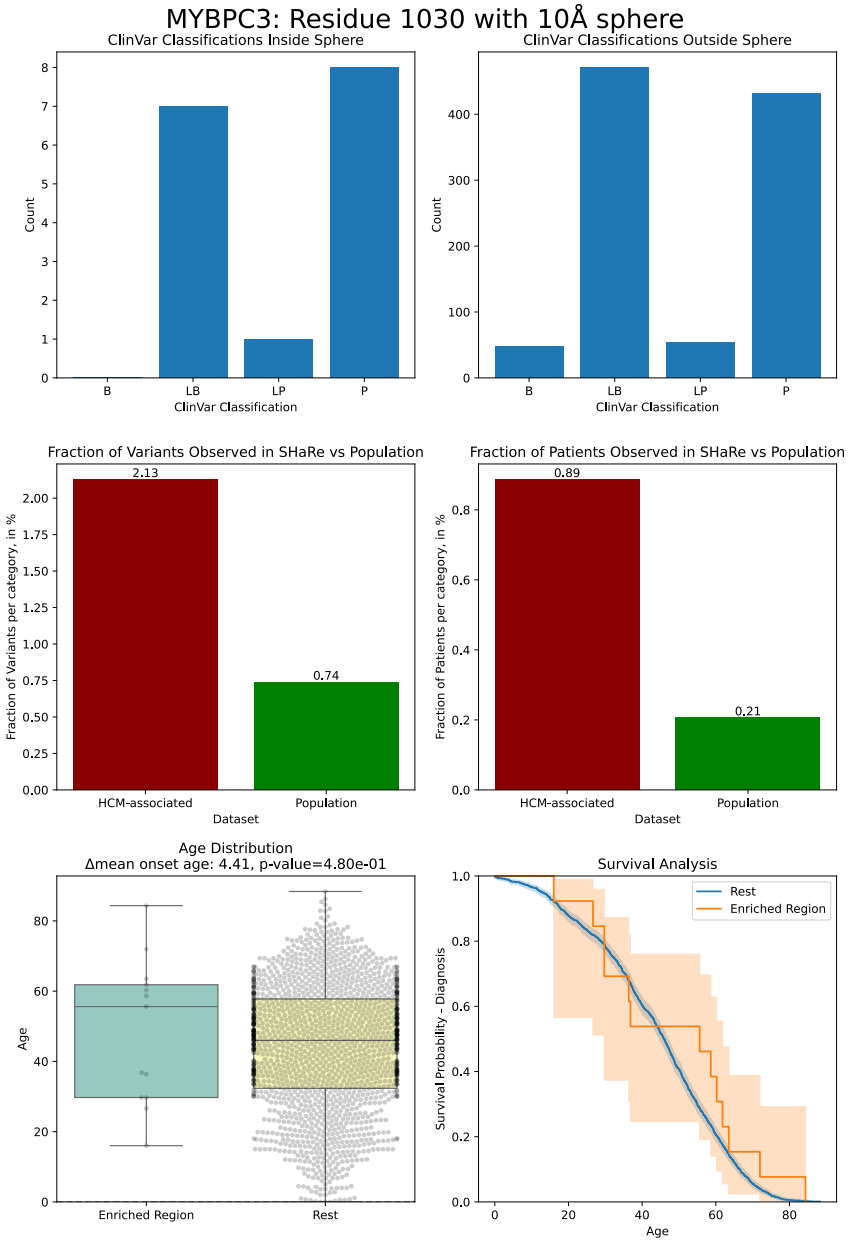

Figure A.7: Validation plot for MYBPC3 hotspot located on residue 1030.

