This electronic thesis or dissertation has been downloaded from the King's Research Portal at https://kclpure.kcl.ac.uk/portal/



Assessing the Impacts of Small Variants in the Giant Sarcomeric Protein Titin Associated with Muscle Disease

Weston, Timir

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT



Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Oct. 2023

Assessing the Impacts of Small Variants in the Giant Sarcomeric Protein Titin Associated with Muscle Disease



Timir Weston

First Supervisor: Professor Franca Fraternali

Second Supervisor: Professor Mathias Gautel

Dissertation submitted for the degree of

Doctor of Philosophy

Randall Centre for Cell & Molecular Biophysics

King's College London

United Kingdom

April 2023

Declaration

I, Timir Weston, declare the work presented herein to be my own, and that no part has been previously submitted in fulfilment of the requirements of any academic degree. I carried out this work at the Randall Centre for Cell & Molecular Biosciences at King's College London, under the supervision of Professor Franca Fraternali (King's College London, University College London) and Professor Mathias Gautel (King's College London).

Certain of the materials and results contained in this dissertation are included in manuscripts for future publication. Abridged versions of Chapters 1 and 2 are separately in the process of submission to academic journals, while a manuscript incorporating elements of Chapter 4 is in preparation. During the course of this thesis, I have also contributed to the paper: "*in silico* and *in vitro* analysis of IL36RN mutations reveals critical residues for the function of the interleukin-36 receptor complex", currently in revision for the Journal of Investigative Dermatology, in collaboration with the laboratory of Prof. Francesca Capon (King's College London).

This research was funded/supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London and/or the NIHR Clinical Research Facility. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Timir Weston

17 April 2023

Acknowledgements

I would like to thank those who have encouraged, supported, and tolerated me over the many years of work that have culminated in the submission of this thesis, both inside and outside of the academic sphere.

To my primary supervisor, Professor Franca Fraternali, I will always be grateful for the time, care, and attention to detail you put towards not only myself and my work, but to all students under your supervision; your investment in our progress, our growth as researchers, and our inheritance of our projects, is an inspiration to all of us, and your constant advocacy on our behalf is an oft-unmentioned yet never unappreciated blessing. I am truly glad and grateful that I undertook my doctorate at the Fraternali lab, and had the opportunity to study with you. To my second supervisor, Professor Mathias Gautel, your sagacity and knowledge in this field forms the bedrock of this work, a solid foundation from which I was able to grow over the course of the past four years. The advice, support, and information you provided was a reassurance in times of doubt and uncertainty, without which this work would not have been.

To those friends and colleagues I have had the privilege of learning from and working together with over the last four years, each and every one of you has given me something to be grateful for. To those who came before: Dr Joseph Chi-Fung Ng, for his tireless contributions to the running of the group and his endless patience for my constant questions; Dr Irene Marzuoli and Dr Marius Kausas, for their respective aid and support with understanding and running molecular dynamics simulations, optimising hardware configurations for performance, and requesting computing time on external clusters; Dr Martin Rees, for his indispensable help on all aspects of the background to this work, and for the annotation of pathogenicity and thermostability; Dr Daniel Koch, for many enlightening conversations around the obscurin protein.

To those who will soon follow: Dongjun Guo, from many panicked emails trying to set up Rosetta on Rosalind to a full-fledged web developer and data scientist, your growth has been marvellous to watch and humbling to be a part of; Shohei Yoshida, your indulgence of my many terrible jokes with hearty laughter is most appreciated; Lutecia Servius, your guidance and excellent advice was material to the completion of the machine learning chapters, and your tolerance of my many terrible jokes is also appreciated; Julian Herrera Braga, though most of our meetings have taken place online, your work and your help both paved the way to these projects' completion; Oriol Gracia I Carmona, though I had only a short time to get to know you, your aid with all things molecular dynamics was instrumental in my data analysis; Timi Ellinas, our long and animated discussions about anything and everything under the sun were a much-needed antidote to all academic doldrums.

To those teachers whose names and lessons remain with me, Keith Mallinson, George Lester, Laura Hutchings, Dr Gary Savage, Dr Timothy Walker, Professor Rebecca Oakey. To all others, past and present, professors and pupils, post-docs and part-timers, I thank you unreservedly, and express my deepest appreciation, no matter how large or how small the contribution.

To my father, Ian Weston, who taught me of the importance of truth, and the need to pursue it at all times.

To my mother, Dr Latha Weston, who taught me of the importance of determination, and self-reflection.

To my sister Maya, the little girl with pigtails who used to hold my earlobes when scared, now a fiercely independent and determined young woman, of whom I could not be more proud.

To my brother Kiran, whom I can no longer pick up and carry so easily, but who has grown into an outstanding young man both in body and mind, and who never ceases to make me smile.

To my grandparents, Dr Narayani Narasimhan and Seshadri Narasimhan, whose achievements and generosity are a shining example which I endeavour to live up to. To Sheina Weston and Roy Weston, whose endurance and fortitude remains a goal to pursue, and a moral standard to never forget.

Finally, to those friends and family that have come and gone, I hope that in some way I may honour you by this dedication.

Abstract

Giant modular sarcomeric proteins such as titin have come under increasing scrutiny for their contributions to diseases such as skeletal and cardiac myopathies. Many variants in titin have been linked with these conditions through computational and experimental studies, and even relatively small, individual missense mutations have been shown to have far-reaching structural consequences. However, assessing the impact of all possible single nucleotide variants (SNVs) in titin is complicated by the immense size of this protein. By combining information on sequence, protein stability, and molecular dynamics features, I propose the generation of a variant classifier able to discriminate between neutral and deleterious variants. I review the disease-associated missense variants in titin reported in the literature and update the centralised resource for information relating to titin variants, TITINdb2, with newly available variants, annotations, structural models, and predictions. From here, I describe the current landscape of understanding of missense variants in titin I compare the coherence of predicted variant impacts with reported impacts in the literature, and investigate the potential of more computationally expensive features, using molecular dynamics, to add information lacking from static representations of these proteins. I found that various

features associated with the movement of the protein domain backbone in simulation were strongly associated with the reported disease-associated missense variants in titin, but that these same features did not differentiate between strongly and weakly stability-altering variants. Furthermore, these results were consistent with machine learning models constructed to predict the pathogenicity or thermostability change associated with titin missense variants; features added to these classifiers from molecular dynamics and deep mutational scanning data improved their ability to discriminate between reported pathogenic and common population variants in titin over classifiers using sequence and structural data alone, but did not facilitate a reliable differentiation between different classes of thermostability-altering variants. These results reiterate the importance of greater diversity in information sources, including dynamical data, for the prediction of pathogenicity in missense variants, while also emphasising the future work required for the accurate determination of thermostability change, particularly for small missense variants.

Contents

Glossary and Abbreviations **1**

Foreword to the Thesis **7**

1 Missense Variants in the Giant Protein Titin **11**

1.1 Synopsis 11

1.2 Introduction 11

1.2.1 Properties of Titin 12

1.2.2 Classification of Skeletal Myopathies 15

1.2.3 Classification of Cardiac Myopathies 16

1.2.4 Truncating Variants and Myopathy 18

1.3 Prioritisation of Variants 21

1.3.1 Determining Pathogenicity of Variants 21

1.3.2 Next-Generation Sequencing and Linkage Studies 24

1.3.3 Experimental Characterisation 25

1.3.4 Expression in Model Organisms 27

1.4 Variant Impact by Sarcomere Region 28

1.4.1 Z-disc 28

1.4.2 I-band 31

1.4.3 N2B 35

1.4.4 N2A 36

1.4.5 A-band 38

1.4.6 M-band . . . 43

1.5 Location-Function Relationships . . . 47

1.5.1 Isoforms . . . 48

1.5.2 Shared Structural Positions . . . 50

1.5.3 HCM/DCM Phenotype . . . 51

1.6 Implications for Diagnosis . . . 52

1.7 Conclusion . . . 54

2 TITINdb2 and Annotating Missense Variants 79

2.1 Synopsis . . . 79

2.2 Introduction . . . 79

2.3 TITINdb Application Notes . . . 81

2.3.1 Updated Features and Data . . . 81

2.3.2 Applications . . . 85

2.4 Methods . . . 87

2.4.1 Data Collection . . . 87

2.4.2 Extraction of Predicted Domain Structures from AlphaFoldDB . . . 89

2.4.3 Validation and Comparison of Model Structures . . . 91

2.4.4 Annotation of Missense Variants . . . 91

2.4.5 Definition of Structural Regions of Domains . . . 92

2.5 Results . . . 94

2.5.1 AlphaFold2 Model Quality Assessment . . . 94

2.5.2 Interface Predictions from AlphaFold2 tandems . . . 96

2.5.3 Properties of Characterised SAVs 98

2.6 Case Studies and Applications 101

2.6.1 Congenital myopathy-associated variants in the A-band 101

2.6.2 A Disease-Associated, Interface-buried Residue in the I-band . . . 105

2.6.3 Prioritisation of Variants from High-Throughput Pipelines 107

2.6.4 Investigating the Mutational Landscape of Titin with Saturation Mutagenesis 108

2.7 Conclusion 113

3 Analysis of Titin Missense Variants using Atomistic Molecular Dynamics 121

3.1 Synopsis 121

3.2 Introduction 121

3.3 Methods 124

3.3.1 Literature and Population Variant Data Sets 124

3.3.2 Thermostability of Variants Data Set 125

3.3.3 Modelling of Domains without Experimental Structure 125

3.3.4 Molecular Dynamics Simulations and Parameters 126

3.3.5 External computing resources 132

3.3.6 Feature Extraction 132

3.3.7 Principal Component Trajectory Analysis 134

3.3.8 Structural Alphabet Encoding and Mutual Information 135

3.4 Simulation of Pathogenic vs Non-Pathogenic Variant Set 136

3.4.1 Features indicative of Pathogenicity 136

3.4.2 PCA Cluster Analysis . . . 141

3.5 Simulation of Thermostability Variant Set . . . 148

3.5.1 Features indicative of Thermostability . . . 148

3.5.2 Case Study: Fn3-70 as a Model System . . . 153

3.6 Discussion . . . 156

3.6.1 Data from MD simulations augment sequence and structural annotation of variants . . . 156

3.6.2 Variants in Ig-169 do not perturb the system *in silico* compared to wild-type . . . 159

3.6.3 Primitive dynamical features do not recapitulate domain thermostability alone . . . 161

3.6.4 Future considerations relating to MD simulation-derived features . 162

3.7 Conclusion . . . 164

4 Prediction of Variant Impacts of Titin SAVs using a combination of Sequence, Structural, and Dynamical Features 173

4.1 Synopsis . . . 173

4.2 Introduction . . . 173

4.3 Methods . . . 176

4.3.1 Data Sets . . . 176

4.3.2 Machine Learning Architectures . . . 178

4.3.3 Hyperparameter Tuning . . . 181

4.3.4 Feature Selection . . . 183

4.4 Prediction of Damaging / Neutral Titin Variant Impacts . . . 185

4.5 Prediction of Titin Domain Stability Changes . . . 196

4.6 Discussion 205

4.6.1 The Importance of Features in Different Prediction Contexts 205

4.6.2 The Contribution of Dynamical Features to Classification of Titin Variants 208

4.6.3 Consideration of Current and Future Machine Learning Strategies . 211

4.7 Conclusion 213

Discussion and Conclusions **219**

References **227**

List of Figures

1 Arrangement of Titin in the Sarcomere 14

2 Major Isoforms of Titin in Adults 17

3 Domain- and Molecular-level arrangement of Titin 19

4 Reported Pathogenic variants of *TTN* in the Z-disc 29

5 Reported Pathogenic variants of *TTN* in the I-band Ig tandem regions . . . 32

6 Reported Pathogenic variants of *TTN* in the I-band N2A/N2B-specific regions 35

7 Reported Pathogenic variants of *TTN* in the A-band 38

8 Reported Pathogenic variants of *TTN* in the M-band 44

9 Reported Disease-Associated Missense Variant Labelling Flowchart 55

10 Screen Capture of TITINdb2 developer build, domain Fn3-119 82

11 Overview of domain model extraction from AlphaFold2 proteome 88

12 Definition of interface-buried residues in TITINdb2 93

13 Experimental structure and predicted model quality metrics for structural classes in TITINdb2 95

14 Structural schematic of domain-domain interfaces in titin 96

15 Positions of interface-buried residues 99

16 Distributions of bioinformatics annotations across titin SAVs in different database populations 102

17 Locations of damaging variants in Fn3-119 and homologous domains . . . 103

18 Locations of predicted domain-domain interface residues in Ig-19 105

19 Saturation mutagenesis heat map of Rhapsody predictions for two disease-associated Ig domains 109

20 Aligned saturation mutagenesis heat map of average per-position Rhapsody predictions for all Ig domains in titin. 112

21 Distributions of dynamical features calculated from trajectories, versus class 137

22 Relative RMSF and relative RMSD comparison of variant systems, by class 138

23 Relative RMSF of variant systems against wild-type PSIC and SASA, by class 140

24 Visual representations of trajectory separation between typical Literature and Population variants 143

25 Separation Distances between WT and mutant trajectories for variants, by Database 146

26 Separation Distances between WT and mutant trajectories for Fn3-119 and Ig-169 variants, by Database 147

27 Feature differences between WT and mutant trajectories for variants, by Class 149

28 Separation of thermostability data set by descriptive features 151

29 Fn3-70 variants with normalised features used for categorisation of thermostability variants 154

30 PCA of backbone coordinates of Fn3-70 variants versus WT 157

31 Calculated feature importances for Random Forest classifiers 188

32 Distribution of predictions made by logistic regression classifiers trained on representative features . . . 190

33 ROC curve for Titin variant classifiers with different architectures . . . 192

34 ROC curve for TitinRF classifiers and select pathogenicity classifiers in common use . . . 194

35 Plots of ΔT_m for variants in thermostability data set against their predicted pathogenicity . . . 197

36 ROC curves for Titin thermostability classifiers with different architectures and feature sets . . . 199

37 ROC curves for Titin classifiers with different architectures and training data, on test set with known pathogenicity and thermostability . . . 204

List of Tables

1 Allele frequencies of reported pathogenic single amino acid variants in titin in populations . . . 57

2 Single amino acid variants in titin reported in the literature . . . 63

3 Experimentally solved Titin Domain structures in Protein Data Bank . . . 115

4 Pathogenicity and stability predictors implemented in TITINdb2 . . . 118

5 Single amino acid variants in titin reported in the literature . . . 167

6 Experimentally characterised variant systems and their reported melting temperatures . . . 169

7 Variants in combined Test Set . . . 178

8 Features used in Classifier Development . . . 215

Glossary and Abbreviations

General Terms	
Compound heterozygous	Different variants in the same gene inherited together, either on the same (*in cis*) or on different (*in trans*) chromosomes
FINmaj	Non-frameshift indel mutation in Ig-169 (EVTW>VKEK)
Fn3	Fibronectin Type-III (domain)
Ig	Immunoglobulin (domain)
Missense	Variants involving the substitution of a single amino acid residue with another within a protein-coding region
SAV	Single Amino acid Variant
SNV	Single Nucleotide Variant
SNP	Single Nucleotide Polymorphism
TTNtv	Titin truncating variant
WT	Wild-type
***TTN* Isoforms**	
FCT	Foetal Cardiac Titin
IC	Inferred Complete titin isoform, all exons
N2A	Predominant titin isoform in skeletal muscle
N2AB	Canonical titin isoform, largest in humans
N2B	Predominant titin isoform in cardiac muscle
Disease and Pathology	
ACM	Arrhythmogenic Cardiomyopathy
ARVC	Arrhythmogenic Right Ventricular Cardiomyopathy
CCD	Central Core Disease
CFTD	Congenital Fibre-Type Disproportion
CM	Cardiomyopathy
CNM	Centronuclear Myopathy
DCM	Dilated Cardiomyopathy

HCM	Hypertrophic Cardiomyopathy
HMERF	Hereditary Myopathy with Early Respiratory Failure
LGMD	Limb-Girdle Muscular Dystrophy
LGMD2J	Limb-Girdle Muscular Dystrophy, type 2J
LVNC	Left Ventricular Non-Compaction
MD	Muscular Dystrophy
MFM	Myofibrillar Myopathy
MmD	Multi-Minicore Disease
NMD	Neuromuscular Disease
PPCM	Peripartum Cardiomyopathy
RCM	Restrictive Cardiomyopathy
T1P	Type 1 Predominance
TMD	Tibial Muscular Dystrophy
Databases and Sequencing	
1KGP	1000 Genomes Project
gnomAD	Genome Aggregation Database
MAF	Minor Allele Frequency
NGS	Next-Generation Sequencing
PDB	Protein Data Bank
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
Proteins and Genes	
ACTN2	α-actinin 2
BglB	β-glucosidase B
CAPN3	Calpain-3
FHL2	Four-and-a-half LIM domains protein 2
OBSCN	Obscurin
PKA	Protein Kinase A
PKG	Protein Kinase G

TCAP	Telethonin
TTN	Titin
Experimental Methods	
Y2H	Yeast-2-Hybrid
AFM	Atomic Force Microscopy
CD	Circular Dichroism
co-IP	Co-immunoprecipitation
DMS	Deep Mutational Scanning
DSF	Differential Scanning Fluorimetry
FRET	Förster Resonance Energy Transfer
GST	Glutathione S-transferase (pulldown assay)
HSQC	Heteronuclear Single Quantum Correlation (NMR)
Iso/PE	Isoprenaline/phenylephrine adrenergic challenge
ITC	Isothermal Titration Calorimetry
MRI	Magnetic Resonance Imaging
NMR	Nuclear Magnetic Resonance
PAGE	Polyacrylamide Gel Electrophoresis
TUBE	Tandem Ubiquitin Binding Entity (assay)
Variant Analysis	
Delta (Δ)	Change in feature between wild-type and mutant
DIS	Domain Interface Score
KF	Kidera Factor
K_u	Equilibrium Constant of Unfolding
PC1	Principal Component 1
PC2	Principal Component 2
PCA	Principal Component Analysis
PSIC	Position-Specific Independent Counts
Q(SASA)	Quotient Solvent-Accessible Surface Area
Rg	Radius of Gyration
RMSD	Root Mean Square Deviation

RMSF	Root Mean Square Fluctuation
SASA	Solvent-Accessible Surface Area
SSE	Secondary Structure Elements
T_m	Melting Temperature
Molecular Dynamics	
MD	Molecular Dynamics
NPT	Constant Number of Particles, Pressure, Temperature
NVT	Constant Number of Particles, Volume, Temperature
P-LINCS	Parallel LINear Constraint Solver
PBCs	Periodic Boundary Conditions
PME	Particle-Mesh Ewald
SPC-216	Simple Point Charge 216 (water model)
TIP3P	Transferable Intermolecular Potential 3P (water model)
Machine Learning	
DL	Deep Learning
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
ML	Machine Learning
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
Funding Bodies and Hardware	
BRC	Biomedical Research Centre
CSD3	Cambridge Service for Data-Driven Discovery
EPSRC	Engineering and Physical Sciences Research Council
HecBioSim	UK High-End Computing Consortium for Biomolecular Simulation
HPC	High-Performance Computing
JADE2	Joint Academic Data science Endeavour 2

NIHR	National Institute for Health and Care Research

Feature Categories

AASM	Amino Acid Substitution Matrix
DYN	Dynamics-derived (legacy)
MD-MOB	MD-derived dynamical
MD-STR	MD-derived structural
SEQ	Sequence-derived
ST-STR	Structure-derived

Foreword to the Thesis

The capability of computational methods to leverage ever-increasing quantities of clean, available data has become integral to our understanding of the natural world; especially in recent decades, they have represented the driving force behind significant shifts in myriad fields. Within the realms of biochemistry and molecular biology, each year brings with it new databases, tools, and other resources that transform what was once hundreds of man-hours of work into automated, transferable, and replicable approaches. However, despite such advances, manual curation of data remains the gold standard in science; computational methods, while greatly expediting existing approaches, have not yet reached such a level of quality as to obviate the need for human interpretation. For the study of proteins and other biomolecular entities, despite the ability of "Big Data" type approaches to produce excellent results when considering the questions in their entirety, they often fall short when considering more specific problems, where the minutiae of individual proteins or protein domains is considered.

In this work, I set out with a focus on a very small subset of similar proteins, the giant sarcomeric proteins; these are proteins of immense size, encoded by single genes, which adopt a modular domain architecture that reflects their purpose in the sarcomere, the contractile unit of muscle tissue. After deliberation, I focused solely on titin, both the largest and most well-studied of the giant proteins; owing to limitations in the validation of disease-associated variants in other proteins, such as obscurin, titin represented the best combination of attributes for detailed analysis of this nature. Aside from the considerable body of computational and experimental analysis already preformed on this protein, titin is not only the largest protein in humans but also highly modular, with substantial similarity within its composition of folded domains. Thus, titin is uniquely suited to this purpose, both in terms of the quantity of data available for this one protein alone, and in terms of the volume of unexplored variation in its mutational landscape to which these conclusions may be applied.

I begin in Chapter 1 with an introduction to the protein titin, in the form of a review of disease-associated variants reported in the literature; as the largest protein in humans, there are a commensurately large number of variants put forward as causative of a variety of inherited conditions. I focus on missense variants, or variants involving the substitution of a single amino acid residue in a protein-coding sequence region with another, as these are not only the most common variation in titin but also in many ways the least well-studied; I also limit the disease associations considered to the myopathies, or diseases of the myocardium, with which titin missense variants are intimately connected. From here, in Chapter 2 I add the data I have collected to TITINdb, the online resource for variation and structural localisation in titin, and discuss the annotation of titin variants, both disease-associated and those found in population genetic databases, and the performance of models designed to predict variant impacts such as pathogenicity or stability change. In Chapter 3, I extend the computational characterisation of titin variants using molecular dynamics (MD) derived features, which are not used in any standard variant impact predictors, and demonstrate the information they can contribute to more accurate annotation of titin variants. Finally, in Chapter 4, I bring together each of these approaches to develop my own models for annotation of variant impacts, and discuss the importance of different features from different data in building a cohesive picture of titin variant consequences.

Hypothesis

Structural and biophysical features associated with known pathogenic missense variants in titin, including features derived from molecular dynamics simulations of titin domains, are predictive of pathogenicity and other variant impacts of titin missense variants. The interplay of these features with commonly used sequence and evolutionary features will improve the prediction of variant impacts of titin missense variants over standard variant impact prediction tools for general use.

Aims

- To collect together all reported pathogenic titin missense variants in the scientific literature, and compare them against neutral variants using existing variant annotation tools.
- To determine features derived from molecular dynamics simulations that are associated with known pathogenicity and thermostability impacts of titin missense variants.
- To develop classification tools based on machine learning architectures that can accurately predict pathogenicity and thermostability impacts of titin missense variants, using existing and novel features associated with these variant annotations.

1 Missense Variants in the Giant Protein Titin

1.1 Synopsis

- In this chapter, I introduce the key physiological and biological properties of titin, and how these underly the problem of small variant impact assessment in this protein.
- I present a compilation of reported titin missense variants in the literature – those that result in a single amino-acid substitution in coding regions – and discuss them in light of the sheer size of titin, their variant position within the sarcomere and their domain, the structural, pathological, and biophysical characteristics that define them, and the methods used for their characterisation.
- I discuss the current knowledge and integration of the multiple fields that have contributed to the understanding of titin-related pathology and offer suggestions as to how these concurrent methodologies may aid the further development in our understanding of titin.

1.2 Introduction

The giant protein titin (alias connectin[1]) is the largest single-chain protein in humans[2–6]. While the best-known components of the sarcomere, the basic contractile unit of striated muscle, are those that make up the thin and thick filaments[7], titin is a crucial element from one end of the sarcomere to the other, to the point of being described as the "third filament"[8–10]. As a molecular ruler, titin determines the length of the sarcomere and those of its various sub-regions[11, 12]. As a molecular spring, titin's I-band region provides passive force to the sarcomere, preventing the overextension of the muscle[13–15]. As a stress sensor, titin measures and communicates changes in the forces impacting the sarcomere to downstream signalling pathways and effecting transcriptional and regulatory changes[16, 17].

As a molecular scaffold, titin provides binding points at its diverse signalling hubs, playing a role in both local signalling cascades and the greater cytoskeletal architecture[18–20]. As a dynamic cellular component, it modulates muscle properties on both long and short timescales, with a range of post-translational modifications that control transient muscle stiffness[17, 21], and longer-term control of stiffness through changing isoform ratios in the make-up of the muscles[6, 19].