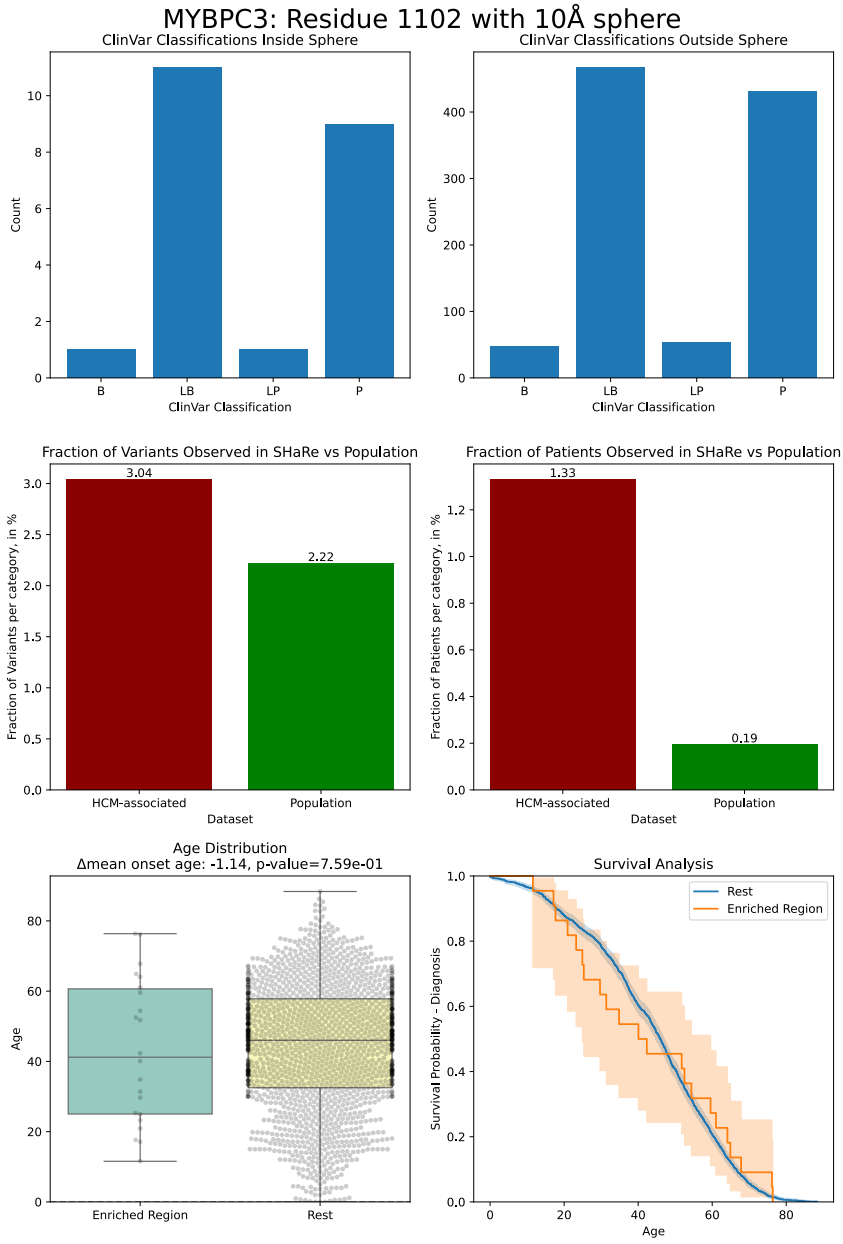

Figure A.8: Validation plot for MYBPC3 hotspot located on residue 1102.

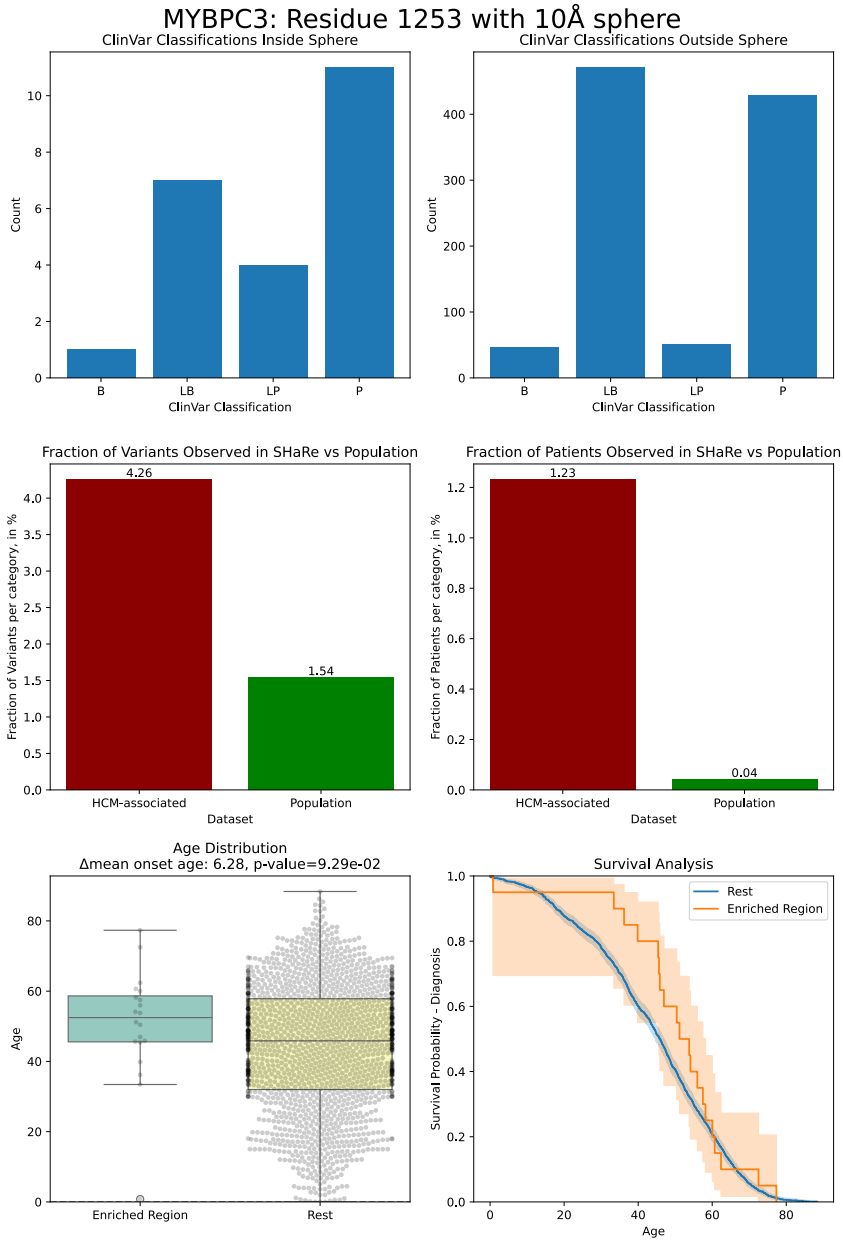

Figure A.9: Validation plot for MYBPC3 hotspot located on residue 1253.