With such an integral role in cardiac and skeletal muscles, it is unsurprising that damage to the titin protein in the form of mutations is associated with the development of muscle disease[22–24]. A panoply of both skeletal and cardiac myopathies have been linked to variants in titin, and the burden that these conditions place on both individuals and health systems worldwide has led to considerable interest in understanding the pathomechanisms that connect the protein to the phenotype. However, while more progress has been made in identifying highly damaging truncating and frameshift variants[22, 25, 26], there have been far fewer missense variants identified and characterised, despite being more common. I present here an overview of the current knowledge regarding missense variants in titin and their associated pathologies, and how they reflect their localisation in titin domains, in sarcomeric regions, and in specific tissues. I will also discuss the different approaches that have been used to identify pathogenic missense variants - variants that cause or increase the risk of developing a particular disease - and the implications that the current body of evidence has for clinical diagnosis.

1.2.1 Properties of Titin

Titin has been described as the molecular backbone that forms the foundation of the sarcomere, the contractile unit of striated muscle tissue[27] (Figure 1). Each molecule extends over 1 micrometre in length, across half of the sarcomere, overlapping with antiparallel titin filaments at each end. Titin is present in all sub-regions of the sarcomere: the Z-disc, where thin filaments are cross-linked by alpha-actinins; the I-band, where the thin filaments progress

into the sarcomere; the A-band, where the motor domains of the bundled myosin molecules in the thick filament are found; and the M-band, where thick filaments are cross-linked at the centre of the sarcomere. Titin, while extremely large, contains only three types of domains, these being I-set immunoglobulin (Ig) domains, fibronectin type-III (Fn3) domains, and the single protein kinase domain[6, 28]. The many Ig and Fn3 domains of titin are not only extremely similar in terms of structure but often of sequence as well, with invariant residues consistently located at the same structural locations. The Ig domains are classified as the I-set type, which have many defining structural features and residues at key positions[29].

The various isoforms of titin are all expressed from the 364-exon TTN gene, whose full protein product encompasses 35,991 amino acid residues, arranged into 169 Ig domains, 132 Fn3 domains, the kinase domain, and extensive unstructured repeat regions (see Figure 3). The inferred complete (IC) isoform[30], a theoretical protein containing all exons in the titin gene, does not exist in nature, and as such all titin proteins in the body exclude some of the 364 exons present. The three main isoforms of titin in adults are the N2BA, N2A, and N2B giant isoforms; at the protein level, the difference between these giant isoforms is found in the I-band region, where the TTN gene encodes three regions of tandem Ig domains separated by varying lengths of unstructured repeats and signalling hub regions, which are differentially spliced to make the different isoforms (see Figure 2). Ig25 and Ig26 are spliced out from the giant isoforms N2BA, N2A, and N2B; these are expressed in the novel exon isoforms novex-1 and novex-2, respectively. Finally, the "tiny titin", novex-3, includes only 46 exons[31]; however, one of these, exon 48, is not found in any of the giant isoforms. While these are the most well-established isoforms of titin, many others are present and continue to be discovered in different physiological contexts; two examples include foetal cardiac titin (FCT)[32], which is found in neonatal and foetal heart muscle, and the recently discovered *Cronos* titin isoform[33], which is up-regulated in developing cardiomyocytes.

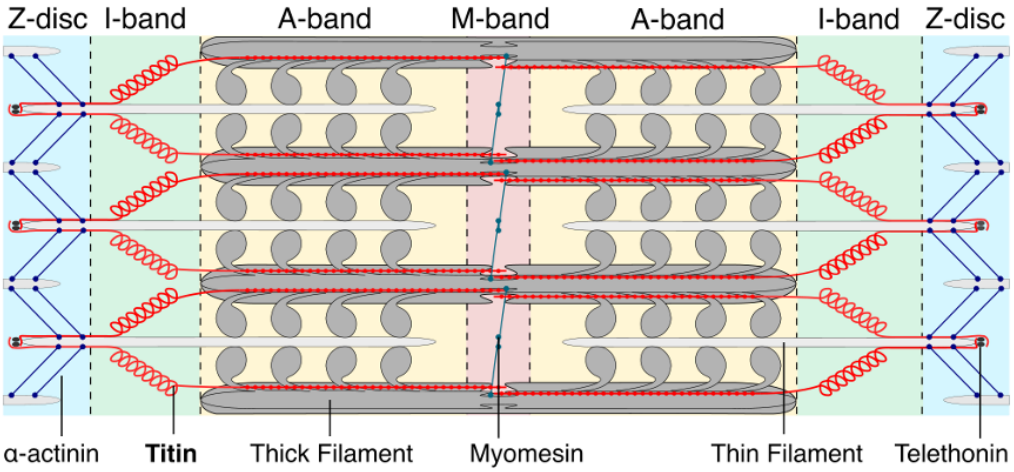


Figure 1: Arrangement of Titin in the Sarcomere

Schematic representing titin (red), its key structural binding partners, and the other major components of the sarcomere. The core of the sarcomere is composed of the myosin thick filaments and actin thin filaments, which together undertake the cross-bridge cycle that allows for contraction and relaxation of the sarcomere, and thus striated muscle. Two antiparallel titin molecules overlap at the centre of the sarcomere, the M-band; they are bound to the thick filament in the A-band, and to the thin filament at the Z-disc, through their interactions with alpha-actinin 2. The titin molecules extend through the Z-disc into the adjacent sarcomere, where they are anchored by telethonin. The unbound I-band region of titin is elastic and provides passive force to the sarcomere, preventing the cross-bridge cycle from overextending the muscle; different isoforms of titin have different levels of elasticity.

1.2.2 Classification of Skeletal Myopathies

The concept of the "titinopathy", a disease of the muscle linked with damaging variants in titin, is one that has become increasingly relevant to the diagnosis of myopathy as more and more damaging variants are identified. Titin variants have been associated with a broad range of neuromuscular pathologies with distinct phenotypic characteristics[34], consistent with the size and number of possible disruptions that may be linked back to the titin molecule. More recently, it has been suggested that titinopathies, rather than being a collection of distinct, albeit partially overlapping disorders, in fact represent a continuum of phenotypes associated with a single condition. This interpretation has yet to be implemented in any clinical setting, but several authors have provided evidence in support of this framework[35–37].

The skeletal muscle diseases linked to titin can broadly be split into muscular dystrophies and congenital myopathies. The former refers to the gradual, progressive degeneration of muscle fibres, leading to worsening health, while the latter refers to those conditions that are associated with early-onset, characteristic damage to the muscle[38], although sometimes the distinction is blurred by slowly progressing myopathies[39]. The most well-documented associations with missense variants in titin are those with muscular dystrophies. Tibial muscular dystrophy (TMD)[40] has been associated with multiple variants in the M-band, particularly the Ig169 domain[41]; homozygous and compound heterozygous variants in these exons are associated with additional, more severe phenotypes, including limb-girdle muscular dystrophy type 2J (LGMD2J)[42]. Another well-established association with titin is that of hereditary myopathy with early respiratory failure (HMERF)[43], which has been shown to result from variants in the Fn3-119 domain[44]. Several titin variants associated with HMERF were originally identified in patients diagnosed with myofibrillar myopathy (MFM), a muscular dystrophy with considerable phenotypic overlap with HMERF[45, 46].

Congenital myopathy phenotypes associated with titin variants include, but are not limited to, centronuclear myopathy (CNM)[47], congenital fibre-type disproportion (CFTD)[35], multi-minicore disease (MmD)[48], central core disease (CCD)[35], type 1 predominance

(T1P)[35], and childhood-juvenile onset Emery-Dreifuss–like phenotype without cardiomyopathy[49]. Typically, these conditions present with significant clinical overlap; diagnosis of individual conditions is mainly achieved by the identification of characteristic sub-cellular structures in the affected muscle fibres. For example, MmD and CCD are collectively known as "core myopathies" for the presence of cores, or distinctive circular structures, within the sarcomere[48]. There can be a considerable heterogeneity in severity and pathological consequences of the conditions even under the same diagnosis; other proteins that are linked with these same conditions may also modulate the clinical presentation.

1.2.3 Classification of Cardiac Myopathies

Cardiomyopathies, or diseases primarily affecting the myocardium, are typically considered as secondary to skeletal myopathies in the context of titin missense variants. While a considerable body of work has demonstrated the strength of the association between truncating titin variants and conditions such as dilated cardiomyopathy (DCM)[50–52], there are few associations between missense variants and cardiomyopathies that have been widespread or definitive enough to attract the same attention as the congenital titinopathies. Even so, a great many missense variants have been put forward as potentially causative of these conditions, and they continue to be actively studied.

The classification of cardiomyopathies remains a subject of considerable debate[53–55]. At the simplest level, genetically inherited cardiomyopathies are partitioned into DCM, hypertrophic cardiomyopathy (HCM), restrictive cardiomyopathy (RCM) and arrhythmogenic cardiomyopathy (ACM). Each of these is so named for the observed consequences to the heart structure: HCM, the most common cardiomyopathy, causes a thickening of the heart muscle walls, and consequently a constriction of the heart chambers, notably the ventricles. DCM, the second-most common, causes an enlargement of the heart, where the heart chambers expand as the muscle walls stretch. RCM does not involve stretching or compressing of the heart muscle, but rather a loss of elasticity and a rigidity of the heart muscle

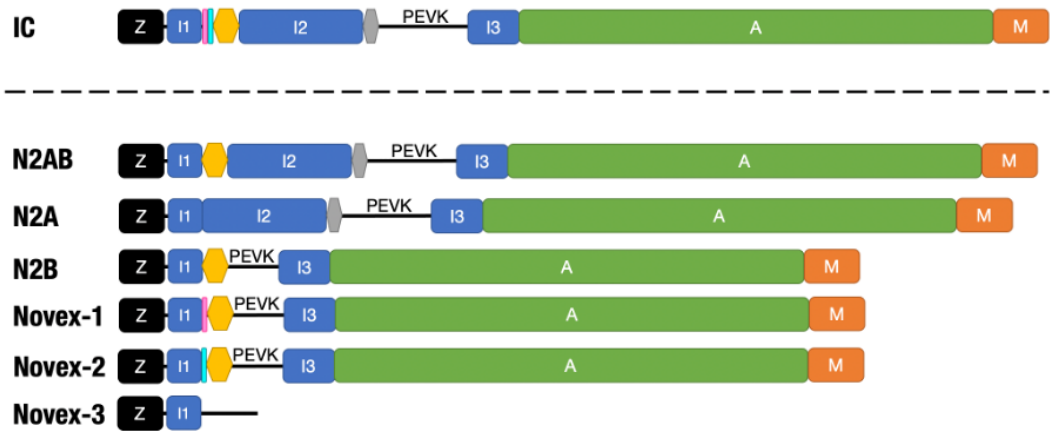


Figure 2: Major Isoforms of Titin in Adults

Major isoforms of titin in adults. The inferred complete isoform, titin-IC, contains all exons in the titin gene. The key differences in exon composition for the giant isoforms N2AB (canonical isoform), N2A (predominant in skeletal muscle), N2B (predominant in cardiac muscle), novex-1 and novex-2 are found in the I-band (blue). All giant isoforms contain a region of tandem Ig domains, from domains Ig10 to Ig24; Ig25 (magenta) and Ig26 (cyan) are spliced out from the isoforms N2BA, N2A, and N2B, and are only expressed in the novel exon isoforms novex-1 and novex-2, respectively. The proximal Ig domains are followed by the "N2B region" (yellow): domains Ig27 and Ig28, connected by a long linker of almost 30 residues, precede a unique, unstructured sequence of around 600 residues in length, followed by Ig29 and Ig30. This N2B region is absent in the predominant skeletal isoform, N2A; instead, what follows Ig30 is a second Ig tandem region, from domains Ig31 to Ig82, found in the N2BA and N2A isoforms only. Also absent from the N2B isoform is the subsequent region, from Ig83 to Ig86, known as the "N2A region" (grey), which also includes a long, 112-residue insertion sequence between Ig83 and Ig84. The PEVK sequence is an elastic, repeating, unstructured region named for the abundance of proline, glutamate, valine, and lysine residues in its constitution. In the N2BA isoform, the PEVK sequence contains almost 2000 amino acids, spanning 31 repeats; by contrast, in the N2B isoform, the PEVK sequence is considerably shorter, with only 186 residues. Finally, in all giant isoforms, a third and final region of closely linked Ig domains follows the PEVK region, from Ig87 to Ig107. After this distal Ig tandem region, the titin molecule moves into the A-band (green).

walls, restricting the cardiac cycle. ACM[56] does not typically involve a dysfunction of the cardiac blood pump, but rather is characterised by arrhythmia; it is typically diagnosed as arrhythmogenic right ventricular cardiomyopathy (ARVC), although cases of ACM involving the left ventricle have also been reported[57, 58].

While the above classification covers most cardiomyopathy-related diagnoses, more recent observations have identified cardiomyopathies which are considered separate pathologies. Most relevant to titin is left ventricular non-compaction (LVNC)[59], a developmental disorder affecting the heart muscle associated with the left ventricle, leading to a dysfunctional, characteristically "spongy" myocardium. Furthermore, within the different subclassifications of cardiomyopathy there is a considerable heterogeneity in consequences, from fatal pathology to asymptomatic, and is likely to be heavily influenced by non-genetic factors. Attempts to refine the classification of cardiomyopathies have been conducted in the past, including the MOGE(S) classification system[60, 61], where individual cardiomyopathies are given a multi-character code based on their morpho-functional properties, organ system involvement, genetic history, genetic aetiology, and optionally the stage of progression. However, to my knowledge this system has seen only limited uptake in a clinical setting, and will not be used when describing cardiomyopathies here.

1.2.4 Truncating Variants and Myopathy

While the scope of this work does not include the contributions of truncating, frameshift, splice-site and nonsense variants to disease, it is worth establishing their significance in genetic contributions to muscle disease, as a lens through which to view those of missense variants. Truncating titin variants, commonly abbreviated as TTNtvs, are thought to be the most important single genetic contributor to DCM[25, 26, 64–66], as well as similar conditions such as peripartum cardiomyopathy (PPCM)[67, 68]. They have also been implicated in congenital skeletal myopathies both with[35, 48, 69, 70] and without cardiac involvement[35, 42, 47]. While the strength of evidence in support of this association is considerable[50–52],

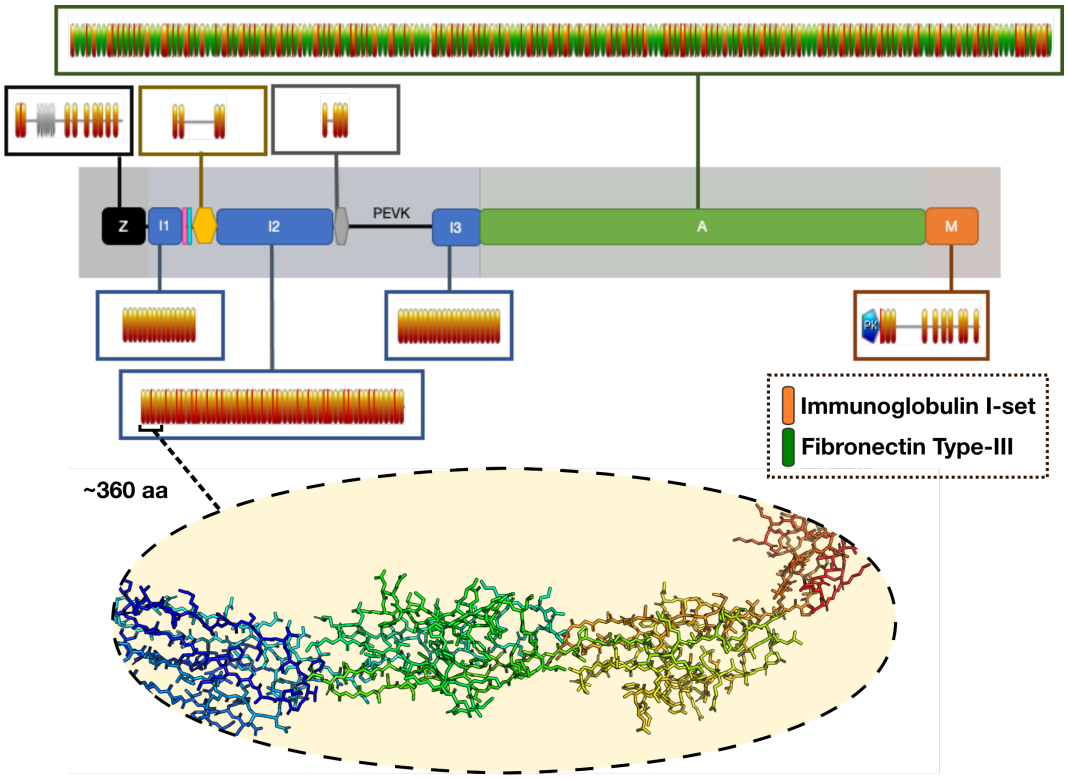


Figure 3: Domain- and Molecular-level arrangement of Titin

Domain- and molecular-level view of the inferred complete isoform of titin. Arrangement of folded globular domains in the main organisational regions of titin is represented schematically (orange: Immunoglobulin I-set (Ig); green: Fibronectin Type-III (Fn3); Titin Kinase represented by blue pentagon). Titin regional schematic (centre) is arranged according to the colour scheme described for Figure 2. Inset, partial molecular-level view of domains Ig31-Ig34 visualised in PyMol[62]. Construction of schematic representation of domains performed using PROSITE[63].

the incidence of TTNtvs in the general population is around 2-3%, considerably higher than the incidence of DCM[69]; this demonstrates that the presence of a TTNtv alone does not necessarily lead to disease, DCM or otherwise.

That a TTNtv, which represents a significant alteration to a vital sarcomeric component, can be found in healthy individuals with no apparent ill effects, has important implications for the mechanisms of pathogenicity not only of TTNtvs but of missense variants as well. TTNtvs in only one allele of the titin gene, or those found in exons that are spliced out from most isoforms found in the myocardium, may permit the carriers to escape the consequences of the variant; even if one isoform of titin integrated into the sarcomere is malformed, a preponderance of healthy isoforms may rescue the functional integrity of the muscle. Consequently, the pathomechanisms of TTNtvs have tended to focus more on haploinsufficiency than on complete loss of function, where the loss of correct-length titin protein results in the destabilisation of the sarcomere. However, recent papers focusing on the pathomechanisms of TTNtvs relating to DCM have presented alternative methods by which these variants may cause disease, focusing on the variant titin proteins themselves.

Two notable papers[26, 66] that have recently detected and quantified the presence of TTNtvs in the hearts of DCM patients offer an insight into the pathomechanisms that may result from TTN variations, both truncating and missense. The first used these quantitative methods to compare the abundance of full-length and truncated titin in DCM explant hearts, demonstrating both the decrease in full-length titin, indicative of haploinsufficiency, as well as the active presence of truncated titin, indicating the integration of both forms of titin in the muscles[66]. The second also found the discrepancy between levels of titin protein in DCM hearts with TTNtvs, but also demonstrated the aggregation of unfolded, truncated titins within the cells, rather than incorporation into the sarcomere, which the authors suggest could result in a proteotoxic response[26]. Thus haploinsufficiency, proteotoxicity, and integration of malformed titins together form a tricast of potential mechanisms that all may influence the consequences, not only of TTNtvs, but also of missense variants. Indeed, recent work

has implicated a proteotoxic response as a consequence of a missense variant in Z-disc titin, albeit as a result of unbound cytoplasmic telethonin[71].

1.3 Prioritisation of Variants

For a variant to be associated with a given pathology, it must first be identified in an individual with that condition. With the advent of next-generation sequencing (NGS) technologies, it is now possible to perform large-scale genomic sequencing on sizable cohorts of patients, rather than single genes or single individuals, greatly increasing the capacity for disease-associated variants to be identified. However, while many such variants are identified from genetic testing of individuals, only a small proportion of these are then proposed as such in the scientific literature, and fewer still are considered as causative of disease with any degree of certainty.

The issue of validation, or how those variants identified within patients can be confirmed to play a causative role in the development of disease, is not one that is unique to titin. However, the size of titin as a protein, and the corresponding number of possible variants that can exist exacerbates this issue considerably. Even so, there exist many techniques that can be used to gather evidence in support of the pathogenicity of a variant, or to investigate the functional consequences of its inclusion in titin.

1.3.1 Determining Pathogenicity of Variants

The ascertainment of pathogenicity of protein variants is never simple, and much guidance has been produced to aid researchers in their classification[53], but the problems that plague the annotation of variants in smaller proteins is magnified greatly when considering those in the giant titin. For example, rare variants – typically defined as those with a population minor allele frequency (MAF) below 1% – are considered first when attempting to identify

likely candidate genes to link to a given condition. However, titin is significantly larger than any other single protein in the human proteome; consequently, the likelihood that any given individual will carry a rare variant in titin is all but certain[72], many such variants will be present in healthy controls[69], and the majority of these variants will be missense variants of unknown significance[73]. Thus, any attempt to associate titin with a muscle disease will always produce many potential candidate variants - both common and rare - which must be filtered based on previous knowledge of likely disease-causing variants.

The danger of considering variants as potentially pathogenic based solely on their rarity has been discussed by many authors[30]. Manrai *et al.* (2016)[74] notably identified several missense variants in sarcomeric proteins associated with hypertrophic cardiomyopathy which, while rare when considering the genetics of the whole population, were far more common in certain ethnic sub-populations, greatly decreasing the likelihood that they were pathogenic. In titin specifically, a study of truncating variants in DCM patients[75] found no enrichment of truncated titin protein in the hearts of patients with African ancestry, unlike the patients with European ancestry who showed a strong association. Moreover, several of the missense variants in titin reported as pathogenic in the literature are present in unaffected individuals in the general population; notably, the ARVC-associated variant H10092Y is present in about 1 in 250 individuals in the gnomAD databases 2.1.1 and 3.1.2, increasing to around 1 in 100 individuals in the Finnish sub-populations (Table 1). This does not necessarily preclude their role in disease, but must be considered in any assessment of their pathogenicity.

Another issue related to the number of variants in titin is that the true causative variant may be overlooked in favour of one with a perceived greater likelihood of pathogenicity. For example, the rare missense variant R34091W, identified in the protein kinase domain of titin in 2 Swedish families[76], was proposed as causative of hereditary myopathy with early respiratory failure (HMERF) based on its reduction of binding affinity of the kinase domain for nbr1[77]. Subsequent studies identified a strong association between variants in a different part of titin with the same condition, and a re-evaluation of the original patients

found that they too had variants in this region[78]. Thus, R34091W is no longer considered a disease-causing variant.

The patterns of inheritance in diseases associated with titin variants are complicated by the multifaceted aetiology of these diseases[79]. In general, disease-associated titin variants can be grouped into three main patterns of inheritance: heterozygous, homozygous, and compound heterozygous. Heterozygous variants present with dominant inheritance and are causative of disease without any other genetic contribution. Homozygous variants present with recessive inheritance and require inheritance in both alleles to cause disease. Compound heterozygous variants are those that are causative for disease when inherited with another compound heterozygous variant in the other allele. Further complicating this picture is the variable penetrance of different variants and their interaction with environmental effects; some variants may not be disease-causing alone but may exacerbate the phenotype of a disease caused by another variant or external factor[80].

Beyond the difficulties with the assessment of genetic data, the environmental component of the diseases in question must be emphasised. Myopathies of the cardiac and skeletal muscles, while they possess a heritable component, are strongly influenced by external and environmental factors, including obesity[81] and level of physical activity[82]. These factors can modulate the disease or the specific pathology alongside genetic factors, and the delineation of the relative contributions of each factor in disease prognosis is in its infancy. Thus, just as many of those with ostensibly damaging variants in titin appear in perfect health, many of those who suffer from the conditions associated with missense variants in titin and the other sarcomeric proteins will not have any clear genetic issues. The purpose of this reiteration is to remind the reader that, while the underlying causes of these conditions cannot be solved through genetic factors alone, understanding the relative risks imposed by different genetic burdens is vital to the protection of long-term health.

1.3.2 Next-Generation Sequencing and Linkage Studies

By far the greatest quantity of known variants are those that have arisen from what will hereafter be referred to as "*in silico*" methodologies – that is to say, those that are proposed from studies of data only. While sequencing is a necessary step in identifying novel variants, NGS and other high-throughput strategies allow for the capture of the full genetic landscape of the individual or cohort. Notably, whole-genome or whole-exome sequencing (WGS/WES) will highlight many more variants than can be related to the pathology in question; for example, healthy individuals who participated in the UK Biobank study carried a median of 1695 rare autosomal variants (MAF < 1%), of which 379 were rare missense variants[83]. This is particularly relevant for titin, as other variants in the genome notwithstanding, there will almost certainly be many variants in the gene itself that do not have any bearing on the individual's condition, given that many such variants are also present in healthy controls[69].

Typically, a range of bioinformatic and computational tools and pipelines will be used to automatically prioritise those variants with the highest likelihood of playing a role in the disease state. These begin with the elimination of common population variants by comparison with known variant databases, such as gnomAD[84, 85], ExAC[86], and 1000 Genomes[87]. If a variant is found to exist in healthy, ethnically matched controls, or has a high population or sub-population incidence in population databases (in other words, common variants with MAF > 1%), it is unlikely to be strongly associated with disease, and can be excluded from further analysis. Additionally, pathogenicity prediction tools can be used to remove from consideration those variants that do not display features indicative of known damaging variants. Various tools exist that allow for all these approaches to coalesce into a single, efficient pipeline, such as ANNOVAR[88], which was used by Begay *et al.* (2015)[89] to filter titin variants in DCM patients to prioritise those most likely to be associated with the condition.

While sequencing alone can prioritise those variants most likely to play a role in pathogenesis, it cannot validate the causal role of any given variant in a pathology alone. The most

common method to support the causative role of a variant in disease is to perform linkage or segregation analysis. Here, a family or body of related individuals both with and without the condition are sequenced, and the pattern of inheritance between the variant of interest and the disease are compared. For segregation analysis, different models of inheritance patterns are fitted to the linkage data to ascertain the link between the condition and variant. Where the variant co-segregates with the disease – present in all affected family members and absent in all unaffected family members – this can be used as evidence of the pathogenicity of the variant. However, segregation analysis is not always possible or unreliable due to availability of family history or variable penetrance of the variant in question.

1.3.3 Experimental Characterisation

The experimental characterisation of variants in titin, hereafter referred to collectively as "*in vitro*" methodologies, involves the attempt to elucidate the biophysical consequences of the variant at the local level, and to then translate these changes to the expected result at the global, organismal level. In general, this requires the establishment of a parameter, such as thermal or mechanical stability, that can be associated with both the wild-type and variant titins, and thus quantitatively compared between the two. Given the sheer size of the titin protein, expressing the entire protein *in vitro* is impossible; typically, only the domain in which the variant is found, or a tandem including the domains either side of it, are expressed. The wild-type and variant constructs can thus be compared independently of the remainder of the titin protein.

The first *in vitro* methodologies were chiefly concerned with the binding properties of titin to other proteins and its disruption by variants of interest; yeast 2-hybrid assays (Y2H)[90] were used to assess the binding of titin to alpha-actinin 2 (ACTN2)[91, 92], telethonin (TCAP)[92], and FHL2[93]. As Y2H has been criticised for a high incidence of false positives, the relevance of these changes for pathogenic potential, absent further validation, remains unclear. A broad range of other techniques to quantify and compare binding ability,

which are considered more reliable indicators of protein-protein interactions, have been used in more recent experimental characterisations, including co-immunoprecipitation (co-IP) assays[94], isothermal titration calorimetry (ITC)[95], glutathione S-transferase (GST) pulldown assays[96] and Förster resonance energy transfer (FRET) microscopy[96].

For variants in domains with no known binding partners, measurements of the stability of purified recombinant protein have been used to characterise the damage done by the variant to the local structure. Native polyacrylamide gel electrophoresis (PAGE) can indicate unfolded protein as decreased electrophoretic mobility through the gel reflects a more partially unfolded domain; bulkier, less compact monomers move more slowly than fully folded ones. Intrinsic fluorescence analysis allows for the detection of naturally occurring fluorophores in a domain; the detection of greater fluorescence intensity in the mutant versus the wild-type spectrum indicates that intrinsic fluorophores, such as tryptophan, are more solvent-exposed, and thus protein fold is less compact. Nuclear magnetic resonance (NMR) approaches can detect the presence or absence of different fold states within a mixed sample. Studies of titin domain thermal destabilisation have used methods such as differential scanning fluorimetry (DSF)[35] and circular dichroism (CD)[35, 95] spectroscopy to uncover the melting temperatures at which the domains are judged to have unfolded, which can be compared between wild-type and mutant systems. Mechanical methods measure the force that must be exerted on a domain to cause unfolding; single-molecule atomic force microscopy (AFM)[97, 98] has been used to probe the destabilisation of domains by missense variants. While protein denaturation methods and protocols vary widely, experiments using chemical, thermal, and mechanical methods on titin domain Ig86 have demonstrated that the unfolding rates were consistent[99]. Thus, comparison across different methods is possible; however, re-folding rates differ slightly.

In addition to experimental characterisation of variant domains alone, mutated titin constructs may be expressed in individual cells, allowing for characterisation in a context closer to that of the native environment. An example of this approach is found in Rees *et al.* (2021)[35], where GFP-tagged dual-domain tandem constructs of both wild-type and mis-

sense variant domains were expressed in neonatal rat cardiomyocytes. The authors compared the expression, localisation, and solubility of wild-type and missense-containing constructs under confocal microscopy, noting that the variant systems showed mis-localisation of the titin protein, as well as evidence of aggregation. These demonstrate how characterised features such as solubility and cellular localisation can be used to indicate the functional impact of a variant in addition to changes to binding affinity and stability.

1.3.4 Expression in Model Organisms

The use of model organisms allows for the pathogenicity of a variant to be tested within a living system – should the phenotype of the human carrier be mirrored in the model upon introduction of the variant, a causal relationship between variant and pathology can be established. However, due to the considerable investment of time and resources required, *in vivo* techniques still only account for a minority of studies on titin missense variants in the published literature.

One of the difficulties in studying a living variant system is that, while laboratory mice are a useful system for studying gene variants, the phenotypes they exhibit are often markedly different from those of human patients with the same variant. This is exemplified by a recently published, long-term study by Jiang *et al.* on the disease mechanisms of the Z-disc variant A178D[71]. Despite previous work demonstrating the capacity of this variant to destabilise its Ig domain, resulting in aggregation and loss of binding to TCAP, no mechanism was offered as to how this could lead to the observed DCM/LVNC phenotype in humans[96]. To investigate this, a knock-in mutant for A178D in mice was developed using CRISPR-Cas9. While the mice homozygous for A178D had a mild DCM phenotype, those heterozygous for the variant showed no phenotypic difference to wild-type mice, despite the dominant inheritance pattern for this variant in humans.

This tolerance of even known damaging mutations is often found for studies using laboratory mice, and in many ways is to be expected; laboratory mice generally live short lives with few stressors which would exacerbate damage to the sarcomere. Even where a different phenotype is produced, it may be a weaker or less noteworthy pathology; for example, the variant T2850I, which has been linked to ARVC, produces only a mild diastolic stiffness phenotype in laboratory mice[100]. For this reason, Jiang *et al.* implemented a long-term approach where the mice were subjected to challenge conditions. The first challenge condition, ageing of mice to 1 year old, did not provoke a difference to wild-type mice. A second challenge condition, infusion with isoprenaline/phenylephrine (Iso/PE) to induce adrenergic challenge, resulted in a hypertrophic response that aggravated the cardiac decline; this allowed for the difference between the variant and wild-type mice to be studied.

1.4 Variant Impact by Sarcomere Region

1.4.1 Z-disc

The Z-disc region of the sarcomere represents the connection between two subsequent sarcomeric units; the characteristic Z-shape that marks the region comes from the cross-linking of anti-parallel actin thin filaments with alpha-actinin. Various other proteins localise to the Z-disc[104], building up a complex network of structural and signalling functions around the lateral cross-links. Titin is one such protein, anchored to the Z-disc through its interactions with telethonin and alpha-actinin[101] (see Figure 4).

Given the unusually strong binding between titin and telethonin[102], as well as the likely role the complex plays in the integrity of the sarcomere[103], several missense variants in Ig1 and Ig2 that impact the binding between the two proteins have been identified and studied in detail. The most extensively studied of these variants, A178D, was first identified through linkage analysis of a three-generation family with DCM and LVNC[96]; the variant was demonstrated to cause destabilisation and unfolding of the Ig2 domain as well as loss

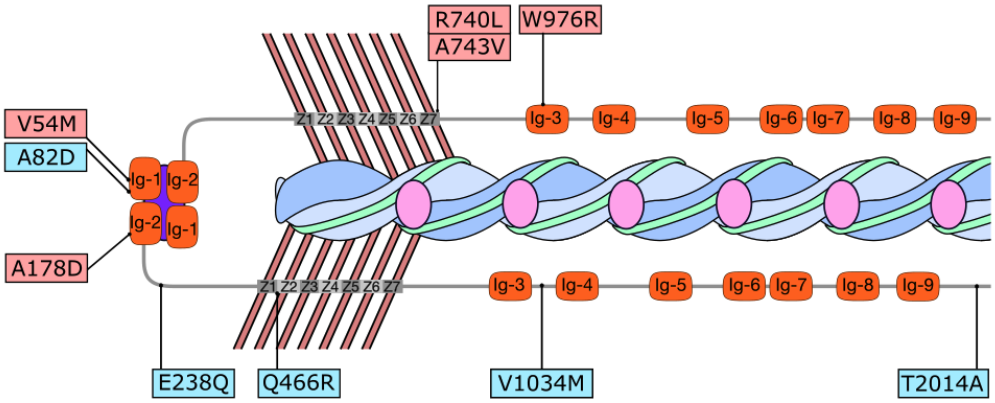


Figure 4: Reported Pathogenic variants of *TTN* in the Z-disc

Titin domains and reported disease-associated missense variants (red = cardiomyopathy-associated, blue = skeletal myopathy-associated) in the Z-disc. The characteristic Z-shape that marks the region comes from the cross-linking of anti-parallel actin thin filaments with alpha-actinin. The N-terminal region of titin passes through the Z-disc and extends through to the other side, where the initial domains Ig1 and Ig2 form a complex with telethonin (purple) with extremely strong binding[101, 102]. Two parallel titin proteins are cross-linked with a single telethonin, forming a dimer[103]. A disordered region of the protein separates Ig2 and Ig3 as titin passes through the Z-disc, including a succession of up to seven Z-repeat sequences of around 45 residues each in length, which are necessary for binding to alpha-actinin. Seven Ig domains (Ig3-Ig9), with long linkers of at least 50 residues between them, conclude titin's presence in the Z-disc.

of binding to telethonin. Interestingly, despite the loss of the titin-telethonin interaction, the integrity of the sarcomere appeared to be relatively unaffected when the variant was expressed in transgenic mice[71]; rather, the pathology appeared to result from the abundance of unbound, unfolded telethonin in the cytoplasm. As telethonin is typically unfolded when not bound to titin, the study suggested that the transgenic mice experienced a proteotoxic response if unable to clear the unfolded protein. In support of this hypothesis, RNA-seq analysis indicated up-regulation of proteasomal and heat-shock proteins, suggesting a response to an excess of unfolded protein, as well as the presence of telethonin in the pulldown fraction of a tandem ubiquitin binding entity (TUBE) assay.

Other variants that have been found in the Ig1-Ig2 region include V54M and A82D. V54M is found in Ig1 and, like A178D, decreases the binding of the Ig1-Ig2 tandem to telethonin *in vitro*, and is also linked with DCM[92]. V54M has also subsequently been extensively characterised computationally, with its effects on titin-telethonin binding probed using molecular dynamics simulations on the solved crystal structure of titin Ig1-Ig2 complexed with telethonin[105]. The simulations indicated that V54M destabilised the Ig1 domain, and also that it decreased the binding affinity of Ig1-Ig2 for telethonin. A82D (Ig1) was found co-inherited *in trans* with the truncating mutation R34807Sfs*7; interestingly, is found at the exact structural position in Ig1 that A178D is located at in Ig2[35]. While no experiment has been performed to assay the binding of telethonin to A82D-titin, it has been demonstrated that the variant substantially lowers the thermal stability of Ig1 and renders the mutant domain insoluble. However, the patient in which this variant was first identified was diagnosed with CFTD and had no evidence of cardiomyopathy or cardiac involvement at age 37 when the study was conducted.

Outside of the Ig1-Ig2 tandem, there are several other missense variants that have been linked with disease. In the Z-repeat region, two variants close to one another were found separately in cardiomyopathy patients, and alter the binding of titin to alpha-actinin 2 – R740L and A743V. Intriguingly, while A743V decreased the binding affinity of titin to alpha-actinin in Y2H assays[92], R740L had the opposite effect, increasing the binding

affinity instead[91]. Moreover, while A743V was linked to DCM, R740L was linked to HCM. This curious difference demonstrates the degree to which variants located only a few residues apart can be associated with surprisingly different pathologies. Another DCM-linked variant, W976R, is located in Ig3 and was found via linkage analysis in a single family with the condition[106].

Finally, several variants have been found in the unstructured regions between Ig domains in the Z-disc. Genetic analysis of a family with a LGMD phenotype identified the variants E238Q and Q466R, both in the region between Ig2 and Ig3, in *cis* within an affected mother and her son, but absent from the unaffected son[107]; intriguingly, a known pathogenic variant in calpain-3 (*CAPN3*:R489Q) was also found in these patients, but did not co-segregate with the condition, despite localising to sub-cellular regions more closely associated with LGMD than the Z-disc of titin. Two more variants have been reported via genetic sequencing of two cohorts of neuromuscular disease (NMD) patients: V1034M was found in the long linker between Ig3 and Ig4[108], compound heterozygous with the A-band variant A18983T; T2014A was identified between Ig9 and Ig10[37], at the very periphery of the Z-disc and I-band, in *cis* with the M-band variant R33903H and compound heterozygous with the splice-site variant c.1800+1G>A. It should be emphasised that E238Q, Q466R, T2014A and V1034M were prioritised solely through bioinformatics pipelines, and their functional consequences have not been experimentally confirmed; this, alongside their location in unstructured regions and identification in patients with muscular dystrophies (MD) rather than with cardiomyopathy, suggests that their disease associations should be treated with caution. Indeed, T2014A occurs in around 1 in 500 individuals of East Asian ethnicity in gnomAD, thus is relatively common in healthy individuals.

1.4.2 I-band

The I-band region of titin is often characterised as "the elastic region"[109, 110]; unlike the other regions of titin, it is not anchored to the greater structure of the sarcomere and can

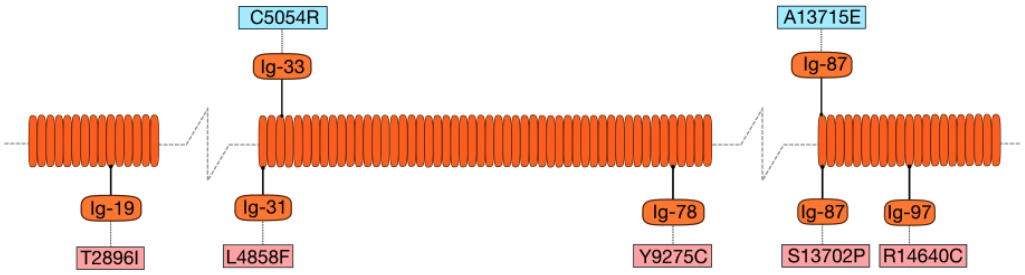


Figure 5: Reported Pathogenic variants of *TTN* in the I-band Ig tandem regions

Titin domains and reported disease-associated missense variants (red = cardiomyopathy-associated, blue = skeletal myopathy-associated) in the I-band tandem regions. The transition between the Z-disc and I-band regions of titin is marked by a short, unstructured region of titin between Ig9 and Ig10, the first domain in the I-band. A region of closely arranged tandem Ig domains connected by short linkers then follows, from domains Ig10 to Ig24. The second Ig tandem region, from domains Ig31 to Ig82, is composed of closely arranged Ig domains from Ig31 to Ig82, in blocks of around 5-6 domains joined by short, 2-3 residue length linkers, themselves connected by longer linkers of 5-6 residues. This region is not found in the N2B isoform and is only present in the N2BA and N2A isoforms. The third Ig tandem region, from Ig87 to Ig107, follows the PEVK region in all giant isoforms.

be sub-divided into multiple different regions, including three Ig domain tandem regions (see Figure 2; Figure 3; Figure 5). The differential splicing of the *TTN* gene in the I-band region distinguishes the major isoforms of titin from one another[12, 111, 112]. In brief, the Ig domains and PEVK region of the I-band adopt a folded, bent arrangement at rest, which straightens and extends as the sarcomere relaxes, including the unfolding of the PEVK; this connection allows the sarcomere to stretch without compromising structural integrity in vertebrate striated muscle, limits the movement of the thin and thick filaments relative to one another, and provides minor passive force to contraction[18].

Within the extensible regions, many variants have been proposed as associated with various diseases, chiefly cardiomyopathies. The most notable of these is T2896I in Ig19, in the first Ig tandem region. It was originally prioritised through whole titin exon sequencing of 38 families with ARVC phenotypes[113], where it co-segregated fully with the disease in a single family of 26 individuals (8 affected carriers) and was selected for further analysis *in vitro*. The variant and WT domains were compared by various methods. The authors conducted analysis of T2896I and WT domains with native PAGE and intrinsic fluorescence,

suggesting that the mutant folded into a less compact state than WT; heteronuclear Single Quantum Correlation (HSQC) NMR analysis showed the sum of two spectra in the tested sample, indicating a combination of correctly folded and partially unfolded structures. Finally, a tandem Ig construct of Ig16-Ig22 experienced a greater degree of proteolysis when incubated with trypsin where mutant Ig19 was used instead of WT. Of the other ARVC-linked variants found in the original patient cohort, one other was found in the elastic I-band: Y9275C in Ig78, in the second, intermediate Ig tandem. However, unlike T2896I, penetrance in this variant was incomplete, and it has not been characterised further.

Following on from these assays, subsequent studies have further characterised the T2896I variant through a combination of *in silico*, *in vitro* and *in vivo* methods, looking at its effects on the single domain, the longer titin chain, and the organism itself. Anderson *et al.* (2013)[98] used atomic force microscopy (AFM) to demonstrate that constructs of 5 tandem mutant Ig19 domains (5-mers) required lower force to unfold than wild-type 5-mers, but that re-folding rates were not significantly affected. They also conducted a trypsin digestion assay, concluding that mutant domains were more susceptible to degradation than wild-type domains. Conversely, Bogomolovas *et al.* (2016)[100] did not detect any significant difference in degradation between wild-type and mutant protein introduced into living mouse skeletal muscle tissue by electroporation. However, they found that the T2896I variant decreased the melting temperature of the Ig19 domain by around $11^{\circ}C$ using differential scanning fluorimetry (DSF). The structure of the mutant domain was solved using X-ray crystallography, and HSQC NMR was used to quantify chemical shifts between mutant and WT; the authors concluded that the variant was causative of a direct increase in the internal flexibility of the domain but did not alter its global structure. X-ray crystallography was again used to solve the structure of the Ig18-Ig20 tandem, which was then used to assess the conformational dynamics of the titin chain using MD simulations, showing that the interface-buried T2896I was likely to alter the flexibility of the chain and thus the relative orientation of the domains. Finally, transgenic mouse models were generated and compared with wild-type mice; while the T2896I KI mice indicated an increase in diastolic stiffness based on echocardiography, they did not present with the full ARVC phenotype. The variant

has since been further characterised using MD simulation and *in silico* stability prediction tools, which suggest that any variants at the T2896 position may be potentially damaging through the same mechanism[114].

The cardiomyopathy most linked to variants in the elastic I-band is DCM; the association has been reported across multiple studies worldwide. The first studies to report the connection were performed on East Asian populations; S4780N was found in Japanese patients with familial DCM along with other variants in the Z-disc and N2A region[92], while G4714D was described in an analysis of Chinese families with DCM[115]. Interestingly, both variants localise to Ig30, a domain that borders the N2B region of the I-band, and is present in all three of the N2A, N2B and N2BA giant isoforms. A third variant, L4854F, was found nearby in Ig31, co-segregating with a severe DCM phenotype when double heterozygous with the laminin A variant K219T in a single Italian family[116]. It should be noted that this variant, located in Ig31, would not be expressed in the predominant cardiac N2B isoform, suggesting that it would play a reduced, likely modulatory role in cardiomyopathy.

An extensive study of TTN missense variants in 147 DCM patients revealed 348 missense variants in N2B and N2BA isoforms across the cohort[89], which were filtered down to 44 priority variants using the bioinformatics prioritisation pipeline ANNOVAR[88]. These 44 variants were then filtered based on segregation analysis, giving 5 variants that co-segregated with the disease in the I-band and A-band. 2 of the co-segregating variants, S13702P (Ig87) and R14640C (Ig97), were found in the I-band, both in the third Ig tandem. However, S13702 showed discordant segregation, with 1 obligate carrier, and R14640C was co-inherited with the A-band variant E29590Q. These two variants were then further characterised by a subsequent study using AFM and chemical denaturation[97]. Both variants reduced the overall thermodynamic stability of their respective domains but did not cause unfolding under physiological conditions. However, R14640C was found to substantially decrease the mechanical stability of the domain, reducing unfolding force from an average of 187 pN to 145 pN, while S13702P only slightly destabilises its domain, from an average of 180 pN to 167 pN. ken together, these results indicate that, despite being prioritised together,

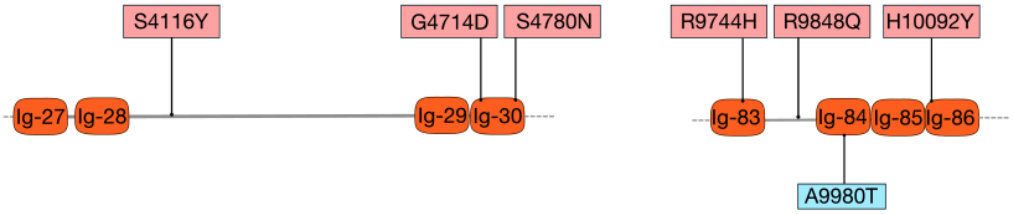


Figure 6: Reported Pathogenic variants of *TTN* in the I-band N2A/N2B-specific regions

Titin domains and reported disease-associated missense variants (red = cardiomyopathy-associated, blue = skeletal myopathy-associated) in the a) N2B and b) N2A regions. The N2B region is found in the N2B and N2BA isoforms and functions both as an extensible region and as a signalling hub for proteins such as heat shock proteins and members of the four-and-a-half LIM domain protein family (FHL). The N2A region is found in the N2A and N2BA isoforms and acts as a signalling hub for proteins such as the ANKRD family.

S13702P is either neutral or close to, while R14640C likely has a much stronger role in DCM pathogenesis, together with E29590Q.

More recently, variants in the elastic I-band linked to skeletal myopathies have been identified and characterised[35]. The variant C5054R, located in Ig33, is linked with a CFTD phenotype; the variant is not found in the N2B isoform, and the patient showed no sign of cardiomyopathy. Conversely, the variant A13715E, linked with a type 1 fibre predominance (T1P) phenotype, is found in Ig87, which is present in the N2B isoform; the patient here did have cardiomyopathy. The authors observed similar trends where truncating variants compound heterozygous with single splice site or nonsense variants in the Ig31-82 region did not produce cardiomyopathy but did produce localised skeletal myopathy, albeit without a full segregation of the phenotype.

1.4.3 N2B

The N2B region is a short, mostly disordered region found in the N2B and N2BA isoforms (see Figure 6a); with that said, its importance both as an extensible region and a signalling hub has long been known and more recently experienced renewed interest[20, 117]. Modifications

to this region can decrease (e.g. phosphorylation by PKA and PKG) or increase (e.g. the formation of disulphide bridges between cysteine residues) titin-based passive stiffness of the sarcomere. It is also the target for several interaction partners, including the heat shock proteins Hsp90 and αB-crystallin, and members of the four-and-a-half LIM domain protein family (FHL)[20].

The interaction between the N2B region and the FHL proteins is thought to be the basis of pathogenicity for the best-characterised variant in the N2B region, S4116Y. First identified in a Japanese patient with HCM[92], it was then characterised *in vitro* in a subsequent paper looking at several HCM- and DCM-linked variants in titin[93]. A Y2H assay indicated that S4116Y significantly reduced the binding of the N2B region to FHL2; another study demonstrated that the variant also decreased the phosphorylation of the N2B region by MAPK/Erk2[118]. The authors concluded that the binding of FHL proteins is an essential part of the stress-sensing function of the N2B region and demonstrated the consequences on the cardiac stress response in *ex vivo* knock-out studies in FHL1 in mice. However, the link between the titin variant and HCM remains tentative.

1.4.4 N2A

The N2A region is, like the N2B region, only present in certain isoforms (in this case N2A and N2BA) and acts as a signalling hub, with many interaction partners and post-translational modifications that modulate its behaviour[20, 119] (see Figure 6b). However, it is described as an inelastic part of the I-band[120], and there is some debate over whether certain described interactions are indeed reflected *in vivo*[121]. The most noteworthy of these is the assertion that N2A binds directly to the actin thin filament, modulated by calcium binding; however, even the most recent studies have drawn conflicting conclusions from the experimental data. Kelly *et al.* (2021)[122] argue that Ig86 alone binds actin in a calcium-dependent manner; Stronczek *et al.* (2021)[123] were unable to find binding sites for either calcium or F-actin in the Ig84-86 tandem; Tsiros *et al.* (2022)[124] argue that only constructs that contain Ig83

and with at least 3 Ig domains can bind F-actin, and that the calcium-dependent binding property is lost when Ig86 is not present. Outside of actin binding, the N2A region binds to chaperones, methyltransferases, and proteases, thought to protect the region from damage and to trigger a response in the event of mechanical stress being detected. Most notably, the ANKRD protein family cross-links adjacent titin molecules by dimerization; a trimeric complex of ANKRD1, titin, and F-actin has also been suggested[125].