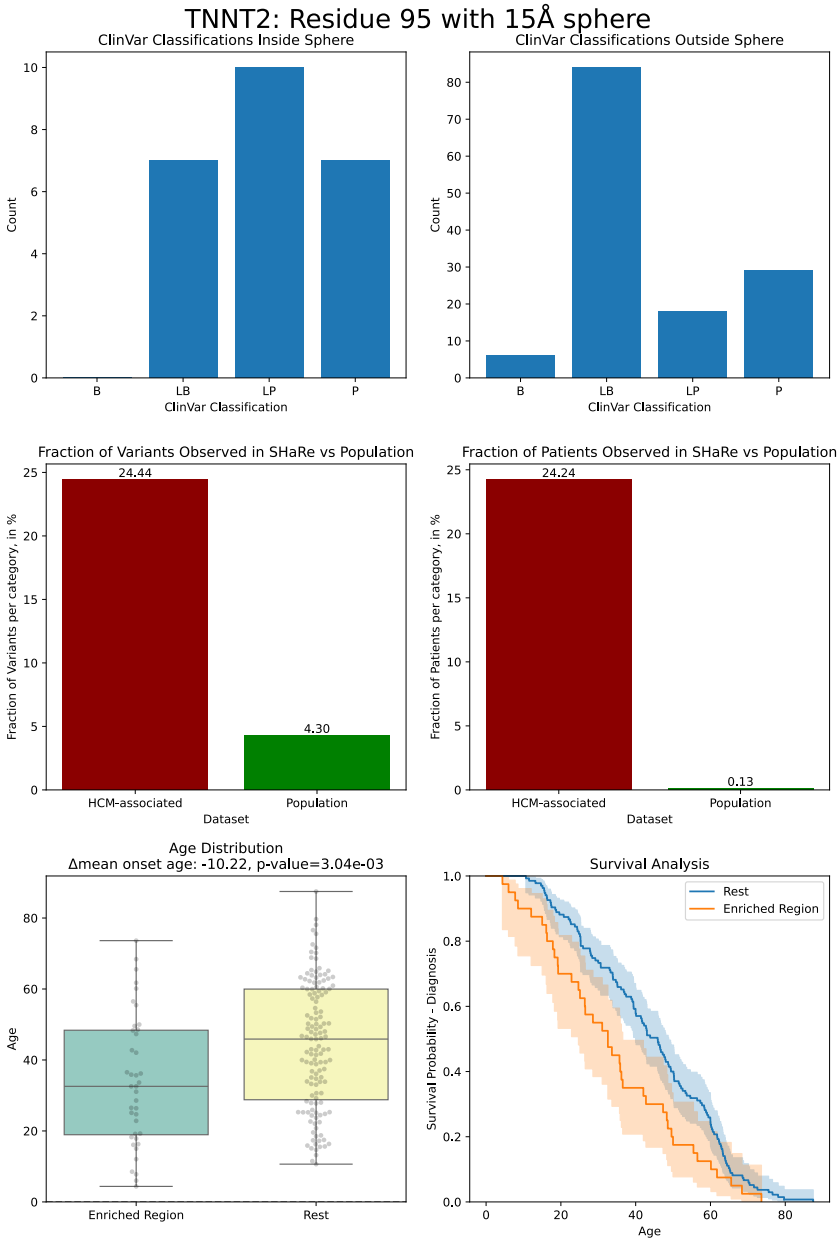

Figure A.10: Validation plot for TNNT2 hotspot located on residue 95.

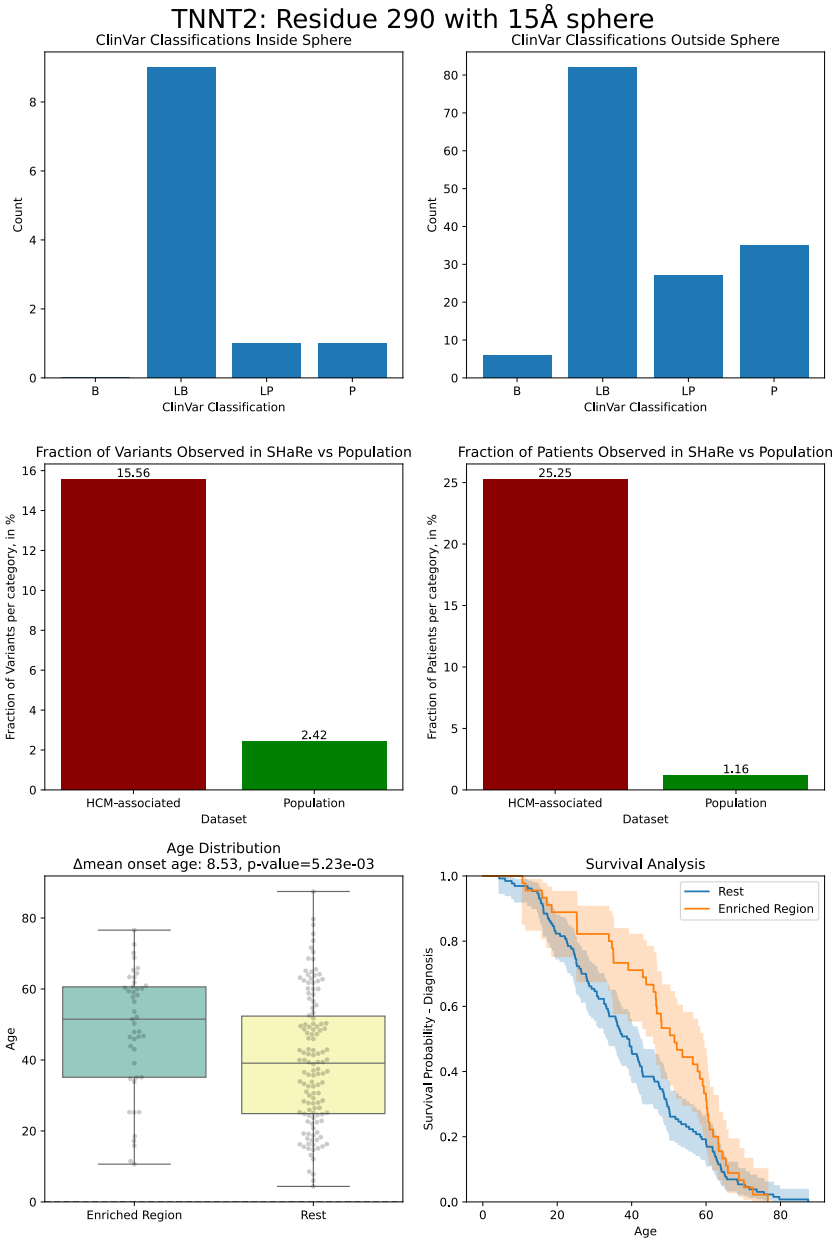

Figure A.11: Validation plot for TNNT2 hotspot located on residue 290.