Like S4116Y in the N2B region, variants in the N2A region have also been associated with HCM, in this case through the disruption of the binding of titin to CARP1, a protein product of the ANKRD1 gene. Genetic sequencing of 384 HCM patients for variants in titin or ANKRD1 identified 5 mutants, 2 in the N2A region of titin and 3 in ANKRD1[94]. The first of these, R9744H, was found in Ig83; the second, R9848Q, was found in the unique sequence between Ig83 and Ig84. A Co-IP assay demonstrated that the binding of titin to CARP1 was increased by both titin variants, and functional studies *in vitro* demonstrated that the subcellular localisation of Myc-tagged CARP1 was altered. A third variant in the N2A region, H10092Y (Ig-86), was found in the same study of ARVC families that first identified T2896I in the I-band[113]; like the other variants in that study, it showed incomplete penetrance, and as mentioned in Section 1.3.1, is present in about 1 in 250 individuals in gnomAD, increasing to around 1 in 100 individuals in the Finnish sub-populations (Table 1). Subsequently, a re-examination of the family in question identified a variant in desmoglein-2 that co-segregated with ARVC[126], likely settling this variant as neutral. Similarly, A9980T (Ig-84) was identified in two siblings with titinopathy and asymmetric facial and limb weakness; it too has a high MAF in gnomAD, present in around 1 in 1000 non-Finnish European individuals, and was reported alongside the variants A19938T and R29293C in the same individuals[127].

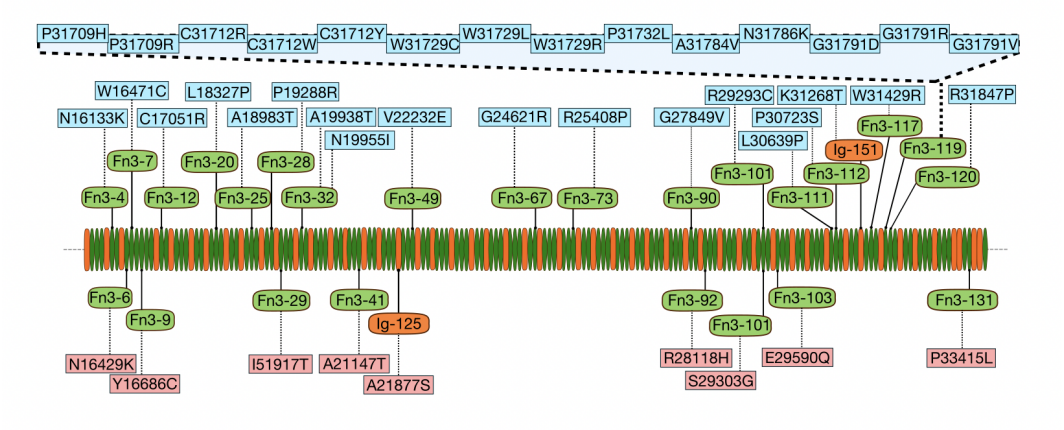


Figure 7: Reported Pathogenic variants of *TTN* in the A-band

Titin domains and reported disease-associated missense variants (red = cardiomyopathy-associated, blue = skeletal myopathy-associated) in the A-band. The section of titin in this part of the sarcomere is closely associated with the myosin thick filaments, thus unlike the I-band titin region does not extend in response to stretch. The region is distinct in its composition, being mostly comprised of tandems of 2-3 Fn3 domains (green) interspersed by individual Ig domains (orange), all closely connected by short linkers to form characteristic super-repeats of 7 (Ig-Fn3-Fn3-Ig-Fn3-Fn3-Fn3) or 11 (Ig-Fn3-Fn3-Ig-Fn3-Fn3-Fn3-Ig-Fn3-Fn3-Fn3) domains.

1.4.5 A-band

The A-band is the region of the sarcomere where the motor domains of the myosin thick filaments are located, and where the active contraction of the sarcomere takes place through the cross-bridge muscle contraction cycle[128]. Unlike the other regions of titin, the A-band section contains both Fn3 and Ig domains (see Figure 7), and is closely associated with the myosin thick filaments, thus unlike the I-band titin region does not extend in response to stretch[4], although a role in thick filament mechanosensation has recently been proposed[129]. Given the great length of the A-band relative to the other regions of the sarcomere – 183 of titin's 302 total domains are in this region – it follows that most known titin missense variants linked to disease fall within this region by sheer coincidence, accompanied by a corresponding variation in the evidence for pathogenicity and the types and severities of pathologies recorded.

Fn3-119, a single fibronectin type-III domain located near to the C-terminal end of the A-band, is perhaps the single most extensively characterised domain in titin in relation to a single condition. Over 10 missense variants in this one domain have been linked to hereditary myopathy with early heart failure (HMERF), with five variants characterised *in vitro*[130]. The first such variant identified, C31712R, is a fully penetrant variant with autosomal dominant inheritance, and has been identified in a wide range of studies and different populations. It was first found to co-segregate with the disease in the 3 unrelated Swedish families tested[131]; another study, published separately but almost concomitantly, found the same variant in 3 British families, again co-segregating with observed pathology[132]. It has subsequently been reported in a number of studies, in different ethnic groups: in individuals of British, Swedish, Finnish, Italian, and Argentinian ancestry[133]; in two individuals, one of Native Canadian and one of Spanish ancestry[134], with the distinct ancestral haplotype of the Native Canadian patient suggesting independent origins in different ethnic groups; in two separate studies of myofibrillar myopathy (MFM), associated with HMERF pathology related to that condition[45, 46]; in a Chinese family in 2015[135], confirming this variant as having a worldwide distribution; in one individual in a British cohort for a study of clinical applications of magnetic resonance imaging (MRI)[136]; in one Afghan, one Spanish, and two unrelated Italian individuals, as well as four individuals in a family of Russian ancestry[44]; more recently, in a Portuguese patient[137] and a Chinese patient[37], both in 2021. Finally, the variant domain has been experimentally characterised *in vitro*, and was found to greatly decrease the solubility and folding of the Fn3-119 domain, where the wild-type domain and those with common variants did not[130]. Thus, C31712R is considered the TTN missense variant with the greatest consensus as being directly causative of disease.

Alongside C31712R, many other variants in Fn3-119 have been linked to HMERF. Variant W31729L was identified in 20 HMERF patients in a single Japanese family and found to co-segregate with the disease in the 9 family members used for linkage analysis[138]. Two more substitutions at the same position, W31729R and W31729C, were found in British and German individuals, respectively, in another study, alongside P31709R in a French

family[133]. W31729C was also found in another individual in the above British MRI cohort[136]. P31709R, W31729R, and W31729C were characterised *in vitro* in the same study that characterised C31712R, and likewise decreased solubility and folding of the Fn3-119 domain[130]. W31729R was also identified in another study of Chinese titinopathy patients[37]. The previously mentioned studies describing C31712R in patients with MFM and HMERF pathology also identified the variant G31791D, as well as three more mutations at the same positions as these two mutations (C31712Y, G31791R, G31791V)[46]; one also identified N31786K as a potential variant, albeit with no segregation analysis[45]. N31786K was also identified in a single Iranian patient[44], alongside the variant P31709H, at the same position as the previously described P31709R, in one Portuguese and one Filipino-Caucasian patient, as well as another example of the W31729C variant in an unrelated Portuguese patient, and the novel variant A31784V in two French families and one Argentinian patient. Finally, a third variant at the same position as C31712R and C31712Y was recently reported; C31712W was identified in a Korean family with HMERF manifesting in old age.

While almost all the above variants are thought to be fully penetrant, causative variants, the variant in Fn3-119 that has been most extensively characterised besides C31712R has much more uncertainty over its pathogenicity. P31732L was the second identified variant in Fn3-119 after C31712R, in a study that used a blind bioinformatics pipeline to prioritise likely causative genetic variants in a cohort of patients with neuromuscular diseases[108]. Like C31712R, P31732L has been identified in multiple patients since its original description, including an Italian family[133], two French-Portuguese families[133], and sporadic cases in British[45], Chinese[135], and Japanese[139] individuals; however, several of these studies made note of the incomplete penetrance observed in these cases, particularly in comparison to C31732R and other fully penetrant variants. Palmio *et al.* (2013)[133] observed that the variant was neither fully dominant nor fully recessive, producing a severe phenotype when homozygous but an unclear one when heterozygous. Similarly, Pfeffer *et al.* (2014)[45] noted that P31732L was also found in the affected patient's unaffected brother. They concluded that the variant may be neutral, of variable penetrance, or late-onset, but ruled out its role as a fully penetrant, dominant variant. However, two separate studies have found that the

variant reduces the thermal stability and solubility of the Fn3-119 domain[35, 130], as with the fully penetrant variants mentioned above. In the later of these two studies, P31372L was homozygous in the two affected individuals[35], supporting the argument that the severe phenotype requires homozygosity or compound heterozygosity to be produced. However, the observed phenotype in these patients was a congenital myopathy with features of HMERF, rather than HMERF.

Outside of the well-established Fn3-119, there are many other variants in the A-band associated with both skeletal and cardiac myopathies. Interestingly, several of these variants occur at the same positions as variants in other domains associated with other conditions, and even represent identical amino acid substitutions; this suggests that there is a similar mechanism that produces different phenotypes depending on the position of the domain within titin. For example, N16133K (Fn3-4) is associated with MmD, and disrupts the solubility and thermal stability of the Fn3-4 domain *in vitro*[35]; it is located at the same structural position as N16429K (Fn3-6), which was reported to co-segregate with atrioventricular block (AVB) in a 3-generation Chinese family[140], as well as the previously mentioned N31786K (Fn3-119). While not the same residue substitution, this position is also the location of N19955I (Fn3-32), which was found compound heterozygous with L30639P (Fn3-111) in an Chinese infant with titinopathy[141]; they are both also compound heterozygous for the splicing variant c.44282-2A>G (inherited in *cis* with N19955I, in *trans* with L30639P), however an *in vitro* expression assay showed no difference in expressed isoform ratios, while confirming a significant decrease in full-length titin compared to healthy controls. W16471C (Fn3-7), C17051R (Fn3-12) and G27849V (Fn3-90) are three more MmD-associated variants, all of which were found to decrease the thermal stability and/or solubility of their respective domains *in vitro*[35]; each of them is located at the same structural position as the Fn3-119 variants W31729C, C31712R, and G27849V, respectively. W31429R (Fn3-117) was identified in a Spanish family with TMD[142], at the same structural position as W31729R (Fn3-119). Finally, while not at the exact same structural position in the two domains, P31732L (Fn3-119) and P33415L (Fn3-131) are located at positions only one residue apart in the same loop region of the fibronectin fold and represent the same residue substitution; however, they

differ in that P33415L is an exposed surface residue whereas P31732L is mostly buried, and that P33415L is associated with the cardiac ARVC phenotype[113] whereas P31732L is associated with HMERF and congenital myopathy.

The remaining known variants are linked with a variety of both skeletal and cardiac myopathies, located throughout the A-band. L18237P (Fn3-20) and R25480P (Fn3-73) are two variants associated with centronuclear myopathy (CNM)[35]; both have been found to destabilise their respective domains *in vitro*. A18983T (Fn3-25) was identified in a patient with MD[108], compound heterozygous with the previously mentioned V1034M at the Z-disc; it has yet to be experimentally characterised. K31268T (Ig151) was identified through WES of a single Estonian family; it co-segregates with inguinal hernia in that family and was not found in ethnically matched population controls[143]. Two further variants associated with MmD not yet mentioned here are R31847P (Fn3-120) and V22232E (Fn3-49), both of which destabilise their domain *in vitro*[35]. The same individual in which V22232E was found has been described in two separate studies[35, 48]; interestingly, it is noted that this variant is implicated in heart disease as well as the regular MmD phenotype. Recently, several variants have been reported on the basis of bioinformatics methods alone; these should not be treated as definitive absent future experimental validation. P19288R (Fn3-28) and G24621R (Fn3-67) were identified as damaging by PolyPhen-2[144] and SIFT[145]; P19288R was found alongside the frameshift variant K4230Nfs*17, and G24621R with the truncating variant Y4418*[37]. A19938T (Fn3-32) and R29293C (Fn3-101) were reported alongside the previously mentioned A9980T (see Z-disc)[127]; however, as with A9980T, R29293C has a high incidence in population data (found in around 1 in 500 individuals of East Asian ethnicity in gnomAD).

Two TMD-linked variants have been found in the A-band, despite the strong association of that condition with variants in the M-band. The first, W31429R, has already been mentioned[142]; however, it has subsequently been reported in a Spanish patient with TMD and DCM[146]. In the former study, it was found to be compound heterozygous with the truncating variant R21209*, whereas in the latter, it was reported homozygous; this distinc-

tion may have some bearing on the differences in phenotype. Similarly, the second variant, P30723S (Fn3-112), was identified in a patient compound heterozygous for this variant and the strongly TMD-linked 'FINmaj' indel (see M-band)[80]. His father, who carried P30723S only, was asymptomatic; consequently, the authors argue that the phenotype of TMD is exacerbated, rather than caused, by these missense variants in conjunction with truncating variants, resulting in greater severity of disease.

For cardiac variants, the single TTN variant associated with RCM, Y16686C (Fn3-9), was found to co-segregate with the condition in a single Jewish family[147]. The variants I19517T (Fn3-29), A21147T (Fn3-41), and A21877S (Ig125) were prioritised computationally as associating with ARVC in the same study that identified the previously mentioned P33415L[113]; however, the variants did not fully segregate with the pathology within the patients' families. Several DCM-linked variants that have been prioritised by bioinformatics approaches also fall within the A-band, namely R28118H, S29303G, and E29590Q; these were identified in a previously mentioned large-scale DCM study (see I-band)[89], and do show co-segregation within families, though they lack additional experimental validation.

1.4.6 M-band

The C-terminal region of titin is located in the M-band, at the very centre of the sarcomere[28] (see Figure 8). Here, the antiparallel myosin molecules that make up the thick filament overlap; while the heads of the molecules extend into the A-band to bind the actin thin filaments, the tails are held in position at the M-band by the members of the myomesin family. Titin, too, is anchored at the M-band, with its C-terminal domains overlapping with those of the antiparallel titin molecules from the opposite end of the sarcomere. In this way, the full sarcomere can be thought of as a symmetrical construct, with two titin proteins extending from either Z-disc to meet at the M-band. Titin binds to its fellow giant sarcomeric protein obscurin or obscurin-like-1 to form a ternary complex with myomesin, forming a link both to the thick filaments and to the outer sarcolemma and sarcoplasmic reticulum[148].

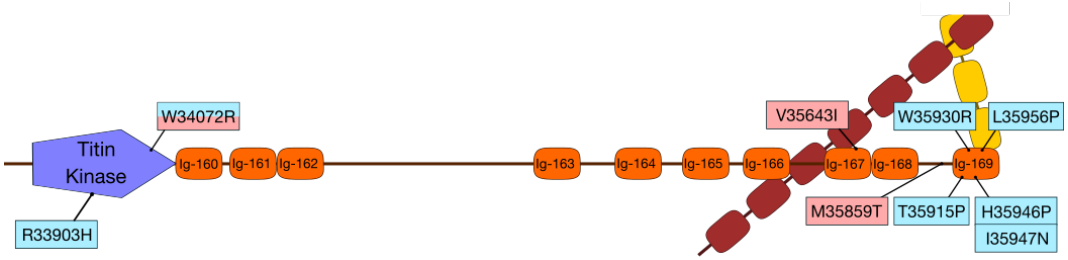


Figure 8: Reported Pathogenic variants of *TTN* in the M-band

Titin domains and reported disease-associated missense variants (red = cardiomyopathy-associated, blue = skeletal myopathy-associated) in the M-band. Here, the C-terminal domains of titin overlap with those of the antiparallel titin molecules from the opposite end of the sarcomere. Titin binds to its fellow giant sarcomeric protein obscurin (yellow) and myomesin (dark red) to form a ternary complex. The kinase domain of titin at the A/M-band junction acts as a signalling hub that monitors the mechanical stresses undertaken by the sarcomere. Note that W34072R (titin kinase) is designated as associated with both skeletal and cardiac myopathies, as it is cardiomyopathy-linked if co-inherited with a cardiac-expressed truncation.

Moreover, the presence of the kinase domain of titin at the border between the A- and M-band regions is thought to act as a signalling hub that monitors the mechanical stresses undertaken by the sarcomere[16, 17, 77].

While the M-band has historically received less attention than other regions of the sarcomere despite the importance of its structural and signalling functions, many of the earliest associations between skeletal myopathies and titin missense variants were made in this region, particularly with respect to tibial muscular dystrophy (TMD) and its links to the final domain in titin, Ig169 (also known as M10). The first such association was found in 2002, where genetic testing of Finnish patients identified a non-frameshift indel of four consecutive substitutions (EVTW>VKEK at position 35927)[41]. This variant, labelled FINmaj, completely co-segregated with TMD with LGMD2J in the 81 individuals from 12 unrelated families tested. The authors attempted to find a comparable mutation in a French family with TMD with LGMD2J; they were able to isolate the variant L35956P, which again co-segregated with the phenotype within that family. Similar studies later identified the variant I35947N in a Belgian family[149], and H35946P in an Italian family[150].

These four mutations were biophysically characterised *in vitro* by Rudloff *et al.* (2015)[95] to compare the magnitude of their effects on the single Ig169 domain. As Ig169 binds to obscurin's N-terminal Ig1 domain at the M-band, the binding affinity of mutant Ig169 for OBSCN-Ig1 was compared to wild-type Ig169 using ITC, as well as comparisons of stability, expression, and aggregation. The FINmaj variant had the greatest effect on the domain, producing severe misfolding, aggregation, the lowest expression, and an abolition of binding to OBSCN-Ig1. The L35956P variant also produced severe misfolding and a loss of OBSCN-Ig1 binding, as well as lowered expression, but did not show aggregation. While the H35946P variant was stable at room temperature, it unfolded at body temperature; otherwise, like L35956P, it had lowered expression, loss of OBSCN-Ig1 binding, but soluble expression. Finally, I35947N showed no difference in stability from WT, as well as normal binding affinity for OBSCN-Ig1, normal expression and no aggregation.

A further study of the Finnish founder mutation FINmaj found that most patients expressed a classic TMD phenotype, whereas only a minority expressed additional phenotypes such as LGMD2J[151]; a study of a 5-generation Chinese family with LGMD2J identified the variant W35930R, which co-segregated fully with the disease in that family[152]. Interestingly, unlike the other variants mentioned, co-segregation of I35947N with TMD was not complete; an asymptomatic family member also carried the variant, albeit with a sub-clinical condition only discovered after subsequent muscle biopsy[149]. A later study also identified I35947N in another Belgian family with TMD[142]; here, the variant was compound heterozygous with the truncating variant Q22507*. This study also identified the variant T35915P in a Tunisian family with TMD, compound heterozygous with the frameshift variant E28338Hfs. Notably, not only were both variants found together with truncating variants, but they also had incomplete penetrance; family members with only one variant and without compound heterozygosity were not affected. Thus, the diversity in phenotype resulting from Ig169 variants may arise in part from truncating or frameshift variants that work in concert with them.

Aside from Ig169, the other element of M-band titin that has received the most interest is the kinase domain. The mechanical stretching of the sarcomere induces conformational change in the kinase, allowing for the binding of the zinc-finger protein nbr2, precipitating a signalling cascade that leads to the activation of the serum response transcription factor SRF[77]. Disruption of this pathway would likely lead to disease as transcriptional changes in response to stress would be altered; however, the variant first identified as doing so, R34091W, was found in two Swedish families with HMERF[76], whose relatives would later be confirmed to also carry the A-band Fn3-119 variant P31732L, which has already been noted as strongly linked to HMERF[78]. This, combined with the relatively high allele frequency of R34091W in non-Scandinavian European populations, has led to its re-classification as benign, although it has been suggested that it may contribute to the phenotypic penetrance of the otherwise incompletely penetrant P31732L variant (see A-band) in monoallelic-bimutational inheritance[153].

The only other disease-associated variant in the kinase domain that has been validated is W34072R, which was originally identified in a patient with MmD-HD (with cardiomyopathy), compound heterozygous with the frameshift variant E2989fs*[48]. It was found to decrease the thermal stability of the kinase domain compared to wild type. Additionally, it was identified a second time in an unrelated patient with non-specific congenital myopathy[35] without cardiomyopathy, compound heterozygous with the truncating variant R7796*. Finally, the variant R33903H was identified in a titinopathy patient in *cis* with the previously mentioned T2014A (see Z-disc) and compound heterozygous with the splice-site variant c.1800+1G>A; as with T2014A, R33903H is relatively common in East Asian populations, also occurring in around 1 in 500 individuals in gnomAD, thus is unlikely to have much role in the disease.

While several cardiomyopathy-associated variants have been reported in the M-band, they are not as comprehensively studied as the variants in Ig-169 and the titin kinase domain, and their associations with disease have not been validated. In the same family in which the previously mentioned ARVC-linked variant A21877S (see A-band) was identified, the

M-band variant M35859T was also reported in the long linker between Ig168 and Ig169[113]. As with A21877S, M35859T has incomplete penetrance with ARVC; moreover, its location in an unstructured region and its high population frequency (around 1 in 100 individuals of South Asian ethnicity in gnomAD) render it unlikely to be connected with disease. A recent paper which aimed to identify individuals in the early stages of familial HCM prior to diagnosis[154] reported the variant V35643I (Ig-167) in a Hui Chinese family with a history of the disease. In a cohort of 18 asymptomatic individuals from this family, the authors report significant echocardiographical differences between those carrying this variant (n = 6) and wild-type individuals (n = 12); they suggest that these differences reflect future onset of HCM, and that such screening techniques can provide early diagnosis of the condition. This is intriguing, as there are no other reports of variants associated with HCM in M-band titin, and relatively few variants in the whole of titin linked to that condition; follow-up studies on this cohort will hopefully shed light on this hypothesis.

1.5 Location-Function Relationships

The vastness of titin and the diversity of its molecular contexts should make it apparent that a variant in one domain or region may be associated with a very different phenotype than one in another. Efforts to map the aetiology of various myopathies with respect to the locations of the associated variants in titin have been mixed[22–24, 155]; while there are extremely strong associations between the development of TMD and HMERF and variants in the very specific modules Ig169 and Fn3-119, respectively, most other conditions are associated with many variants spread across titin.

The reverberations of the pathogenic missense mutation to the local domain structure, the dynamics of the titin chain, and ultimately health of the patient carrying the variant, reflect several factors related to its location in titin. Depending on the structural location of the variant in its parent Fn3 or Ig domain, the domain may still fold correctly, or partially or even fully unfold at physiological conditions, especially if the variant disrupts the hydrophobic

core of the domain. Alternatively, if it occurs on the domain surface, where it is less likely to affect domain stability, it may still disrupt or destabilise interactions with native binding partners. While the relationship between a variant's capacity to destabilise its parent domain and its pathogenicity to the carrier is not fully understood, the destabilisation of certain domains, most notably Ig169 and Fn3-119, has clear and reproducible consequences[78, 95]. These potential ramifications of the location of variants in titin are discussed below.

1.5.1 Isoforms

A clear consequence of the location of the missense variant in the titin sequence is whether it is expressed in a particular isoform. Notably, the Ig31-82 tandem, N2A region, and most of the PEVK region are absent from the N2B isoform, while the N2B region is absent from the N2A isoform[30]; thus, the consequences of pathologies may be modulated by the expression patterns of isoforms in different tissues. For example, the variant W34072R (titin kinase) has been identified in two individuals; in the first, where the variant was found compound heterozygous with E2989Efs*4, the patient presented additionally with cardiomyopathy[48], whereas the second, compound heterozygous with R7796*, did not[35]. It is notable that the former truncating variant occurs prior to the N2B region in the TTN sequence, while the latter occurs after, potentially explaining this discrepancy. Similar considerations may play a role in the often very restricted effects of striated muscle diseases, such as TMD, and also the variability between cases of the same condition[40].

The N2B isoform, predominant in cardiac muscle, is shorter than the N2A and N2BA isoforms; it is expected that missense variants associated with cardiomyopathies would not occur in regions which N2B lacks. This is not always the case; for example, the DCM-linked variant L4854F (Ig31) and the ARVC-linked variant Y9275C (Ig78) are both located in the Ig31-82 tandem region. However, L4854F was found as a double heterozygote with LMNA variant K219T[116], and Y9275C presents with incomplete penetrance[113]; hence, it is feasible that both variants have only a modulatory role in disease progression. These

variants would be present in affected hearts in the N2BA isoform, but as the expression ratios of the N2B and N2BA isoforms in hearts can vary significantly between species, between individuals, and between muscles[156, 157], identifying the proportion of titins affected by the variants in question is not simple. In mammals, the N2BA isoform is predominant in atrial muscle, but the N2B isoform represents the majority of titin in the left ventricle[112]. However, the murine models typically used for study of cardiomyocytes have far less N2BA isoform in their muscle than larger mammals, and consequently murine myocardium tends to be much stiffer[156].

It should be clarified that, when discussing the differential expression of titin isoforms, the designation of N2A, N2B, and N2BA reflects the earliest elucidations of titin's presence in muscle, where the separation of isoforms by the presence or absence of the key signalling hub regions contributed to their identification[111, 112]. More recent research has shown that alternative splicing can diversify the available isoforms through adding or removing specific exons[158]; moreover, it has been shown that certain muscles will up-regulate the expression of certain titin exons in transcripts relative to others; notably, the predominant isoforms in the foetal and neonatal myocardium are longer, more extensible isoforms such as foetal cardiac titin (FCT), which are replaced during development with the shorter, less extensible N2B isoform[32, 159]. Thus, the muscle tissue in one region of the body may incorporate far more of the translated titin with the mutated region than another. Some authors have suggested that muscle stiffness may be modulated dynamically through titin expression, especially as a response to heart disease[160–162]; this further complicates the association of variants with specific conditions or time of onset. Future studies that incorporate the analysis of transcriptomic data, particularly the comparison of expression of missense variant exons in affected and unaffected muscles, would further elucidate this issue.

1.5.2 Shared Structural Positions

As the 302 individual domains in titin are composed of only 3 distinct domain folds, it is not uncommon to see similar or identical residues at the same structural position in different domains, with similar contributions to the stability of the domain fold. Intriguingly, several missense variants identified in titin represent the exact same amino acid substitution at the same domain position, differing only in the domain in question. Notably, there is the A-band domain Fn3-119, strongly associated with HMERF; of the 13 currently described Fn3-119 missense variants associated with HMERF, 5 have seen the exact same residue substitution at the same structural position in 5 other Fn3 domains in titin, associated with other conditions.

While it is not surprising that the same substitution at the same structural position would cause the same damage to the structure, questions remain over whether the mechanisms governing the pathogenicity of the variant itself can be directly compared. For example, the variants A82D and A178D represent the same residue substitution at identical structural positions. In both cases, a buried hydrophobic core residue is replaced by a hydrophilic, charged residue, destabilising the domain core. The A82D variant was found to decrease the melting temperature of the Ig1 domain by around 37°C *in vitro*[35], and the A178D variant to decrease the binding affinity of titin for TCAP[96]. However, while the A178D variant has been associated with the cardiomyopathies DCM and LVNC, the A82D variant is associated with the skeletal myopathy CFTD, without cardiac involvement, and is co-inherited with the truncating variant R34807Sfs*7.

Both Ig1 and Ig2 are expressed in all major isoforms of titin, including the 'tiny titin' novex-3 isoform, and both are involved in the intransient interaction with TCAP that anchors titin to the Z-disc. However, most interactions between TCAP and titin involve the Ig2 domain, with no similarly extensive bond networks stabilising the interaction between Ig1 and TCAP[101, 103], potentially explaining the discrepancy. This is thought to be a consequence of the directionality of muscle stretch, as the titin-TCAP complex can bear remarkably heavy loads applied to the C-terminal end both *in silico*[105] and *in vitro*, but

not when the same loads are applied to the N-terminus[101]. In summary, the destabilisation of adjacent domains, with similar physiological roles, in an identical manner, does not imply that the consequent pathology or pathomechanisms will apply to both.

1.5.3 HCM/DCM Phenotype

A curious observation of missense variants in the Z-repeat region of the Z-disc concerns the two cardiomyopathies, HCM and DCM. As previously mentioned, the HCM-linked variant R740L reportedly increases the binding affinity of the Z-repeat region for ACTN1 *in vitro*[91], whereas the purported DCM-linked variant A743V decreases it[92]. Moreover, the increase in binding affinity for a particular binding partner is a consistent trait of HCM-linked variants; the N2B-region variant S4116Y increases the affinity of the N2B region for FHL2[93], and the N2A-region variants R9744H and R9848Q increase the affinity of the N2A region for CARP[94]. By contrast, the two other DCM variants that are linked with changes in binding affinity, V54M in Ig1 and A178D in Ig2, both decrease the affinity of titin for TCAP[35, 92].

While no firm conclusions may be drawn from the small sample size of variants in titin, it is interesting to consider the implications in light of the pathologies in question. HCM is characterised by the thickening and expansion of the heart muscle, leading to a decreased volume of the heart chambers, whereas DCM is characterised by a stretching and elongation of the heart muscle, leading to an increased volume of the heart chambers[54]. In many respects, the pathologies resemble opposite extremes, alike the opposing consequences of the Z-repeat variants. It should be reiterated, however, that the majority of HCM-linked genetic associations come from the myosin thick filament and its binding partners; thus, any aetiology involving titin does not account for all cases of this disease with genetic underpinnings. Nevertheless, this concept of opposite extremes may be an important consideration in future studies of pathogenicity. Rather than pathogenicity and domain instability being proportionally linked, increasing the stability of domains or binding interactions may be

equally damaging through an opposing mechanism. The idea for stabilising mutations to be pathogenic remains a subject of ongoing study[163], but has yet to be brought into focus for titin. It should be reiterated, however, that the majority of HCM-linked genetic variants reside in the myosin thick filament and its binding partners, and no consistent genetic or biochemical evidence has yet been provided for a role of titin in HCM.

1.6 Implications for Diagnosis

An emerging diagnostic paradigm is one that seeks to address the current classification of variants into categories based on evidence for pathogenicity, and the limitations that arise from titin variants that seem to differ in damage caused based on their molecular context. Most titin missense variants that are strongly associated with disease possess an autosomal dominant inheritance pattern[64, 164]; this is to be expected, as more tenuous disease associations require more evidence to demonstrate satisfactorily. Less attention, therefore, has traditionally been paid to recessive alleles, which may be common in the general population, but unnoticed except in the presence of another variant. However, incomplete phenotypic penetrance does not preclude the role of a variant in disease, causative or otherwise. Increasingly, variants in titin are reported whose phenotypic effects are only demonstrable in concert with other variants[80, 116, 142], and this list is likely to expand as we continue to seek a genetic basis for inherited myopathies without a clear monogenetic association.

In a recent study of congenital myopathy patients[35], it was found that most of the patients were compound heterozygous for two or more titin variants, often one missense and one truncating; in this mutational context, the damaging consequences of a misfolded region of titin can be seen in earnest at the organismal level. The authors suggest, therefore, that destabilising titin missense variants compound heterozygous with truncating variants may be sufficient to cause disease. This paradigm must be explored further, but offers an intriguing insight into some of the disease associations discussed here. For example, the TMD-associated hotspot domain Ig-169 contains multiple variants, but not all these variants

are equivalent; I35947N is noticeably less damaging, both in experimentally characterised properties[95] and in patterns of inheritance[142, 149], compared to other characterised variants. It is also the only one of these variants in which a truncating mutation was found alongside the missense, and indeed, the disease phenotype co-segregated only with those individuals who were compound heterozygous for both the missense and truncating variants.

This issue is one that is especially pertinent for titin. Not only is it larger than any other protein in the human genome, and thus practically certain to include multiple variants on both alleles[22, 23], but it is also a gene of myriad functions, and any one variant in titin may have minimal interplay with another, through the compartmentalisation of the gene into domains and the different isoforms expressed in different muscle tissues. Consequently, the standard approach for identification of pathogenic variants, where variants are labelled based on their likelihood of being causative for disease independently of any other genetic factors, risks eliminating key information that distinguishes the consequences of genetic variants with the same disease associations. For example, the online database for pathogenic variants, ClinVar[165, 166], lists 11 single amino acid substitutions in titin as definitively pathogenic; notably, these include the incompletely penetrant variants I35947N (Ig-169) and P31732L (Fn3-119) in the same category as fully penetrant variants in these same domains. While these are certainly disease-associated and well-characterised, the clear differences between these variants and those more damaging variants are lost, giving an unclear representation of damaging titin variants.

To compensate for this weakness in the annotation of titin variants and those of other giant proteins and genes, I have additionally labelled the variants reported in the literature (Table 2) with a simple key that broadly categorises the variants based on the strength and nature of the disease association. The variants are ranked from 0-5 on the evidence presented in their characterisation ("Quality of Evidence"), based on the quality and quantity of evidence to support classification. They are then further annotated with either a "+" (disease-causing with full penetrance) or "-" (modulatory role or recessive action) indicating the most likely role played by the variant based on that evidence ("Pattern of Inheritance"). The evidence

metric is additive, such that multiple sources of validation are considered cumulatively rather than separately. A schematic flowchart of the labelling procedure is given in Figure 9.

The purpose of this annotation should be to give a brief but recognisable indication of the reliability of the disease associations described. Any variant with a Quality of Evidence value of 3 or greater is considered to be of close to certain pathogenicity; those variants with a value of 4-5 reflect those that have been additionally validated by multiple studies or different methods. Variants with a value of 2 should be treated with caution, while those with a value of 1 should be treated as requiring further validation. A value of 0 indicates that no evidence supports this disease association despite the reported pathogenicity. This approach is limited by the quantity of data on titin missense variants currently reported in the literature; future sequencing studies that record the presence of recessive missense variants in titin-linked pathologies would lay the groundwork to delineate the combinations of variants that are causative for disease. Encouragingly, the existence of titin-specific databases of variants, including TITINdb[30] for protein variants and CardioDB[69] for genetic data, will aid in the processing of these large quantities of data in future.

1.7 Conclusion

The evidence for the role of titin missense variants in disease, while still lagging behind that of truncating variants, is now more pronounced and better understood. Biophysical characterisation of variants and their parent domains in vitro has given us an appreciation of how these variants affect their local structure, and the expression of variants in vivo has begun to shed light on the pathomechanisms that have so far been shrouded in uncertainty. The data collected here represent, to the best of our knowledge, the most complete resource for disease-associated titin missense variants, and the evidence in support of those associations. We have also discussed the different validation methodologies, both experimental and computational, that have been used in the past to investigate these variants, and the quantity and quality of evidence that should be required to support a given disease association. Thus, we have

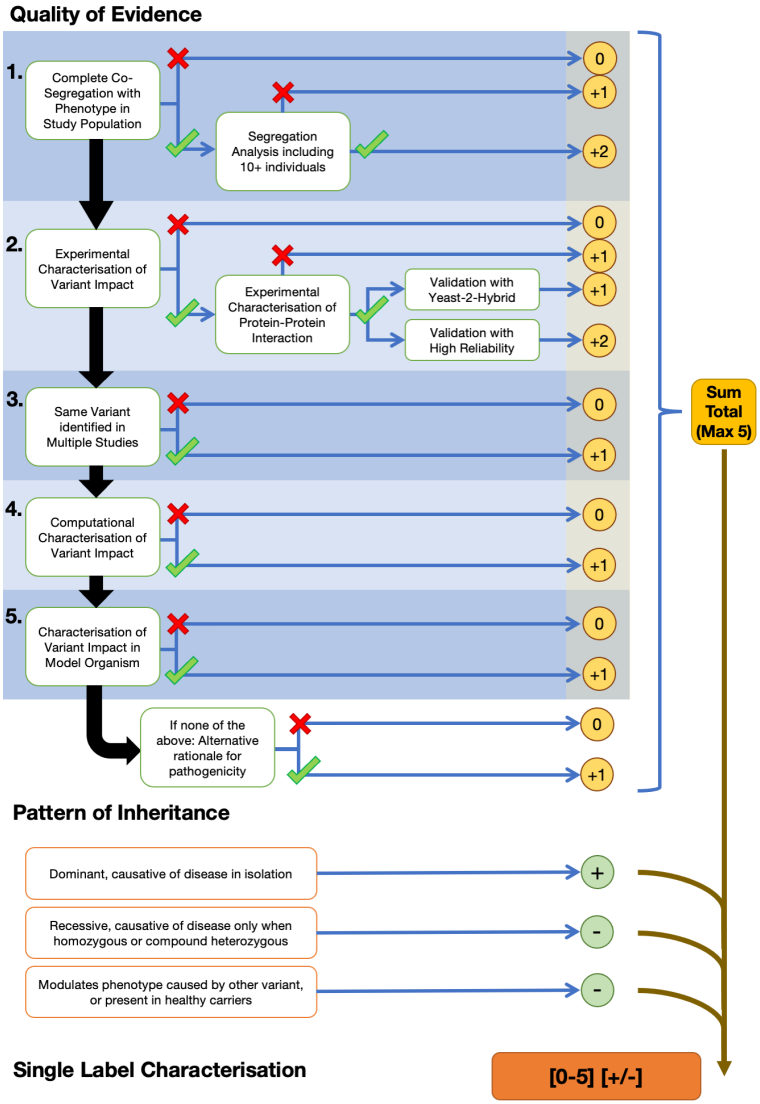


Figure 9: Reported Disease-Associated Missense Variant Labelling Flowchart

Flowchart elaborating upon our annotations of titin missense variants with a single label pathogenicity descriptor, as provided in Table 2. Each label consists of a Quality of Evidence score (0-5) and a Pattern of Inheritance (+/-). The Quality of Evidence score is a summation of scores from each of five categories of evidence typically presented for validation of the disease association, with a maximum value of five. If the score for a variant is equal to 0 after summation of all five categories, an alternative rationale for pathogenicity may be put forward, and the Quality of Evidence score set as 1; this is for scenarios where none of the five categories apply, but preponderance of other evidence is weighted towards favouring the disease association. For example, the variant occurs in a known disease-associated domain or location, or a known phenotype is exacerbated in the presence of both the variant in question and a second, causative variant.

identified those variants which may be reliably considered as pathogenic, and those which will necessitate further study to validate their hypothesised mechanisms, and to enable their use in disease-associated variant data sets.

While debates continue over the severity, penetrance, and validity of individual missense variants, the mutational map of titin continues to be filled, and patterns can begin to be drawn. However, the evidence gathered thus far is dwarfed by the evidence that is needed; large regions of titin remain uncharacterised and with functions unknown, and the patterns of isoform expression are not fully understood. The continued unravelling of the veil of titin-linked myopathies depends in large part on our continued ability to generate clear, reliable data on titin missense variants.

Table 1: Allele frequencies of reported pathogenic single amino acid variants in titin in populations

	Global	AFR	FIN	NFE	SAS	EAS	AMR	ASJ	Other
V54M	4.25E-05	.	.	.	.	5.53E-04	2.82E-05	.	.
A82D	.	.	.	.	.	.	.	.	.
A178D	.	.	.	.	.	.	.	.	.
E238Q	.	.	.	.	.	.	.	.	.
Q466R	2.39E-05	3.69E-04	.	.	.	.	.	.	.
R740L	.	.	.	.	.	.	.	.	.
A743V	.	.	.	.	.	.	.	.	.
W976R	.	.	.	.	.	.	.	.	.
V1034M	8.16E-04	.	1.91E-03	1.12E-03	6.86E-04	.	2.54E-04	.	1.11E-03
T2014A	1.45E-04	.	.	.	.	2.05E-03	.	.	.
T2896I	.	.	.	.	.	.	.	.	.
S4116Y	.	.	.	.	.	.	.	.	.
G4714D	8.23E-05	.	.	1.56E-05	.	1.08E-03	.	.	.
S4780N	8.93E-05	.	.	7.82E-06	1.31E-04	1.03E-03	.	.	.
L4854F	4.02E-06	.	.	.	.	.	2.90E-05	.	.

C5054R	.	.	.	.	.	.	.	.	.
Y9275C	6.43E-05	.	.	1.41E-04	.	.	.	.	.
R9744H	1.78E-05	4.14E-05	.	3.12E-05	.	.	.	.	.
R9848Q	1.21E-05	.	.	.	3.27E-05	1.11E-04	.	.	.
A9980T	6.03E-04	.	.	7.82E-06	3.00E-04	8.06E-03	.	.	1.41E-04
H10092Y	4.04E-03	2.89E-04	8.83E-03	6.01E-03	2.71E-03	.	5.09E-04	1.16E-03	2.94E-03
S13702P	.	.	.	.	.	.	.	.	.
A13715E	.	.	.	.	.	.	.	.	.
R14640C	2.82E-05	.	.	6.23E-05	.	.	.	.	.
N16133K	.	.	.	.	.	.	.	.	.
N16429K	.	.	.	.	.	.	.	.	.
W16471C	3.88E-04	4.14E-05	4.83E-04	6.51E-04	.	.	1.42E-04	9.74E-05	8.47E-04
Y16686C	.	.	.	.	.	.	.	.	.
C17051R	4.03E-06	.	.	.	3.27E-05	.	.	.	.
L18237P	6.43E-04	2.07E-04	1.20E-04	1.13E-03	.	.	6.22E-04	.	8.43E-04
A18983T	9.32E-05	.	.	1.81E-04	.	.	5.68E-05	.	1.41E-04
P19288R	.	.	.	.	.	.	.	.	.
I19517T	4.03E-06	.	.	8.92E-06	.	.	.	.	.

A19938T	8.07E-06	.	.	8.92E-06	3.27E-05	.	.	.	.
N19955I	.	.	.	.	.	.	.	.	.
A21147T	5.39E-04	.	4.03E-05	2.62E-04	3.30E-03	.	1.15E-04	2.94E-04	1.14E-03
A21877S	1.02E-04	.	.	5.41E-05	5.91E-04	.	.	.	1.67E-04
V22232E	.	.	.	.	.	.	.	.	.
G24621R	8.07E-06	.	.	.	3.27E-05	5.61E-05	.	.	.
R25480P	.	.	.	.	.	.	.	.	.
G27849V	.	.	.	.	.	.	.	.	.
R28118H	3.63E-05	.	.	4.46E-05	.	5.60E-05	8.71E-05	.	.
R29293C	1.93E-03	.	.	1.64E-04	1.63E-04	2.60E-02	.	.	8.42E-04
S29303G	.	.	.	.	.	.	.	.	.
E29590Q	4.02E-06	.	.	8.88E-06	.	.	.	.	.
L30639P	.	.	.	.	.	.	.	.	.
P30723S	6.18E-05	.	1.22E-04	1.03E-04	.	.	.	.	1.43E-04
K31268T	4.79E-04	.	1.92E-03	6.33E-04	.	.	.	.	7.02E-04
W31429R	.	.	.	.	.	.	.	.	.
P31709H	.	.	.	.	.	.	.	.	.
P31709R	.	.	.	.	.	.	.	.	.

C31712R	4.11E-06	.	.	9.10E-06	.	.	.	.	.
C31712Y	.	.	.	.	.	.	.	.	.
C31712W	.	.	.	.	.	.	.	.	.
W31729C	.	.	.	.	.	.	.	.	.
W31729L	.	.	.	.	.	.	.	.	.
W31729R	.	.	.	.	.	.	.	.	.
P31732L	1.21E-05	.	.	1.78E-05	3.27E-05	.	.	.	.
A31784V	.	.	.	.	.	.	.	.	.
N31786K	.	.	.	.	.	.	.	.	.
G31791D	.	.	.	.	.	.	.	.	.
G31791R	.	.	.	.	.	.	.	.	.
G31791V	.	.	.	.	.	.	.	.	.
R31847P	.	.	.	.	.	.	.	.	.
P33415L	.	.	.	.	.	.	.	.	.
R33903H	2.42E-04	8.27E-05	7.64E-04	6.26E-05	3.27E-05	2.00E-03	.	.	.
W34072R	8.11E-06	.	.	1.78E-05	.	.	.	.	.
V35643I	6.07E-05	.	.	6.24E-05	.	4.61E-04	.	.	.
M35859T	1.77E-03	8.26E-05	.	8.81E-04	8.82E-03	5.12E-05	2.54E-04	8.02E-03	2.52E-03

T35915P	.	.	.	.	.	.	.	.	.
W35930R	.	.	.	.	.	.	.	.	.
H35946P	.	.	.	.	.	.	.	.	.
I35947N	4.01E-06	.	.	8.85E-06	.	.	.	.	.
L35956P	.	.	.	.	.	.	.	.	.

Allele frequencies of reported variants in this review in gnomAD-2.1.1, both from the global population and from reported sub-populations. Sub-population ethnicities are given as follows: AFR *African;* FIN *Finnish;* NFE *European (excluding Finnish);* SAS *South Asian;* EAS *East Asian;* AMR *Latin American;* ASJ *Ashkenazi Jewish.*

Table 2: Single amino acid variants in titin reported in the literature

Variant	Domain	Associated Condition	Validation	Penetrance	Co-inheritance	Label	Reference
V54M	Ig-1	DCM	(+1) Y2H PPI assay (+1) MD simulation	Unknown		2+	Itoh-Satoh *et al.* (2002)[92] Kumar *et al.* (2017)[105]
A82D	Ig-1	CFTD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	R34807Sfs*7	3-	Rees *et al.* (2021)[35]
A178D	Ig-2	DCM, LVNC	(+1) Segregation analysis (+1) 10+ individuals (+2) GST PPI assay (+1) *in vivo* murine model	Full (8 affected, 14 total)		5+	Hastings *et al.* (2016)[96] Jiang *et al.* (2021)[71]
E238Q	NA	LGMD	(+1) Segregation analysis	Full (2 affected, 3 total)	Q466R	1-	Peddareddygari *et al.* (2022)[107]

Q466R	NA	LGMD	(+1) Segregation analysis	Full (2 affected, 3 total)	E238Q	1-	Peddareddygari *et al.* (2022)[107]
R740L	NA	HCM	(+1) Y2H PPI assay	Unknown		1+	Satoh *et al.* (1999)[91]
A743V	NA	DCM	(+1) Y2H PPI assay	Unknown (2 affected, 2 total)		1+	Itoh-Satoh *et al.* (2002)[92]
W976R	Ig-3	DCM	(+1) Segregation analysis (+1) 10+ individuals	Incomplete (11 affected, 35 total; 4 unaffected carriers)	R21201*	2-	Gerull *et al.* (2002)[106] Herman *et al.* (2012)[25]
V1034M	NA	MD	Association only	Unknown (2 affected, 3 total)	A18983T	0	Vasli *et al.* (2012)[108]
T2014A	NA	Possible titinopathy	Association only	Unknown	R33903H, c.1800+1G>A	0	Huang *et al.* (2021)[37]

T2896I	Ig-19	ARVC	(+2) Thermostability assay (+1) MD simulation (+1) *in vivo* murine model	Full (7 affected, 11 total)		4+	Taylor *et al.* (2011)[113] Anderson *et al.* (2013)[98] Bogomolovas *et al.* (2016)[100] Fleming *et al.* (2021)[114]
S4116Y	NA	HCM	(+1) Y2H PPI assay	Unknown		1+	Itoh-Satoh *et al.* (2002)[92] Matsumoto *et al.* (2006)[93]
G4714D	Ig-30	DCM	(+1) Segregation analysis	Full (2 affected, 2 total)		1+	Liu *et al.* (2008)[115]
S4780N	Ig-30	DCM	Association only	Unknown		0	Itoh-Satoh *et al.* (2002)[92]
L4854F	Ig-31	DCM	(+1) Segregation analysis	Incomplete/severe only (5 affected, 16 total)	LMNA:K219T	1-	Roncarati *et al.* (2013)[116]

C5054R	Ig-33	CFTD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	Y24599Mfs*12	3-	Rees *et al.* (2021)[35]
Y9275C	Ig-78	ARVC	Association only	Unknown		0	Taylor *et al.* (2011)[113]
R9744H	Ig-83	HCM	(+2) Co-IP PPI assay (+1) *ex vivo* histopathology (+1) Multiple studies	Unknown		4+	Arimura *et al.* (2009)[94] Lopes *et al.* (2013)[167]
R9848Q	NA	HCM	(+2) Co-IP PPI assay (+1) *ex vivo* histopathology	Unknown (2 affected, 2 total)		3+	Arimura *et al.* (2009)[94]
A9980T	Ig-84	Novel titinopathy	Association only	Unknown (2 affected, 3 total)	A19938T, R29293C	0	Sasaki *et al.* (2020)[127]
H10092Y	Ig-86	ARVC	Association only	Unknown		0	Taylor *et al.* (2011)[113] Costa *et al.* (2021)[126]

S13702P	Ig-87	DCM	(+1) Segregation analysis (+0) Mechanostability assay (+1) *in silico* assessment	Incomplete (2 affected, 3 total)		2-	Begay *et al.* (2015)[89] Zuo *et al.* (2021)[97]
A13715E	Ig-87	T1P	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown (2 affected, 2 total)	R11022*	3-	Rees *et al.* (2021)[35]
R14640C	Ig-97	DCM	(+1) Segregation analysis (+1) Mechanostability assay (+1) *in silico* assessment	Full (2 affected, 3 total)	E29590Q	3-	Begay *et al.* (2015)[89] Zuo *et al.* (2021)[97]
N16133K	Fn3-4	MmD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	R5308*	3-	Rees *et al.* (2021)[35]
N16429K	Fn3-6	AVB	(+1) Segregation analysis	Full (5 affected, 7 total)		1+	Liu *et al.* (2020)[140]