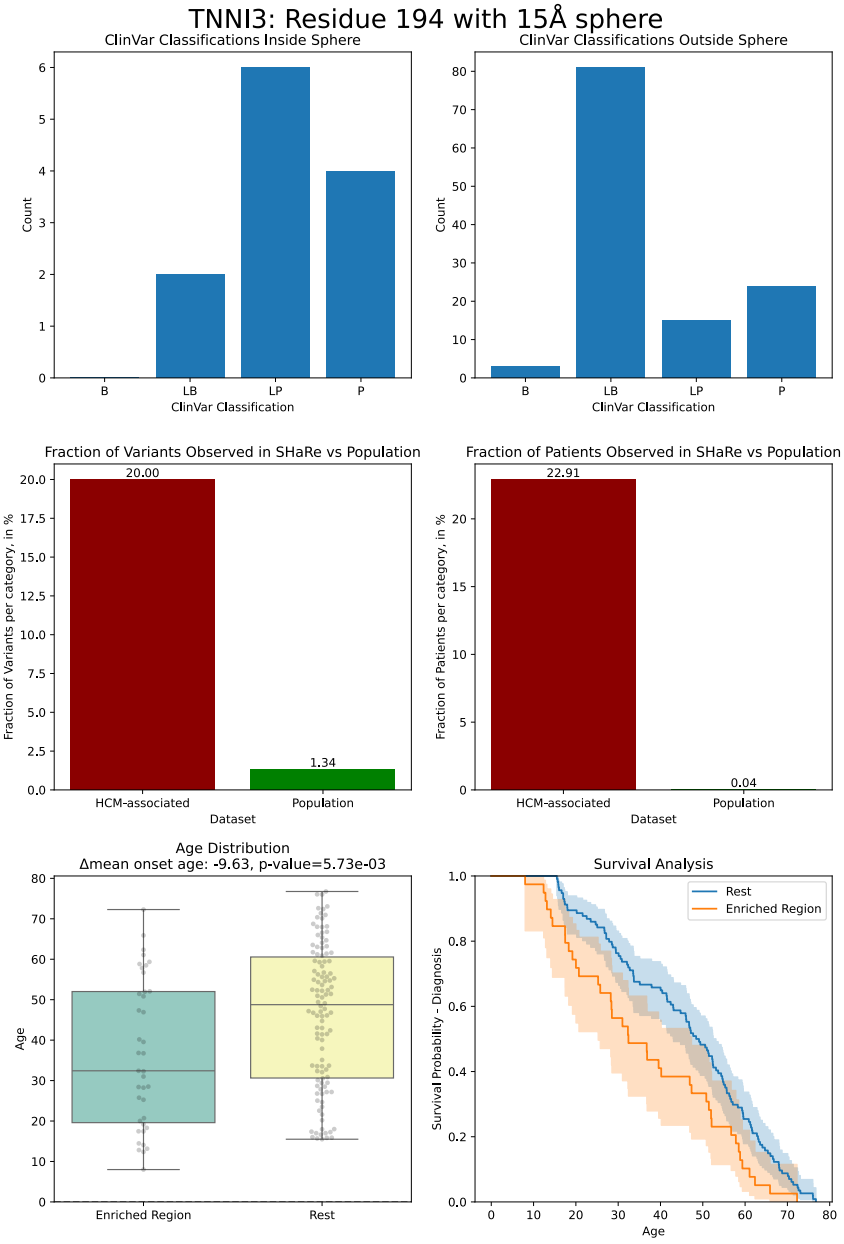

Figure A.12: Validation plot for TNNI3 hotspot located on residue 194.

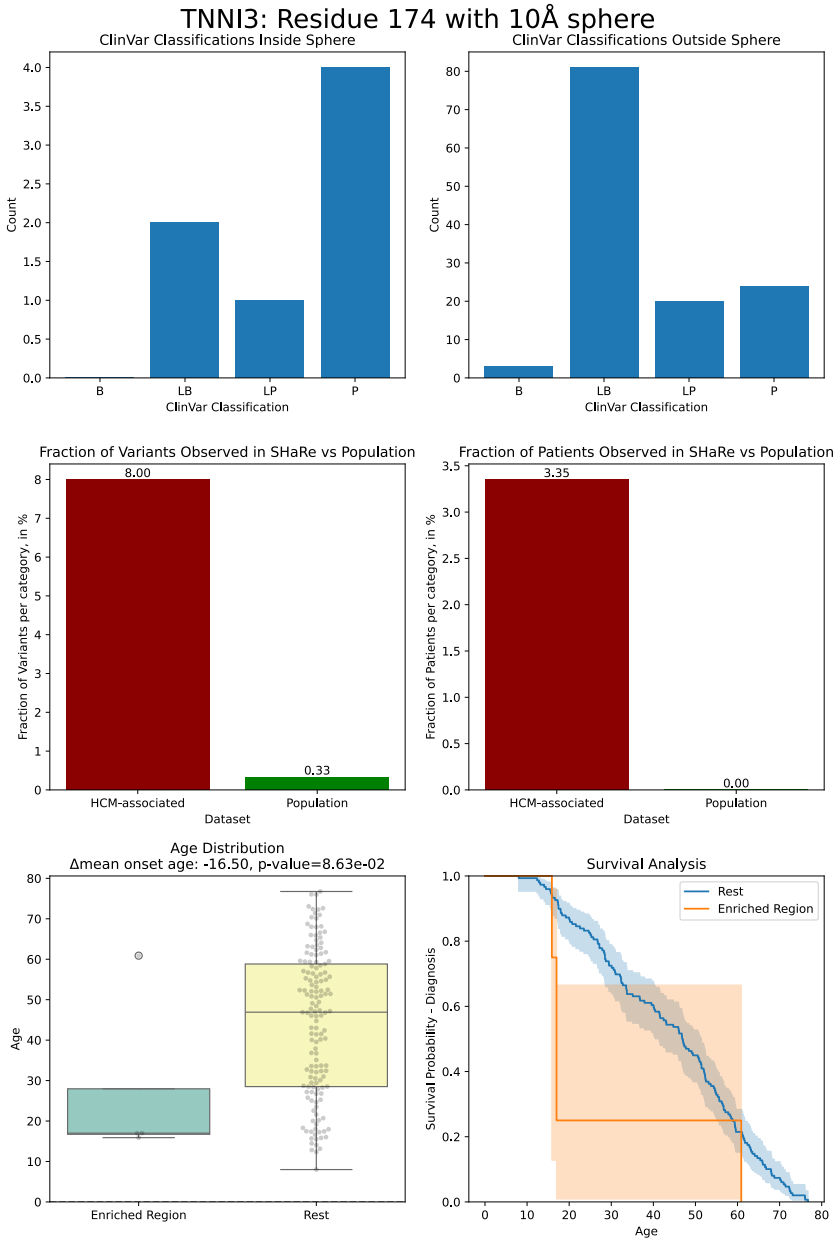

Figure A.13: Validation plot for TNNI3 hotspot located on residue 174.