W16471C	Fn3-7	MmD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Full (2 affected, 2 total)	R17337*	3-	Rees *et al.* (2021)[35]
Y16686C	Fn3-9	RCM	(+1) Segregation analysis (+1) 10+ individuals	Full (5 affected, 13 total)		2+	Peled *et al.* (2014)[147]
C17051R	Fn3-12	MmD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Full (2 affected, 2 total)	Q32711Lfs*22	3-	Rees *et al.* (2021)[35]
L18237P	Fn3-20	CNM	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	D32441Ffs*1, W35453*	3-	Rees *et al.* (2021)[35]
A18983T	Fn3-25	MD	Association only	Unknown (2 affected, 3 total)	V1034M	0	Vasli *et al.* (2012)[108]

P19288R	Fn3-28	Possible titinopathy	Association only	Unknown	K4230Nfs*17	0	Huang et al. (2021)[37]
I19517T	Fn3-29	ARVC	Association only	Incomplete		0	Taylor *et al.* (2011)[113]
A19938T	Fn3-32	Novel titinopathy	Association only	Unknown (2 affected, 3 total)	A9980T, R29293C	0	Sasaki *et al.* (2020)[127]
N19955I	Fn3-32	Novel titinopathy	(+1) Muscle biopsy	Unknown	L30639P, c.44282-2A>G	1-	Yu *et al.* (2019)[141]
A21147T	Fn3-41	ARVC	Association only	Unknown	M35859T	0	Taylor *et al.* (2011)[113]
A21877S	Ig-125	ARVC	Association only	Unknown		0	Taylor *et al.* (2011)[113]

V22232E	Fn3-49	MmD-HD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Recessive (1 affected, 2 total)	N34020Tfs*9	3-	Chauveau *et al.* (2014)[48] Rees *et al.* (2021)[35]
G24621R	Fn3-67	Possible titinopathy	Association only	Unknown	Y4418*	0	Huang *et al.* (2021)[37]
R25480P	Fn3-73	CNM	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	R14679*	3-	Rees *et al.* (2021)[35]
G27849V	Fn3-90	MmD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	R21201*	3-	Rees *et al.* (2021)[35]
R28118H	Fn3-92	DCM	(+1) Segregation analysis (+1) *in silico* assessment	Full (2 affected, 2 total)		2+	Begay *et al.* (2015)[89]

R29293C	Fn3-101	Novel titinopathy	Association only	Unknown (2 affected, 3 total)	A9980T, A19938T	0	Sasaki *et al.* (2020)[127]
S29303G	Fn3-101	DCM	(+1) Segregation analysis (+1) *in silico* assessment	Full (2 affected, 2 total)	LMNA:c.936G>A	2-	Begay *et al.* (2015)[89]
E29590Q	Fn3-103	DCM	(+1) Segregation analysis (+1) *in silico* assessment	Full (2 affected, 3 total)	R14640C	2-	Begay *et al.* (2015)[89]
L30639P	Fn3-111	Novel titinopathy	(+1) Muscle biopsy	Unknown	N19955I, c.44282-2A>G	1-	Yu *et al.* (2019)[141]
P30723S	Fn3-112	Novel titinopathy	(+1) Segregation analysis	Incomplete (1 affected, 3 total)	FINmaj (35927delins-EVTW>VKEK)	1-	Evilä *et al.* (2014)[80]

K31268T	Ig-151	IH	(+1) Segregation analysis	Incomplete (5 affected, 9 total)		1-	Mihailov *et al.* (2017)[143]
W31429R	Fn3-117	TMD, DCM	(+1) Segregation analysis (+1) Multiple studies	Recessive (1 affected, 5 total)	R21209*	2-	Evilä *et al.* (2017)[142] Lopez-Bravo *et al.* (2021)[146]
P31709H	Fn3-119	HMERF	(+1) Hotspot domain	Unknown		1+	Palmio *et al.* (2019)[44]
P31709R	Fn3-119	HMERF	(+1) Segregation analysis (+2) Western blot assay (+1) Hotspot domain	Full (3 affected, 6 total)		4+	Palmio *et al.* (2014)[133] Hedberg *et al.* (2014)[130]

C31712R	Fn3-119	MFM, HMERF	(+1) Segregation analysis (+1) 10+ individuals (+1) Multiple studies (+2) Western blot assay (+1) Hotspot domain	Full (multiple studies)	5+	Ohlsson *et al.* (2012)[131] Pfeffer *et al.* (2012)[132] Toro *et al.* (2013)[134] Palmio *et al.* (2014)[133] Hedberg *et al.* (2014)[130] Pfeffer *et al.* (2014)[45] Uruha *et al.* (2015)[46] Yue *et al.* (2015)[135] Palmio *et al.* (2019)[44] Morais *et al.* (2020)[137] Huang *et al.* (2021)[37]
C31712W	Fn3-119	HMERF	(+1) Hotspot domain	Unknown (2 affected, 2 total)	1+	Yeo et al. (2021)[168]
C31712Y	Fn3-119	MFM, HMERF	(+1) Hotspot domain	Unknown	1+	Uruha *et al.* (2015)[46]

W31729C	Fn3-119	HMERF	(+1) Segregation analysis (+2) Western blot assay (+1) Hotspot domain (+1) Multiple studies	Full (2 affected, 4 total); Full (4 affected, 4 total)	5+	Palmio *et al.* (2014)[133] Hedberg *et al.* (2014)[130] Bugiardini *et al.* (2018)[136] Palmio *et al.* (2019)[44]
W31729L	Fn3-119	MFM, HMERF	(+1) Segregation analysis (+1) 10+ individuals (+1) Hotspot domain	Full (5 affected, 10 total)	3+	Izumi *et al.* (2013)[138]
W31729R	Fn3-119	HMERF	(+2) Western blot assay (+1) Hotspot domain (+1) Multiple studies	Unknown (1 affected, 1 total)	4+	Palmio *et al.* (2014)[133] Hedberg *et al.* (2014)[130] Huang *et al.* (2021)[37]

P31732L	Fn3-119	CM, HMERF	(+1) Segregation analysis (+1) 10+ individuals (+1) Multiple studies (+2) Western blot assay, thermostability assay (+1) Cellular expression, solubility, localisation (+1) Hotspot domain	Incomplete (multiple studies)	5-	Vasli *et al.* (2012)[108] Palmio *et al.* (2014)[133] Hedberg *et al.* (2014)[130] Pfeffer *et al.* (2014)[45] Yue *et al.* (2015)[135] Rees *et al.* (2021)[35] Sano *et al.* (2022)[139]
A31784V	Fn3-119	HMERF	(+1) Segregation analysis (+1) Hotspot domain	Full (6 affected, 8 total)	2+	Palmio *et al.* (2019)[44]
N31786K	Fn3-119	HMERF	(+1) Segregation analysis (+1) Multiple studies (+1) Hotspot domain	Full (2 affected, 7 total)	3+	Pfeffer *et al.* (2014)[45] Palmio *et al.* (2019)[44]
G31791D	Fn3-119	MFM, HMERF	(+1) Hotspot domain	Unknown	1+	Uruha *et al.* (2015)[46]

G31791R	Fn3-119	MFM, HMERF	(+1) Hotspot domain (+1) Multiple studies	Unknown		2+	Toro *et al.* (2013)[134] Uruha *et al.* (2015)[46]
G31791V	Fn3-119	MFM, HMERF	(+1) Hotspot domain	Unknown		1+	Uruha *et al.* (2015)[46]
R31847P	Fn3-120	MmD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation	Unknown	R5430*	3-	Rees *et al.* (2021)[35]
P33415L	Fn3-131	ARVC	Association only	Unknown		0	Taylor *et al.* (2011)[113]
R33903H	Kinase-1	Possible titinopathy	Association only	Unknown	T2014A, c.1800+1G>A	0	Huang *et al.* (2021)[37]
W34072R	Kinase-1	CM, MmD-HD	(+2) Thermostability assay (+1) Cellular expression, solubility, localisation (+1) Multiple studies	Recessive (1 affected, 2 total); Unknown)	E2989Efs*4 / R7796*	4-	Chauveau *et al.* (2014)[48] Rees *et al.* (2021)[35]

V35643I	Ig-167	HCM	(+1) Segregation analysis	Unknown (6 affected, 18 total)		1+	Luo *et al.* (2022)[154]
M35859T	NA	ARVC	Association only	Unknown	A21877S	0	Taylor *et al.* (2011)[113]
T35915P	Ig-169	TMD	(+1) Hotspot domain	Recessive (3 affected, 4 total)	E28338fs	1-	Evilä *et al.* (2017)[142]
W35930R	Ig-169	LGMD2J	(+1) Segregation analysis (+1) Hotspot domain	Recessive (3 affected, 6 total)		2-	Zheng *et al.* (2016)[152]
H35946P	Ig-169	TMD, LGMD2J	(+1) Segregation analysis (+2) Thermostability assay (+1) Hotspot domain	Full (5 affected, 7 total)		4+	Pollazzon *et al.* (2010)[150] Rudloff *et al.* (2015)[95]

I35947N	Ig-169	TMD, LGMD2J	(+1) Segregation analysis (+2) Thermostability assay (+1) Hotspot domain	Incomplete (6 affected, 9 total, 1 unaffected carrier); Recessive (2 affected, 3 total)	Q22507*	4-	Van den Bergh *et al.* (2003)[149] Rudloff *et al.* (2015)[95] Evilä *et al.* (2017)[142]
L35956P	Ig-169	TMD, LGMD2J	(+1) Segregation analysis (+2) Thermostability assay (+1) Hotspot domain	Full (3 affected, 7 total)		4+	Hackman *et al.* (2002)[41] Rudloff *et al.* (2015)[95]

Missense variants in titin reported in the literature, with relevant annotations described in this review. Each variant is identified by its residue change, position according to the inferred complete isoform, parent domain, and reported condition with which it is associated. We include a summarised notation of validation methods used, as well as the reported penetrance of the variant and the size of the cohort(s) used to make that determination. Any variants co-inherited with the titin variant in question in the affected individuals are included. Finally, a labelling scheme is provided to summarise the confidence of the label of damaging variant, and the magnitude of the pathological consequence; further explanation may be located by reference to Figure 9.

2 TITINdb2 and Annotating Missense Variants

2.1 Synopsis

- In this chapter, I introduce the annotation of identified missense variants in titin using computational and bioinformatics methods, and the advantages and disadvantages of this approach.
- I introduce a new version of the online variant annotation database for titin missense variants, TITINdb, as a publicly accessible repository for the known population and literature-sourced variants in titin, with collected annotations from a wide array of variant impact predictors, as well as improved structural coverage at the single and tandem domain level.
- I demonstrate the application of computational prediction of variant impacts through a series of case studies of analysis of titin missense variants, and discuss the need for improvement of predictors to accommodate the increasing quantity and breadth of data available.

2.2 Introduction

I have previously established the importance of titin as a major disease-related gene, as well as the particular challenges with ascribing pathogenicity to single missense variants associated with disease. In Chapter 1, I dissected the various reported pathogenic variants in the scientific literature and the evidence presented in support of each of their roles in disease, and the mechanisms by which they have been proposed to act. However, these 60 variants are only a tiny fraction of the 683,829 theoretical single amino acid variants (SAVs) present across the 35,991 residues of the full-length titin gene; the sheer size of the protein, and the number of variants that will inevitably be found in even healthy individuals, precludes

the experimental characterisation of all variants identified in a given population cohort. This bottleneck between the identification and accurate annotation of variants identified in individuals with disease is the greatest obstacle in uncovering the specific genetic elements contributing to that condition.

Given that the process of experimental characterisation of specific variants in titin associated with disease is both time- and labour-intensive, this represents an ideal system upon which computational methods may be used to boost efficiency. Here, rather than expressing each mutant protein separately *in vitro* or *in vivo*, the consequences of variants are assessed based on their inherent molecular properties, using the knowledge generated from experimental studies as both guide and validation. To elaborate, many features of variants can serve to support or else eliminate the candidacy of that variant for causation of pathogenic consequences: the frequency of that variant in both global and ethnically matched populations; the evolutionary propensity of the amino acid change to occur at homologous positions; the local structural context of the residue position; the assessment by the many existing variant impact algorithms and tools that incorporate these features.

The principal difficulty in parsing the wealth of possible data is in the quantity of different features that may be used to annotate even a single variant, and how to collect and interrogate them in a single space. In order to address these issues, the web server TITINdb was previously established[30]; here, a user could visualise any residue in the 302 titin domains (169 immunoglobulin, 132 fibronectin type-III, 1 protein kinase) on an atomic-level experimental structure or model of that domain, and to extract information on its contacts, solvent accessibility, and the frequency of observed variation at that position. Furthermore, this functionality was integrated with both population variant data and reported pathogenic variants in the literature, allowing for any non-synonymous missense variant of interest to be mapped to its respective position on its parent domain, with the goal of facilitating the prioritisation of candidates for subsequent experimental characterisation.

Since its introduction, TITINdb has been extensively used as a resource by both experimental and computational groups working on titin (examples [35, 169]). In order to develop a more systematic overview of the mutational landscape of titin than reported pathogenic variants alone, the original TITINdb was used as a base; however, with the introduction of new population databases, new methods of protein structure prediction, and many new variant impact or pathogenicity predictors since the original release of TITINdb, it was decided that the existing web server should be updated with a comprehensive overhaul of existing data and new features to better consolidate the wealth of information available. To summarise, the current build of TITINdb2 includes (i) an increase in computational annotation from 4 to 19 variant impact predictors, all available through a new custom data table dialogue; (ii) thorough structural coverage of single domains with AlphaFold2 predicted models and new tandem-domain models where applicable; (iii) new predicted domain-domain interface annotations, marking residues buried by adjacent domains in the full titin protein; (iv) an expanded *in silico* saturation mutagenesis assessment incorporating 3 impact predictors for all canonical domain residues; (v) a comprehensive overhaul of available data, including population data sources and variants reported pathogenic in the literature.

2.3 TITINdb Application Notes

2.3.1 Updated Features and Data

Data Sources: The original release of TITINdb included known SAVs in titin from three sources: variants reported as pathogenic in the literature (via PubMed), population variants from the Genome Aggregation Database (gnomAD)[84, 85] and population variants from the 1000 Genomes Project (1KGP)[87]. The new release includes these previous sources as well as including newly collected data; as gnomAD has since released a subsequent, separate short variant data set, both the original (gnomAD release 2.1.1; exomes + genomes) and newer (gnomAD release 3.1.2; genomes only) data sets have been included as separate categories. Thus, in total TITINdb2 contains 25,135

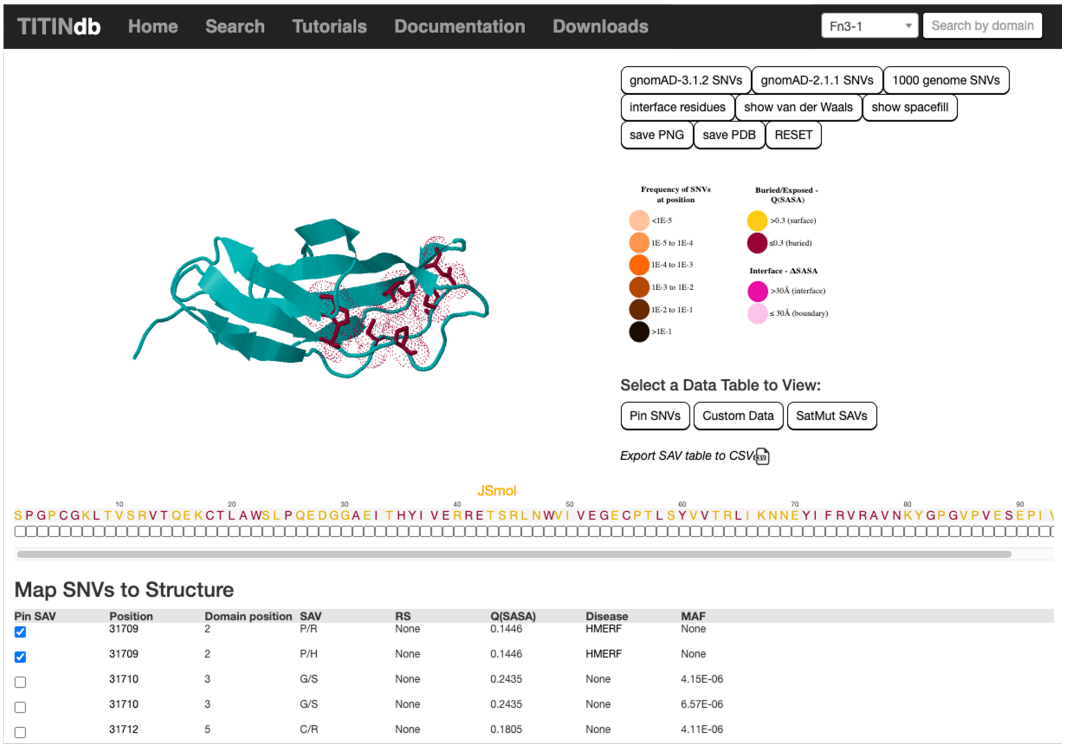


Figure 10: Screen Capture of TITINdb2 developer build, domain Fn3-119

Screen capture of the Domain-level page for titin domain Fn3-119. At the top of each page is a representation of the 3D structure of the domain in Jmol; this can be rotated, enlarged, and manipulated with simple mouse commands. To the right of this window is a control panel which allows for the population of the 3D structure with relevant subsets of known variants or different visual representation. The data tables, which are found below the structure window, allow the user to view and compare information from different annotations with respect to SAVs both in existing databases, and across all possible mutational variation. Importantly, the user can highlight the 3D positions of particular variants by selecting the respective check box, either in the "Pin SNV" table for known SAVs, or the sequence selection dialog underneath the structure window.

distinct, annotated missense variants, of which 74 variants are reported pathogenic in the literature.

AlphaFold2 Structural Coverage. The structural coverage of titin domains has been supplemented by the release of AlphaFold2 predictions for the human proteome[170, 171]; however, titin as a complete protein is too large to be represented in a single model. Thus, I have extracted the full predictions for titin from the complete human proteome predictions available at AlphaFoldDB[171]; here, titin is represented as 1400-residue overlapping "contigs" as input for which AlphaFold2 structure prediction. The structures of individual domains were extracted from each contig according to the domain boundaries established in the previous release of TITINdb[30]; each individual domain was typically represented in 6 to 7 individual contigs, which together form an ensemble of predicted structures.

In addition to single domain models and structures, tandem Ig domains, where two Ig or Fn3 domains are connected by a short (1-8 residue) linker, are included as separate structures to enable the investigation of residues buried at domain interfaces. Thus, TITINdb2 offers rapid access to the AlphaFold2 domain structures extracted from the H. sapiens predicted proteome, alongside quality assessment metrics and mapping. In addition, should the user wish to access the original AlphaFold2 prediction for the 1400-residue contig, the full PDB structures are available for easy download by clicking on the "AlphaFold construct" entry for a given domain.

Domain-Domain Interface Annotation. A newly added feature in TITINdb2 is the annotation of domain-domain interface residues. Single-domain models allow for annotation of residues as "exposed" or "buried" based on their quotient solvent-accessible surface area (Q(SASA), as calculated by POPS[172]); however, this can give an inaccurate picture of their actual solvent accessibility. The titin protein *in vivo* is composed of a chain of many closely linked domains, such that residues located at the interfaces between two titin domains may be buried by other residues in the linker or adjacent domain. Thanks to the provision of AlphaFold2 structural models, TITINdb2 now includes models of all tandem two-domain constructs in the whole titin

protein, including annotations of these interface residues.

$$ResidueLabel = \begin{cases} EXPOSED & Q(SASA)_{single} > 0.3 \\ BURIED & Q(SASA)_{single} \leq 0.3 \\ INTERFACE & SASA_{single} - SASA_{tandem} \geq 30\text{Å} \end{cases} \quad (1)$$

In brief, a residue is labelled as "interface-buried" based on the difference between the calculated SASA for a residue in the domain alone and that same residue in the multi-domain tandem. Analysis indicates that 7.4% of domain residues in titin are interface-buried by this definition, as well as 2 of the 55 reported pathogenic missense variants in domains in the literature (see Methods).

Annotation with new Variant Impact Predictors. In addition to the previous annotation of SAVs with the prediction tools Condel[173], mCSM[174], DUET[175], and PolyPhen2[144], I have updated the database with a broader range of variant impact predictors, including newer tools that incorporate different data and architectures to these more established methods. Notably, I have computed pathogenicity prediction scores for all possible SAVs with Rhapsody[176], for all domains represented in the canonical N2BA isoform (i.e. excluding domains Ig25, Ig26). Each possible amino acid substitution is given a predicted score between 0 (most likely neutral) and 1 (most likely damaging), giving a broad view of the mutational landscape across titin. Moreover, I have performed this same *in silico* saturation mutagenesis with PolyPhen2. Thus, TITINdb2 provides computational saturation mutagenesis predictions calculated from three tools: the sequence-based pathogenicity predictor PolyPhen2; the structure-based stability predictor mCSM, and the combined sequence, structure, and network-based pathogenicity predictor Rhapsody. Each of these predictions can be leveraged to assess the predicted impact of any given variant on the domain in question.

Variants in the database have additionally been annotated with 14 pre-computed pathogenicity prediction scores sourced from dbNSFPv4[177], which greatly expands the range of annotations available for titin variants. All the available predictors can

be selected or de-selected in the visible table using the new Custom Data Table functionality. A full list and information on the predictors included is given in 4.

2.3.2 Applications

Visualise Structural Positions with "Pin SNVs". As in the previous version, TITINdb2 retains the ability to view the structural location of a variant of interest by "pinning" the position in the Jmol[178] molecule viewer. This can be used to highlight locations in a domain where reported pathogenic variants are found, as seen in Figure 10; the A-band domain Fn3-119, a well-established hotspot for pathogenic variants associated with hereditary myopathy with early respiratory failure (HMERF)[44, 78] . Further options for visualisation are provided via the buttons to the right of the molecule viewer, including colour annotation with the frequency of variation at the designated position in each variant database, and the newly implemented "interface residues" functionality which indicates those positions that are typically buried in the presence of adjoining titin domains.

In Figure 10, those structural positions in the Fn3-119 domain where reported damaging variants are found are highlighted using the "Pin SNVs" table. The user can view which variants have been associated with HMERF and can select or deselect these residues as desired. They can then compare these different locations. Alternatively, should the user be interested in a specific domain, disease, or transcript position, TITINdb allows the user to search for these key data directly; for example, if a novel variant were to be identified in a particular isoform, by searching for that position in TITINdb, the user can identify any reported SAVs at that location, and the domain in question, in a single step.

Access Variant Annotations with "Custom Data Table". The field of variant impact prediction has continued to expand since the original conception of TITINdb, with a vast array of proposed prediction methods making use of different features and

computational approaches to distinguish damaging from neutral variants. As such, I have expanded the range of annotations for known titin SAVs by combining data from pre-computed variant prediction databases with in-house annotation pipelines; these can be selected as desired, as well as other annotations including sub-population allele frequencies and disease-related comments, using a new dialog option under "Custom Data Table". A summary of prediction methods included in TITINdb is given in 4.

Variant impact predictors are commonly used in variant prioritisation pipelines, where likely candidates for disease-associated variants are filtered from whole-genome or whole-exome sequencing data; a good example of such an approach is described in Begay *et al.* (2015)[89], where the annotation pipeline ANNOVAR[88] was used to identify likely candidate variants in titin from sequencing data from patients with dilated cardiomyopathy (DCM). Two variants were identified in titin's I-band, S13702P (Ig-87) and R14640C (Ig-97); interestingly, bioinformatics annotation tools available to TITINdb predict that R14640C is more damaging (Rhapsody (0.867), DUET ($-1.68 kcal mol^{-1}$), mCSM ($-1.83 kcal mol^{-1}$)) than S13702P (Rhapsody (0.647), DUET ($-0.375 kcal mol^{-1}$), mCSM ($-0.372 kcal mol^{-1}$)). Indeed, these two variants were subsequently characterised by atomic force microscopy[97], which revealed that while R14640C significantly destabilised its parent domain, S13702P showed only moderate destabilisation. This demonstrates an application of TITINdb as a source for prioritisation of novel variants using easily accessible frequency and impact prediction annotations.

Compare Variant Impact across Positions with "SatMut SAVs". While annotations of individual variants give an idea of the impacts and characteristics of that specific variant, they must be contextualised by comparison to other variants in the same and other domains. Thus, *in silico* saturation mutagenesis offers a way to view the full mutational landscape of titin, with pre-computed values to greatly reduce the time taken for this analysis. For example, one can view the different predicted impacts of all possible amino acid substitutions at a position of interest; the three available prediction values for this approach are calculated using Rhapsody (pathogenicity prediction),

PolyPhen-2 (pathogenicity prediction), and mCSM (protein stability prediction). These values are available for all possible amino acid substitutions within the single domain boundaries of all titin domains in the canonical N2AB isoform (i.e. excluding Ig25, Ig26).