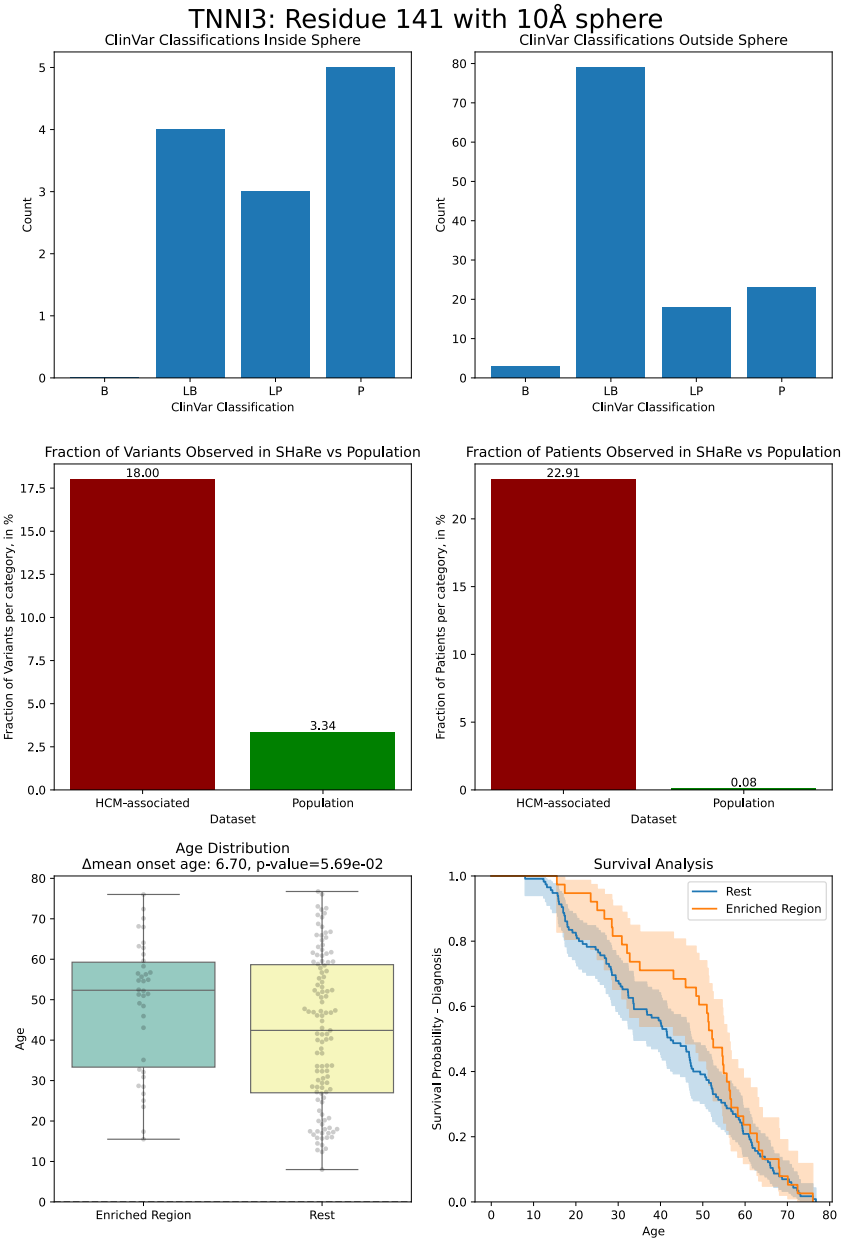

Figure A.14: Validation plot for TNNI3 hotspot located on residue 141.