2.4 Methods

2.4.1 Data Collection

Population missense variant data is sourced from three databases, the 1000 Genomes Project (1KGP)[87], gnomAD-2.1.1 and gnomAD-3.1.2[84]. In the previous build of TITINdb, missense variants in the 1KGP data set were retrieved using the online data slicer[87] and mapped to the inferred complete isoform; these originate from 2504 nominally healthy individuals. The gnomAD short variant data sets v2.1 and v3.1 are both aggregate databases comprising many large-scale genome sequencing studies; the v2.1 data set includes data from 125,748 exomes and 15,708 genomes from 141,456 unrelated individuals, while the v3.1 data set includes data from 76,156 genomes from unrelated individuals. Both data sets were filtered using the online browser tool to include missense variants only, then were downloaded separately; the two data sets are mostly non-overlapping but contain some duplicate information. As such, allele frequencies and sub-population frequencies for both the v2.1 and v3.1 populations are given in separate data entries within TITINdb; note that the variant frequency given has been modulated to account for sequencing coverage within the raw data.

While individuals with severe paediatric disease and their close relatives are excluded from gnomAD, the data sets also include disease-specific genetic studies, so their inclusion in the population data does not necessarily preclude pathogenicity. However, it should be stressed that the population frequency for these variants should not be higher than that in the general population, and the gnomAD data set should not be enriched for damaging

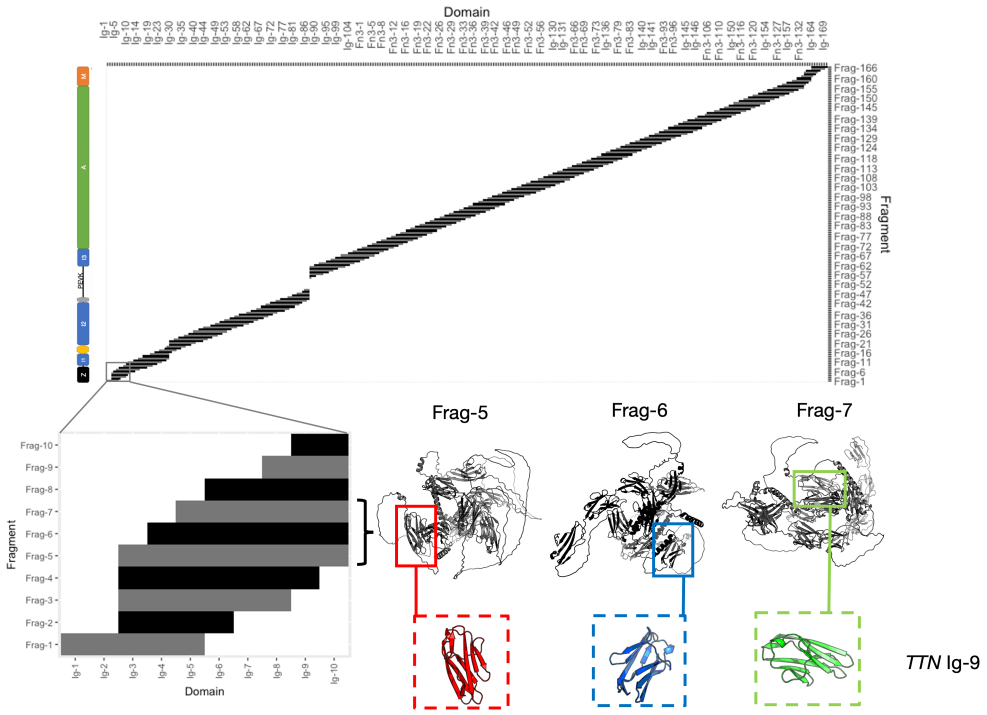


Figure 11: Overview of domain model extraction from AlphaFold2 proteome

A workflow demonstrating the extraction of AlphaFold2 domain structures from whole protein fragment structure files. Titin is represented in AlphaFoldDB by way of 166 overlapping fragments of 1400 residues; this can be visualised as a coverage map of all structured domains in the TTN gene. The zoom-in window in the bottom left shows this in greater detail; for example, the Z-disc domain Ig-9 is represented in fragments 5-10 in this window, while both Ig-1 and Ig-2 are only represented in fragment 1. By going into each fragment structure and extracting the residues within the domain boundaries established, one structural model is obtained for every domain in titin for each fragment in which it is represented (Ig-9 in fragments 5, 6, and 7 shown here).

variants. Finally, while all three databases give sub-population allele frequencies, the ethnic populations considered for each database differ; most notably, the sub-population "European" is given for 1KGP variants, whereas both gnomAD v2.1 and v3.1 split this sub-population into "Finnish" and "Non-Finnish European". Further information can be found on the online Documentation page.

2.4.2 Extraction of Predicted Domain Structures from AlphaFoldDB

The coverage of experimentally solved titin domains in the literature has continued to expand, with 52 entries in the PDB corresponding to domains in titin, covering 35 domains in total (see 3). However, there are still 267 domains in titin that lack structural coverage. In the previous build of TITINdb, a computational pipeline was constructed to define the domain boundaries for all detected domains in titin's inferred complete (IC) isoform, and then to model those domains using the template-based modelling software, Modeller. In brief, domains were defined in the IC isoform by scanning with HMMER[179] against Pfam seed libraries and validated by generation of sequence logos (Weblogo) from multiple sequence alignments of Ig/Fn3 domains thus defined (T-Coffee) and alignment with existing experimental structures of homologous titin domains. These individual domain sequences in FASTA format were then supplied to Modeller as input for template search, homology modelling, and model assessment steps. The full pipeline, including schematic figure, can be found in the original paper[30].

Recent advances in the field of protein structure prediction have culminated in the release of AlphaFold2, a Deep Learning-based software capable of *in silico* predictions of protein structural models from sequence data comparable in quality to solved experimental structures[170]. Moreover, the authors have released the protein structure predictions for all proteins in the proteomes of many species for the benefit of the structural biology community, via the online database AlphaFoldDB[171]. However, the size of titin renders any attempt to model the protein in its entirety impractical; the maximum size of protein that can be predicted within the memory limitations of available GPUs, approximately 1400 residues, is considerably smaller than titin's 35,991 residues. Thus, predictions for titin domains are not readily accessible through the web portal.

The full proteome predictions for *H. sapiens* were downloaded from AlphaFoldDB and the entries corresponding to the UniProt entry for titin's N2AB isoform (UniProt ID Q8WZ42) were extracted. These are organised into what I denote "contigs" of overlapping

predictions; a sliding window of 1400 residues is taken from titin and used to compute a predicted structure, moving in "steps" of 200 residues. Thus, the whole titin N2AB sequence is represented in 166 contigs, with each individual domain represented fully in 6-7 contigs. The coverage of the 302 domains in titin across the 166 contigs is visualised in Figure 11; the structures extracted confirm the previous assessment of locations of domains in titin.

A pipeline was developed to identify and extract all single and tandem domains from these contigs, using the previously established domain boundaries. Tandem domains were defined as successive domains separated, according to the domain boundaries, by a linker of 10 or fewer residues. The pipeline, written in Python 3.8 with the use of the PyMol module[62], took the sequences for the single and tandem domains in FASTA format as input and searched for a sequence match within the sequences for the supplied contigs; if a successful match was found, that segment of the contig was extracted and saved as a new PDB file. The pipeline produced a total of 1925 single-domain and 1624 tandem-domain structures. As each of the models for a given titin domain were produced in a slightly different chain context, there are differences between them; this is especially pronounced for tandem domains with long linkers. As such, the web server provides for visualisation of variants on all the single and tandem domains extracted from contigs, as well as the capability to download the original contigs from the structure selection page.

As the UniProt canon sequence was used by AlphaFold2 to model all contigs, there is no coverage of the novel exon domains Ig-25 or Ig-26. To remedy this, I used an adaptation of the AlphaFold2 pipeline distributed through Google Colaboratory, ColabFold[180], which not only allows for remote access to GPUs for prediction, but also incorporates the slightly less extensive but much faster multiple sequence alignment tool MMseqs2[181], rather than the jackhmmer[179] algorithm used by AlphaFold2. The models produced are also not relaxed by an Amber force field, as per the standard prediction pipeline, because of memory limitations.

2.4.3 Validation and Comparison of Model Structures

In order to compare the AlphaFold2-derived models to the previously generated homology models and PDB experimental structures, I calculated two measures of model quality using a local version of the MolProbity[182] pipeline in Python and Phenix. The first of these, the MolProbity score, is an estimate of the effective resolution of the model were it an experimentally solved structure; it incorporates the MolProbity clash score with Ramachandran and rotamer outliers to give a single value which reflects the experimental resolution at which such a score would be on average. Thus, the closer the value to zero, the better the assessed quality of the model. The second measure, Ramachandran Z-score[183], summarises the Ramachandran outliers/favoured in a single value, where $abs(Z) \leq 2$ would generally be considered to represent a good quality model. The MolProbity score was calculated for all AlphaFold2 and Modeller models and compared against the actual experimental resolution of X-ray crystal structures in the literature.

2.4.4 Annotation of Missense Variants

Missense variants from population databases and disease-associated variants in the literature have been annotated with various computational tools reflecting structural data and predicted impact on the parent domain. Most of these variants are sourced from general databases with pre-computed predictions; on the TITINdb website, variants that localise to each domain are annotated with the pathogenicity prediction scores Condel[173] (via FannsDB), REVEL[184], and MVP[185] (via dbNSFPv4). Values for 12 other pathogenicity predictors from external databases are included in the raw database available through download (see 4).

As with the first build of TITINdb, the predicted protein stability change on mutation in $kcalmol^{-1}$ has been predicted for all SAVs resulting from nsSNVs with the DUET[175] software, and for all possible SNVs at all domain positions with mCSM[174]. This *in silico* saturation mutagenesis gives a picture of the impact of different mutations at different posi-

tions across the whole breadth of possible mutations in titin. Thus, in addition to mCSM, I computed a saturation mutagenesis for two additional predictors of pathogenicity, giving a saturation mutagenesis table of three predictors. The first of these, PolyPhen-2[144], is a popular sequence-based predictor with limited use of structural data; the second, Rhapsody[176], combines the sequence-based features computed by PolyPhen-2 with additional structural and network dynamics features calculated from supplied protein structures to give a final prediction. As Rhapsody requires an input structure, I built a simple pipeline around the Rhapsody python distribution to compute Rhapsody scores for all positions within the AlphaFold2 models with the optimal MolProbity score for each titin domain. These predictions exclude the non-canonical domains Ig-25 and Ig-26, which are absent from the canonical UniProt sequences required to calculate PolyPhen-2 and Rhapsody scores.

In addition to predicted pathogenicity scores and protein stability change, variants in domains are annotated with predictions relating to protein-protein interactions. Each variant is assessed for whether it is predicted to map to a protein-protein interaction site with SPPIDER[186], which is expressed as a binary classification. Variants in those domains where an experimental structure exists for that domain in complex with an interaction partner (PDB 1YA5 for Ig-1/Ig-2 in complex with TCAP; PDB 3KNB for Ig-169 in complex with OBSL1; PDB 4UOW for Ig-169 in complex with OBSL1; PDB 4C4K for Ig-169 in complex with OBSCN) have been assessed with mCSM-PPI[174] to predict the impact of that mutation on the binding affinity between the two proteins.

2.4.5 Definition of Structural Regions of Domains

In the previous build of TITINdb, the quotient solvent-accessible surface area – Q(SASA) – was calculated for all residues in all domains of titin, and used to annotate the variants in the database. This was performed automatically by the tool POPS[172], using the default probe radius of 1.4Å. Residues with $Q(SASA) > 0.3$ were defined as exposed/surface residues, while those with $Q(SASA) \leq 0.3$ were defined as buried/core residues. With the availability

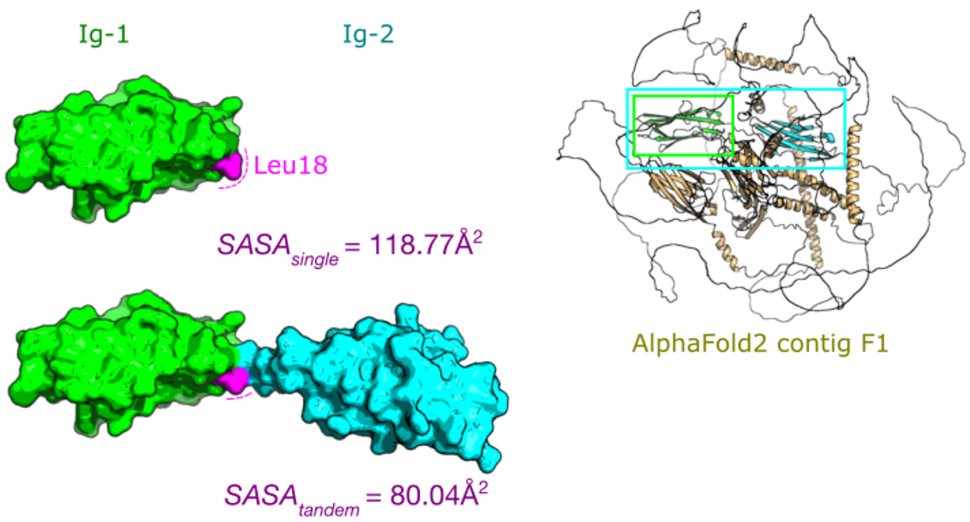


Figure 12: Definition of interface-buried residues in TITINdb2

Each residue within folded domain boundaries in titin was assessed for its presence in a domain-domain interface within AlphaFold2 structural models. As seen above, in each "contig" or "fragment" structure supplied by AlphaFoldDB, a representation of multiple successive domains are modelled; by taking the solvent accessibility of a residue in a single structure (example here: leucine-18 in domain Ig-1) and comparing it to the solvent accessibility of the same residue in the two-domain structure, the consequences of the presence of the second domain on solvent accessibility can be quantified. Here, leucine-18 has a solvent-accessible surface area of 118.77Å^2 *in the single domain, and* 80.04Å^2 *in the tandem; thus, its* $\Delta SASA$ *is calculated as* 38.73Å^2*, and as such is considered an interface-buried residue.*

of tandem titin domains through AlphaFold2, I additionally calculated interface regions for all tandems defined according to the previous boundaries. For each residue in the tandem, the SASA value of the residue in the tandem domain model was subtracted from the SASA value of that same residue in the single domain to give the change in SASA, or $\Delta SASA$, an adaptation of the method used in POPSCOMP[187] for protein complexes. Thus, a value of $SASA_{single} - SASA_{tandem} = \Delta SASA \geq 30$Å for a particular residue would be considered an interface-buried residue. An example of this workflow is provided in Figure 12.

2.5 Results

2.5.1 AlphaFold2 Model Quality Assessment

The single domain AlphaFold2 models compared very favourably to the homology models and experimental structures, with the best models approaching an estimated resolution of around 0.5 Å, and all models under 2.5 Å (see Figure 13a). Similarly, the AlphaFold2 models are considered very good models based on the Rama-Z score, although here the homology models are comparable in quality to the average AlphaFold2 model, and only significantly lower quality than the AlphaFold2 models judged best (see Figure 13b). For the ColabFold models, the quality assessment is similar to that of the AlphaFold2 models, with MolProbity scores of 1.57 (Ig-25) and 1.70 (Ig-26). Overall, the provision of AlphaFold2 models has increased the quality of the structural coverage for titin.

I caution the reader that Ig-112 remains a problematic domain for modelling. This domain contains a large insert within a loop region, which is poorly modelled by all methods; in the previous build of TITINdb, it could not be modelled to an acceptable standard using Modeller, and as such was modelled with I-TASSER. The AlphaFold2 models for Ig-112 satisfy geometric constraints adequately, and do not appear as outliers in terms of their MolProbity and Rama-Z scores; however, be aware that its prediction for the loop does not reflect any existing experimental knowledge.

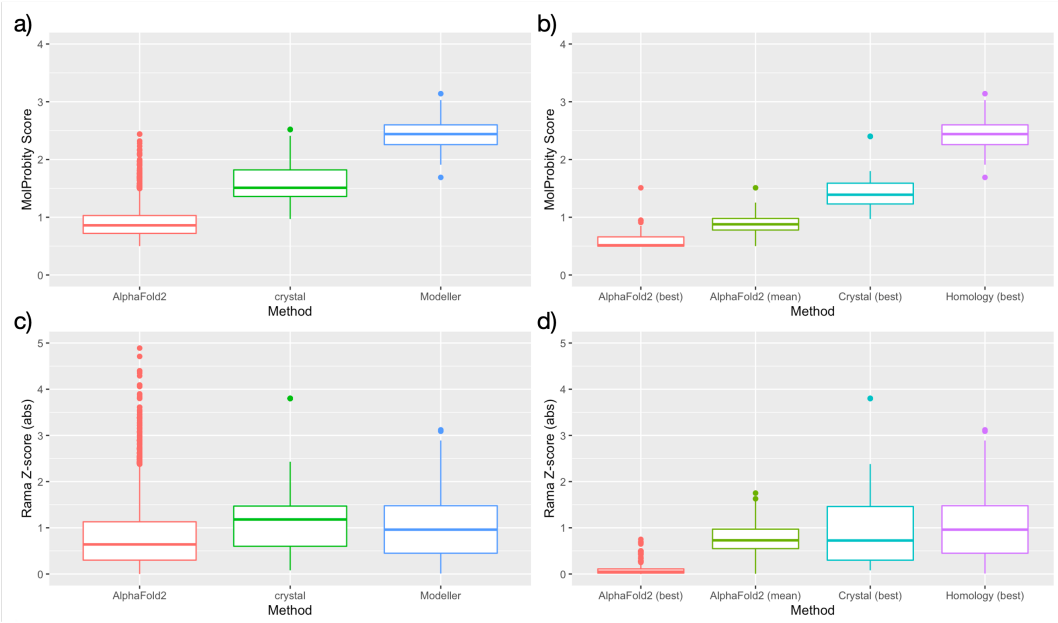


Figure 13: Experimental structure and predicted model quality metrics for structural classes in TITINdb2

A comparison of quality assessment metrics for AlphaFold2 models, template-based homology models generated by MODELLER, and experimentally solved crystal structures of titin domains. Both the MolProbity score (a, b) and the Rama-Z score (c, d) were produced for for all models in each category (a, c) and for the "best" assessed models in each category (b, d). For MolProbity score, there is a visible improvement from homology models to AlphaFold2 models, and even average AlphaFold2 models tend to perform better than the best homology models; the assessments are comparable to and even exceed the assessments for solved crystal structures. The Rama-Z score shows far less difference between all structures, with only the best AlphaFold2 models having a noticeably lower abs(Z) than the other categories.

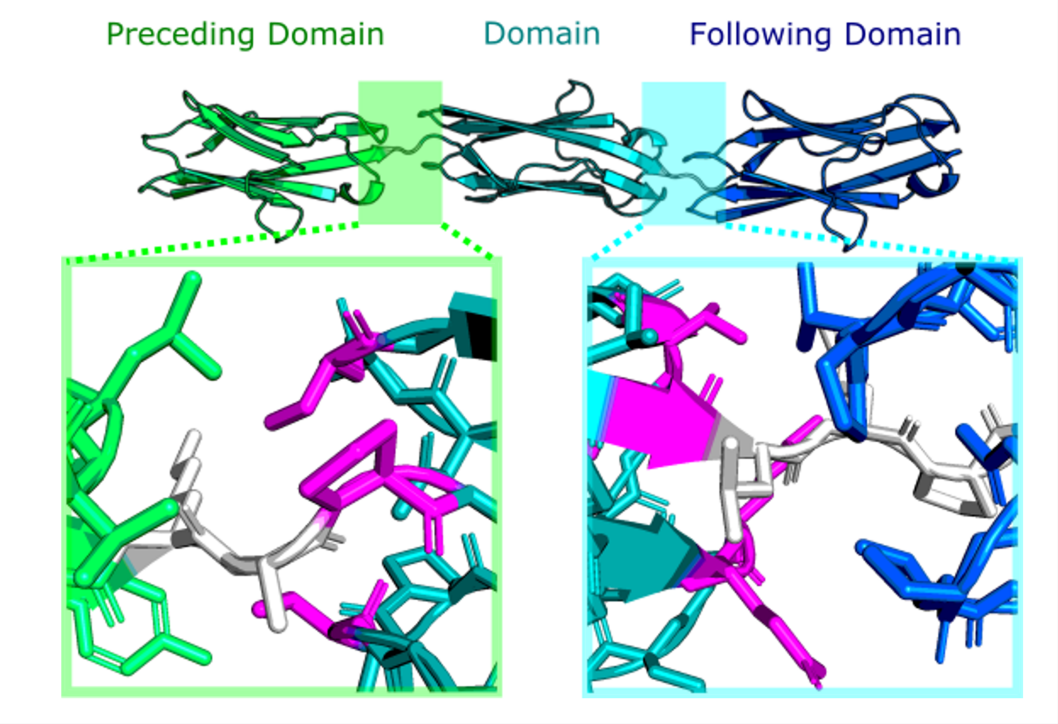


Figure 14: Structural schematic of domain-domain interfaces in titin

A visualisation of the arrangement of domain-domain interfaces in tightly connected tandem repeat regions of titin. A domain of interest (cyan) is preceded by a "Preceding" domain (green) and followed by a "Following" domain (blue). Residue positions in the domain of interest that are solvent-accessible in the single domain model, but buried when the neighbouring domains are considered, are considered interface-buried (magenta).

2.5.2 Interface Predictions from AlphaFold2 tandems

The predicted interface values were calculated across all available AlphaFold2 models for which a prediction of the tandem model was also available. For sake of simplicity these were split into "preceding" and "following" models (see Figure 14. In the "preceding" model, the single titin domain was compared to the two-domain tandem model in which it represented the N-terminal domain, and in the following model, where it was the C-terminal domain. This allows for the isolation of the residues that were typically involved in the interfaces between the domain of interest and both the previous domain in the titin chain and the one after. As each individual titin domain is represented in up to 7 different predicted AlphaFold2 constructs, I implemented a consensus approach to determining whether a given residue is

annotated as interface-buried: if the residue was classified as interface-buried in more than half of the available contigs ($Prop > 0.5$), it would be annotated as an interface residue. Encouragingly, most interface-buried residues were consistently classified as such across the different AlphaFold2 contigs; of the 2206 residue positions where the residue was classified as interface-buried at least once, the interface classification was the majority classification at 2103 positions (94.8%). The remaining positions, where interface-burial was either the minority or accounted for half of the classifications, were marked as non-interface. Thus, of the 28350 residue positions in titin within tandem domain boundaries, 2103 positions are considered interface-buried, or approximately 7.4% of these positions.

As seen in Figure 15a-b, the residue positions annotated as interface-buried were consistent across both the Ig and Fn3 domains. In Ig domains, residues at positions 11-15 (corresponding to the A'-B beta turn), positions 64-66 (corresponding to the E-F beta turn) and positions 90-92 (corresponding to the domain C-terminus) were typically calculated as buried in the interface with the following domain in the titin chain, while residues at positions 1-2 (corresponding to the domain N-terminus), positions 23-28 (corresponding to the B-C beta turn) and positions 77-80 (corresponding to the F-G beta turn) were calculated as buried in the interface with the prior domain. In Fn3 domains, residues at position 1 (corresponding to the domain N-terminus), positions 27-31 (corresponding to the long loop between beta strands B and C) and positions 80-83 (corresponding to the long loop between beta strands F and G) are buried in the interface with the prior domain; interestingly, and unlike Ig domains, only a minority of Fn3 domains have calculated interface-buried regions in these loop regions with the following domain, although residues in positions 96-100 (corresponding to the domain C-terminus) are consistently calculated as interface-buried.

As the Ig and Fn3 folds in titin are highly similar, the consistency in locations of interface-buried residues is to be expected; the construction of domain-domain tandems is likewise similar between modules. The differences in the exact position of interface-buried residues, and the presence in some domains but not in others, can mostly be attributed to the differing length and aligned positions of these residues in different domains, and the differing lengths

of linker sequences that connect domains. However, I wish to draw attention to two potential sources of confusion in these annotations. The first is that, in a minority of Ig domains, there are calculated interface-buried regions outside of the locations previously mentioned; these reflect the fact that certain tandem domains, particularly those that act as "hinges" in the I-band region, are predicted by AlphaFold2 in highly compact, bent conformations, which results in interfaces along the length of the two domains. It is unclear from experimental data whether such interfaces are physiologically present *in vivo*; as such, these interfaces have been retained in TITINdb, but I caution users against using AlphaFold2 predicted models as reliable evidence of biological interfaces.