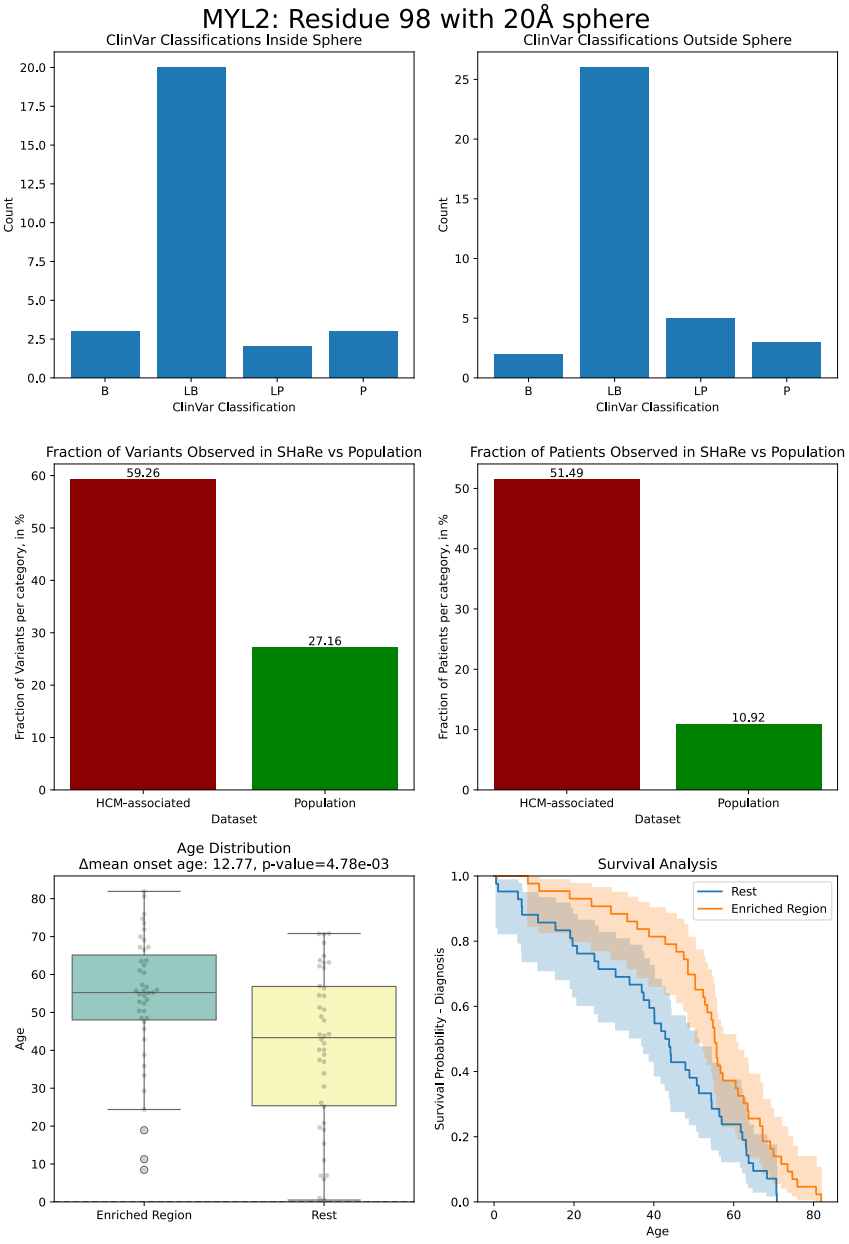

Figure A.15: Validation plot for MYL2 hotspot located on residue 98.

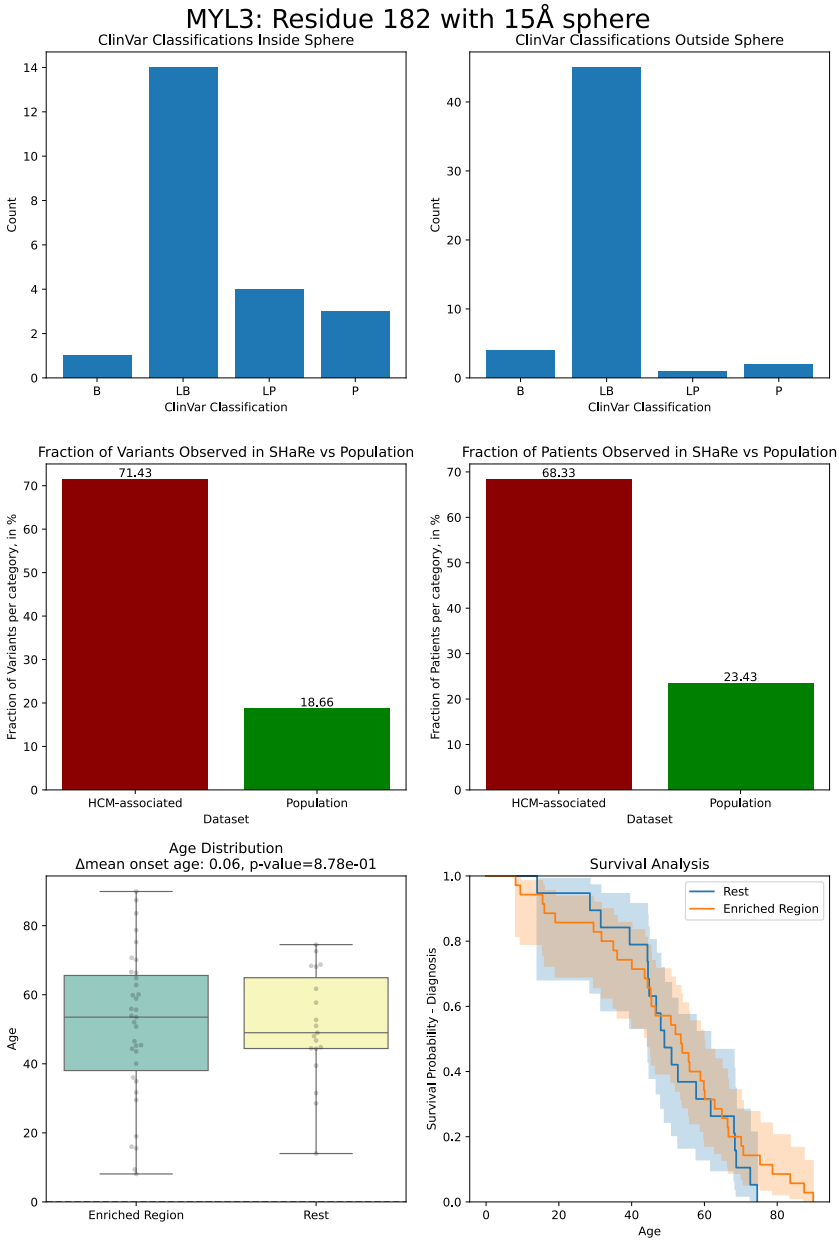

Figure A.16: Validation plot for MYL3 hotspot located on residue 182.