The second note relates to the N- and C-terminal interface residues. In the conception of the interface-buried residue annotation, the intention is to highlight residues that are in contact with the domain-domain interface itself; however, with the definitions I have used above, the terminal residues can be predicted buried owing to their contact with the linker. Thus, I delineated two separate categories for interface-buried residues; those that were buried in the domain-domain interface were labelled as "interface", whereas those that were adjacent to the linker regions were given a separate class, "boundary". To automatically assign these classes, any residue labelled as "interface" in the proximal or terminal residues of a given domain is instead labelled as "boundary". In the graphical view of TITINdb2, "interface" residues are labelled in magenta and "boundary" residues in pink (see Figure 18).

2.5.3 Properties of Characterised SAVs

Saturation mutagenesis of SAVs in titin allows for both the annotation of variants in population and disease-associated databases, and the comparison between them. They have become increasingly popular in experimental characterisation of variants[188], and computational methods may also be used for *in silico* saturation mutagenesis. In Figure 16, the distributions of predictions for the three methods used for saturation mutagenesis of SAVs, Rhapsody, mCSM and PolyPhen2, and for the saturation mutagenesis of nsSNVs, Condel, are compared

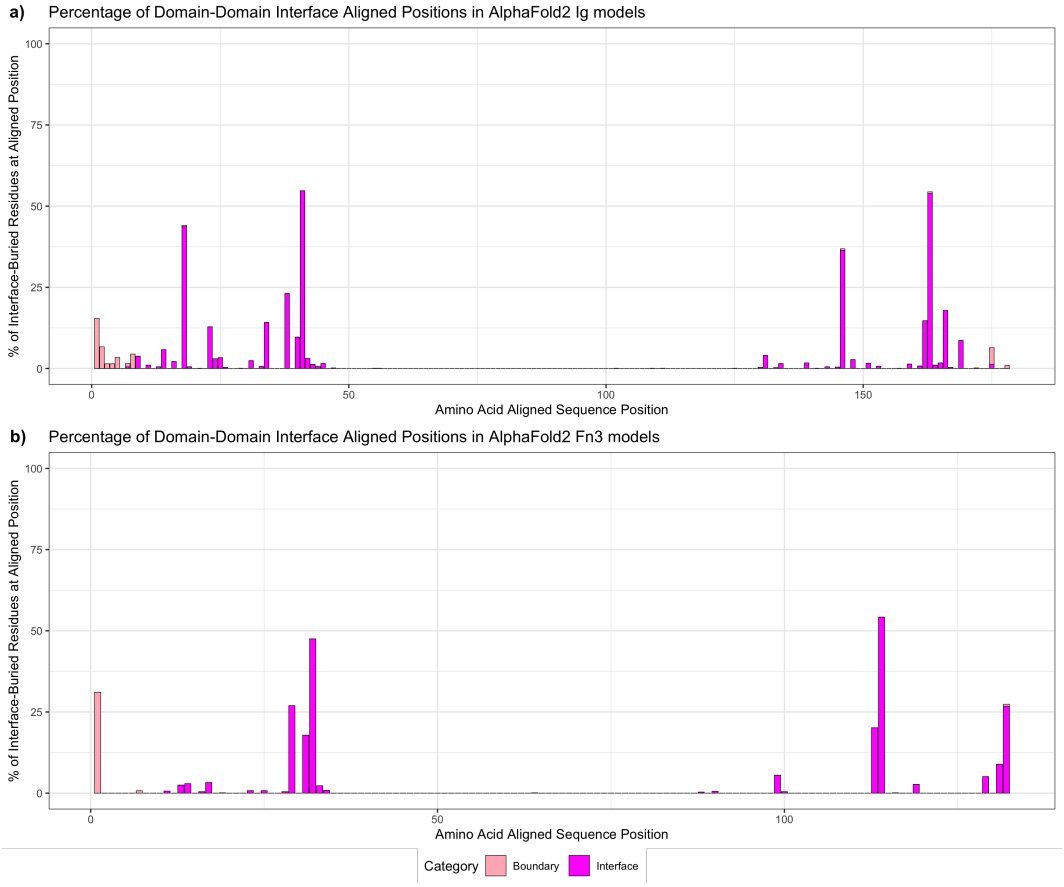


Figure 15: Positions of interface-buried residues

The positions of residues most frequently annotated as interface-buried when domain sequences are aligned, both for a) Ig domains and b) Fn3 domains. The frequencies are expressed as the percentage of residues in that aligned residue position, across all AlphaFold2 models of titin domains for that respective fold, which are annotated as interface-buried (magenta; aka "Interface") or linker-buried (light pink; aka "Boundary"). The regions of highest frequency represent the interface-adjacent regions in the two domain folds.

against the four databases for which SNVs have been sourced: 1000 Genomes, gnomAD-2.1.1, gnomAD-3.1.2, and the reported pathogenic variants in the scientific literature. The population databases are expected to have an even distribution of reported scores, while the literature variants should be enriched for potential pathogenicity or destabilisation.

As can be seen in the plots in Figure 16a-d, the distributions mostly follow the expected patterns. Rhapsody, the pathogenicity prediction algorithm, scores variants on a scale from 0 to 1, where 1 represents the greatest likelihood of pathogenicity and 0 the least, with a label of "deleterious" given where the score exceeds 0.55, or "neutral" where the score is under 0.45. The median values for the population databases are close to 0.5, as expected, and the median for the reported pathogenic literature variants is significantly higher than all of them. This suggests that the reported variants are indeed enriched for pathogenic variants where the population databases are not. Likewise, the protein stability prediction tool mCSM, which predicts a protein stability change $\Delta\Delta G$ in $kcalmol^{-1}$ for a given variant in a structure, does not predict a significantly different median $\Delta\Delta G$ between the population databases, but does predict a significantly lower median $\Delta\Delta G$ for literature variants.

Like Rhapsody, the pathogenicity prediction algorithms Condel and PolyPhen-2 predict a score between 0 and 1, where 1 represents the greatest likelihood of pathogenicity and 0 the least, and both also show the literature variants to be significantly different to the population variants in distribution. However, the distributions themselves invite comment due to the differences between them, and to that of Rhapsody. The plot of Condel values shows a very small interquartile range for each of the population databases, indicating that many variants in these databases are being predicted very close to 0.5, where the deleteriousness cut-off value is typically taken to be 0.52. Thus, Condel tends to predict scores within a very narrow margin around the prediction cut-off, which may undermine the annotation of variants. For PolyPhen2, the phenomenon is reversed; an extremely wide interquartile range demonstrates how the values are most often found at the extremes, giving very clear binary predictions. The median value for population variants is very high, around the 0.95 cut-off for PolyPhen2. These observations should be taken into consideration when sourcing

pathogenicity predictions for novel variants, as well as the potential risks in relying on a single pathogenicity prediction tool.

As can be seen in Table 1, many of the variants reported pathogenic in the literature also appear in population genetic databases such as gnomAD-3.1.2. This is to be expected; however, in some cases the population minor allele frequency (MAF) approaches or even exceeds 1%, which would class them as common variants in most definitions, and provide strong evidence against their role in any pathology. Most notably, variant H10092Y has a MAF of over 0.1% in both gnomAD-3.1.2 as a whole and in almost all ethnic sub-populations, notably including Finnish individuals, in which it surpasses 1%. Indeed, Rhapsody, PolyPhen2 and Condel all agree on the neutrality of this variant, in stark contrast to the other variants reported pathogenic. That said, H10092Y was reported with incomplete penetrance and only has a minor association with arrhythmogenic right ventricular cardiomyopathy, so this result is to be expected. The variant M35859T is the only other variant with a sub-population MAF over 1%, but this variant does not fall within a domain and so only Condel score is available, which is predicted neutral. Other variants with high sub-population MAF are also only reported based on association; a notable exception is L18237P, which is both present in non-Finnish Europeans with a MAF over 0.1%, but also predicted, like H10029Y, below 0.5 in all available pathogenicity scores. This is of interest as this variant was found in a cohort of congenital myopathy patients and has been demonstrated to decrease the melting temperature of its parent domain by around $17°C$ compared to the wild-type domain.

2.6 Case Studies and Applications

2.6.1 Congenital myopathy-associated variants in the A-band

A recent study by Rees *et al.* (2021)[35] demonstrates the use of TITINdb in characterising novel missense variants identified in patient cohorts and is summarised here with permission from the original authors. A recruitment of patients with both congenital myopathies and

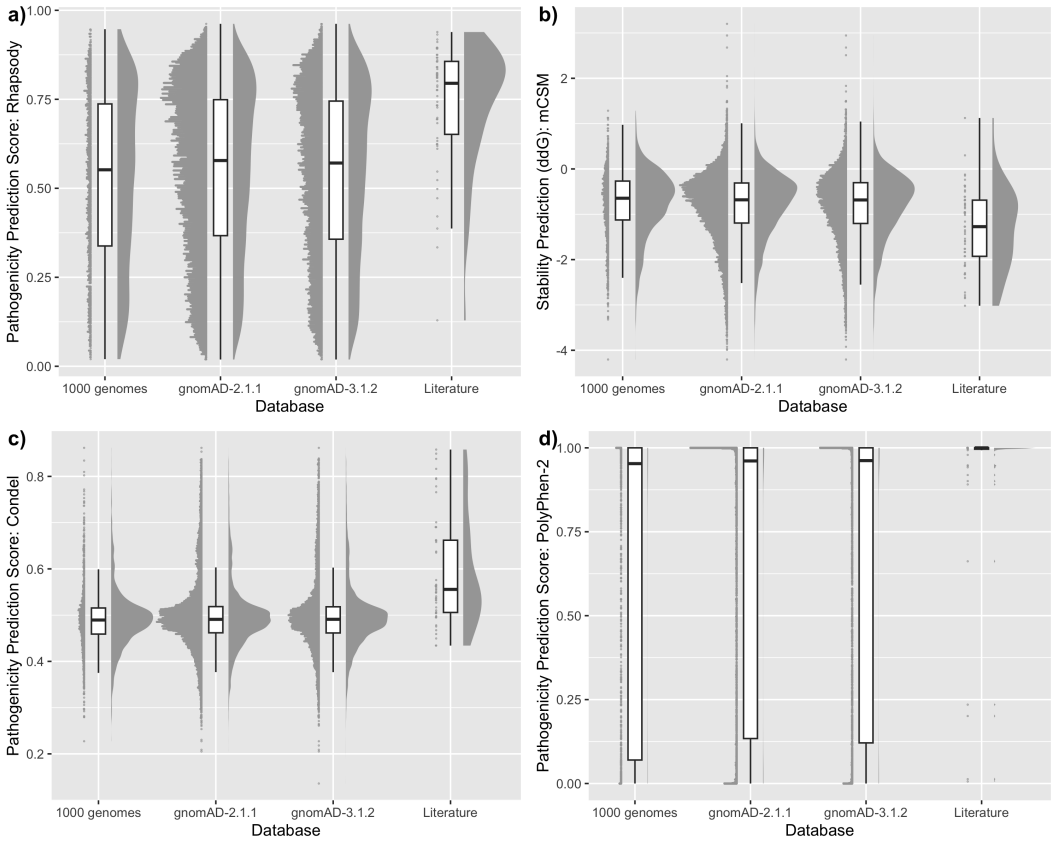


Figure 16: Distributions of bioinformatics annotations across titin SAVs in different database populations

Plots of annotation values for SAVs sourced from different original distributions. Three distributions are from population databases - 1000 Genomes, gnomAD-2.1.1, and gnomAD-3.1.2. The fourth distribution is SAVs reported as pathogenic in the literature. Annotations for four variant impact predictors are shown: a) Rhapsody, b) mCSM, c) Condel, d) PolyPhen-2. Three representations of the distribution of annotated values are shown: a bubble plot (left), indicating absolute quantities of variants from different distributions in discrete bins; a box plot (centre), indicating the median and inter-quartile ranges of values; a half-violin plot (right) indicating the comparative distributions of variants from different sources.

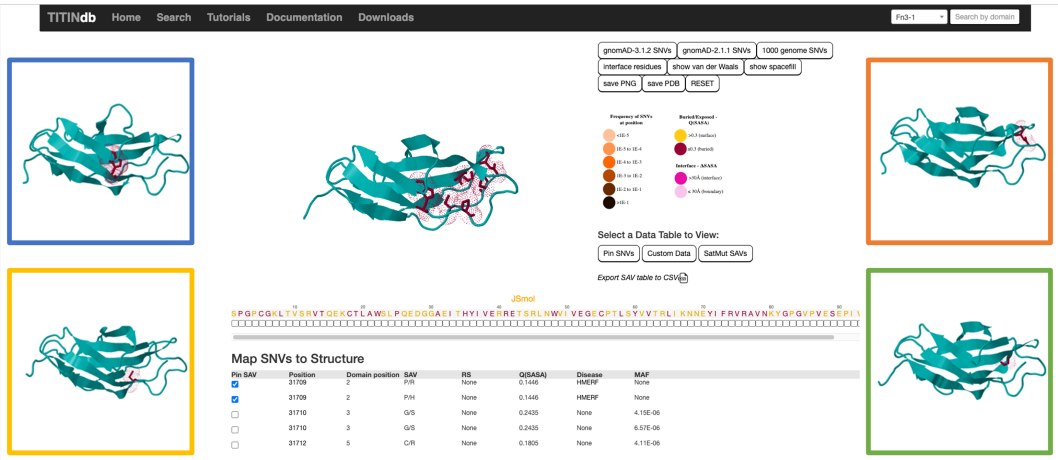


Figure 17: Locations of damaging variants in Fn3-119 and homologous domains

A screen capture of TITINdb2 illustrating reported pathogenic variants in Fn3-119 associated with HMERF, and inset, molecule viewer captures of homologous Fn3 domains in titin with reported pathogenic variants in the same structural positions: W16471C (Fn3-7, blue) at the same position as W31729C (Fn3-119); C17051R (Fn3-12, yellow) at the same position as C31712R (Fn3-119); N16133K (Fn3-4, orange) at the same position as N31786K (Fn3-119); G27849V (Fn3-90, green) at the same position as G31791V (Fn3-119).

at least two rare missense or truncating variants in titin were described for their clinical, histopathological, and genetic features. The missense variants found were identified and mapped to the parent domains using TITINdb; the structural homology models provided were downloaded and used to visualise the location of the missense variants, while the values for Q(SASA), mCSM and Condel predictions, and gnomAD minor allele frequencies were used to annotate each of the variants.

The concept of the "hotspot", as it relates to titin, refers to a position or positions in a molecule that are highly intolerant to mutation – typically, these can be identified when many variants with damaging consequences are found to localise to these common positions. As titin is a large, modular protein, the identification of hotspots is somewhat less intuitive – variants in different regions of the protein may have greatly different effects, potentially realised only in specific tissues. Thus, it is useful to compare newly identified, putatively pathogenic variants against known damaging variants, to determine if they correspond to a domain or structural location with known disease associations. Hotspots in titin can be categorised in two ways: firstly, hotspot domains, where many damaging variants within a

single domain or closely situated domains result in a similar phenotype; secondly, hotspot positions, wherein many damaging variants may be located at or around a similar structural location in different domains with the same fold.

It is well-established in the literature that variants in the Fn3-119 domain of titin are strongly associated with hereditary myopathy with early respiratory failure (HMERF)[78], with 13 reported missense variants in the literature alongside the condition. Many of these variants occur at the same location in the domain; for example, C31712R and C31712Y both at position 5 in Fn3-119. This suggests that within the fibronectin domain fold are specific residues integral to the stability of the domain and poorly tolerant of mutations; this observation was extended by Rees *et al.* with the identification of disease-associated variants in other domains that closely mirrored those already identified in Fn3-119. To elaborate, four A-band variants identified in patients with multi-minicore disease (MmD) represented the exact same residue change, at the exact same structural location, as variants reported in HMERF patients in Fn3-119: C17051R in Fn3-12 mirrors C31712R in Fn3-119; W16471C in Fn3-7 mirrors W31729C in Fn3-119; N16133K in Fn3-4 mirrors N31786K in Fn3-119; G27849V in Fn3-90 mirrors G31791V in Fn3-119 (see Figure 17). Intriguingly, this suggests that not only are these integral positions conserved in other titin Fn3 domains, with the same residue changes having a commensurate effect, but also that the destabilisation of different titin domains in different regions of the A-band can result in different pathologies.

A similar case was observed in Ig domains, where the variant A82D in Ig-1, which is associated with congenital fibre type disproportion (CFTD), is found at the same structural location as the well-characterised variant A178D in Ig-2. A178D was first identified in a single 3-generation family with dilated cardiomyopathy (DCM) with features of left ventricular non-compaction (LVNC), where it co-segregates with the condition (of 14 individuals tested, all 8 individuals who carried the A178D variant were affected)1. The variant was shown to destabilise the Ig-2 domain *in vitro*, and caused a reduction in the binding affinity of the titin N-terminus for TCAP. A subsequent study found that, when the same variant was expressed in knock-in mouse models, a DCM phenotype was unmasked when the mice were

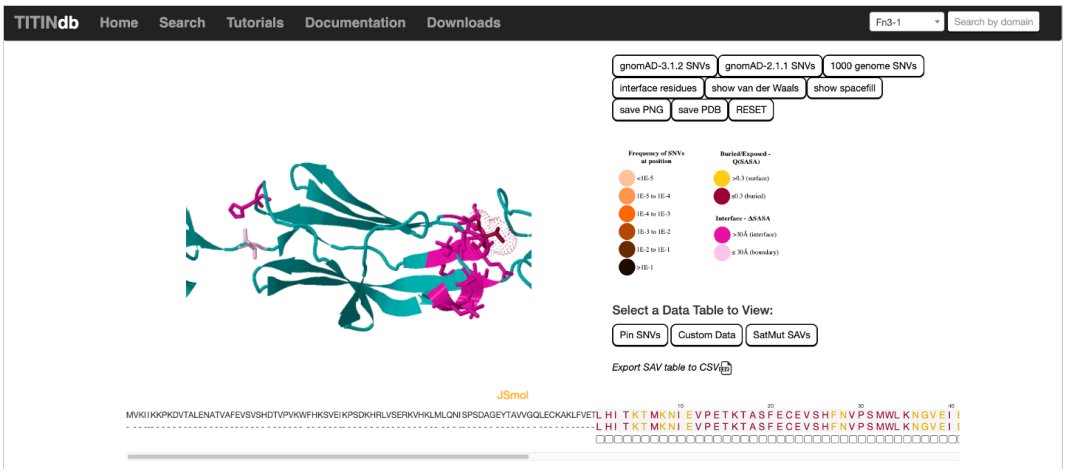


Figure 18: Locations of predicted domain-domain interface residues in Ig-19

Screen capture of the Domain-level page for Ig-19, with domain-domain interface residues toggled on. Here, those residues predicted by AlphaFold2 structural representations to be buried by contact with a domain external to the parent domain are labelled as "interface-buried" (magenta), while those buried by contact with linker residues are labelled as "boundary" (light pink). These may be viewed for all domains connected to other domains in titin by short linkers.

placed under challenge conditions, which the authors suggest was due to the proteotoxic effect of unbound, unfolded TCAP remaining in the cytoplasm[71]. In the same way, the A82D variant identified in Rees *et al.* also destabilised the parent domain *in vitro* and the mutant domain was insoluble; however, the A82D variant is associated with CFTD rather than DCM, as with A178D. It is notable that A82D was found to be compound heterozygous with the truncating variant R34807Sfs*7, and the patient in question showed no evidence of cardiomyopathy when the study was conducted.

2.6.2 A Disease-Associated, Interface-buried Residue in the I-band

The variant T2896I in the domain Ig-19 is reported damaging, associated with arrhythmogenic cardiomyopathy (ACM), on the basis of genetic, computational, and functional studies[98, 100, 113]. This variant localises to the A'-B beta turn (the loop that connects the A' and B beta strands in the immunoglobulin fold); interestingly, several other reported titin variants in the literature localise to this same beta turn (G4714D in Ig-30; S13702P in Ig-87; A21877S

in Ig-125). Thus, this structural position may be of interest to determine the extent of its role in the pathomechanisms of these variants.

By navigating to the specific page for the domain Ig-19, the structure or model upon which to visualise this variant may be selected (see Figure 18; structure used is the X-ray crystal structure 5JDD). By selecting the box at position 15 in the domain sequence, the SNV table may be filtered to bring up those variants found at this position. Then, selecting the variant T1896I will pin the variant onto the structure in the molecule viewer - in other words, highlighting it with a "licorice" structural representation - as well as accessing the annotations available through the SNV table. These include the scores from pathogenicity and stability prediction tools, links to the original literature, and population data, if available. Note that if the variant exists in multiple databases (for example, if it were present in both the literature and as a population variant in 1KGP), it would appear as multiple entries in the SNV table; the allele frequencies and other database-specific information are taken from the corresponding source.

An interesting observation to make of this variant is that it is interface-buried; its solvent-accessible surface area is reduced when flanking domains are considered, versus Ig-19 alone. Thus, while it would be considered an exposed residue when mapped to only the parent domain, mapping to the tandem illustrates it as buried. This concurs with computational studies suggesting that the variant induces an effect on the conformational dynamics of the titin chain[100]. Thus, this variant has been used as a model for the effect of functional impacts beyond the focus of a single domain, specifically with reference to the titin chain context. An example is that of Fleming *et al.* (2021)[114], who more closely investigated the domain interface itself as a potential hotspot for damaging variation.

The authors investigated the impact of T2896I in the Ig19-Ig20 tandem using molecular dynamics (MD) simulations of both the wild-type system and different missense mutations at the affected residue position. Curiously, although the THR to ILE substitution showed the greatest deviation from the wild-type when principal component analysis (PCA) was

conducted, all other variants at the same position showed a similar transition from the straight wild-type conformation to a more twisted conformation, or alternation between straight and twisted conformations, with the exceptions of T2896V and T2896W. The authors proposed a Domain Interface Score (DIS) based on computational characterisation of stability change for all possible variants in tandem titin domains. Similarly, the characterisation of interface residues by solvent accessibility, which covers all titin tandem domains rather than Ig19-Ig20 only, will allow for reported pathogenic variants in these regions to be easily identified as interface-buried, and thus for pathogenicity related to titin chain context to be more readily explored.

2.6.3 Prioritisation of Variants from High-Throughput Pipelines

Next-generation sequencing (NGS) and other high-throughput approaches generate huge quantities of data, including the genetic data from patients with inherited diseases. In the case of titin, not only have variants in titin been associated with a wide range of muscle diseases, but the sheer size of titin also means that all individuals will contain some missense variants of varying significance within their titin sequence[69]. By extension, this also means that healthy individuals will have missense variants, often those which would be considered rare in the general population. This complicates the prioritisation of variants, where the great quantity of variants identified in a patient population must be filtered to produce a candidate list of a size amenable to further experimental validation.

Bioinformatics annotation pipelines such as ANNOVAR[88], VEP and SnpEFF are often used to identify variants of interest from within large gene sequencing data sets, by annotating the individual variants with bioinformatics tools and then filtering by specific criteria. TITINdb2 can perform a similar function for titin missense variants, but with certain advantages over these more general approaches. Consider a data set of titin variants taken from a cohort of inherited myopathy patients; one would be interested in determining whether any variants of interest, with the possibility of association with the disease state, are present

that could be studied further. By filtering the data set to remove all non-coding variants, and then removing all but non-synonymous SNVs, a list of missense variants is obtained that can be queried against TITINdb2.

TITINdb, like other annotation pipelines, provides minor allele frequency (MAF) values for the population variant databases, as well as those of reported sub-populations; this allows for variants already reported in healthy individuals, or beyond a certain MAF threshold, to be filtered out. Where TITINdb simplifies the process is that it provides the necessary population and computational information in a single location, without the need for separate software or converting between variant formats. For example, the availability of pre-computed predictions for all possible variants by mCSM, Rhapsody, and PolyPhen2 allows for assessment of the likelihood for a given variant to be damaging, without running these programs for each variant of interest, even where these variants have never been reported in SNV databases – which is particularly important for NGS data analysis where novel variants may be detected.

In addition to the availability of variant effect prediction scores, variants of interest can be mapped to domain structures, even if no solved experimental structure exists; this allows for the Q(SASA) of the residue at that position to be predicted from the model, giving an idea of the solvent accessibility associated with the mutational change. In addition, if a single variant is pinned to a structure within TITINdb, the SNV table will automatically update to bring related or identical SNVs (where the same variant appears in multiple sources) to the top of the table. These variants will also be annotated with predictive tools and coloured according to Q(SASA).

2.6.4 Investigating the Mutational Landscape of Titin with Saturation Mutagenesis

Rhapsody[176] is a pathogenicity prediction tool that aims to improve on long-standing and well-established bioinformatic tools by incorporating more sources of data pertinent to the impact of the variant of interest. Sequence conservation features are the most commonly

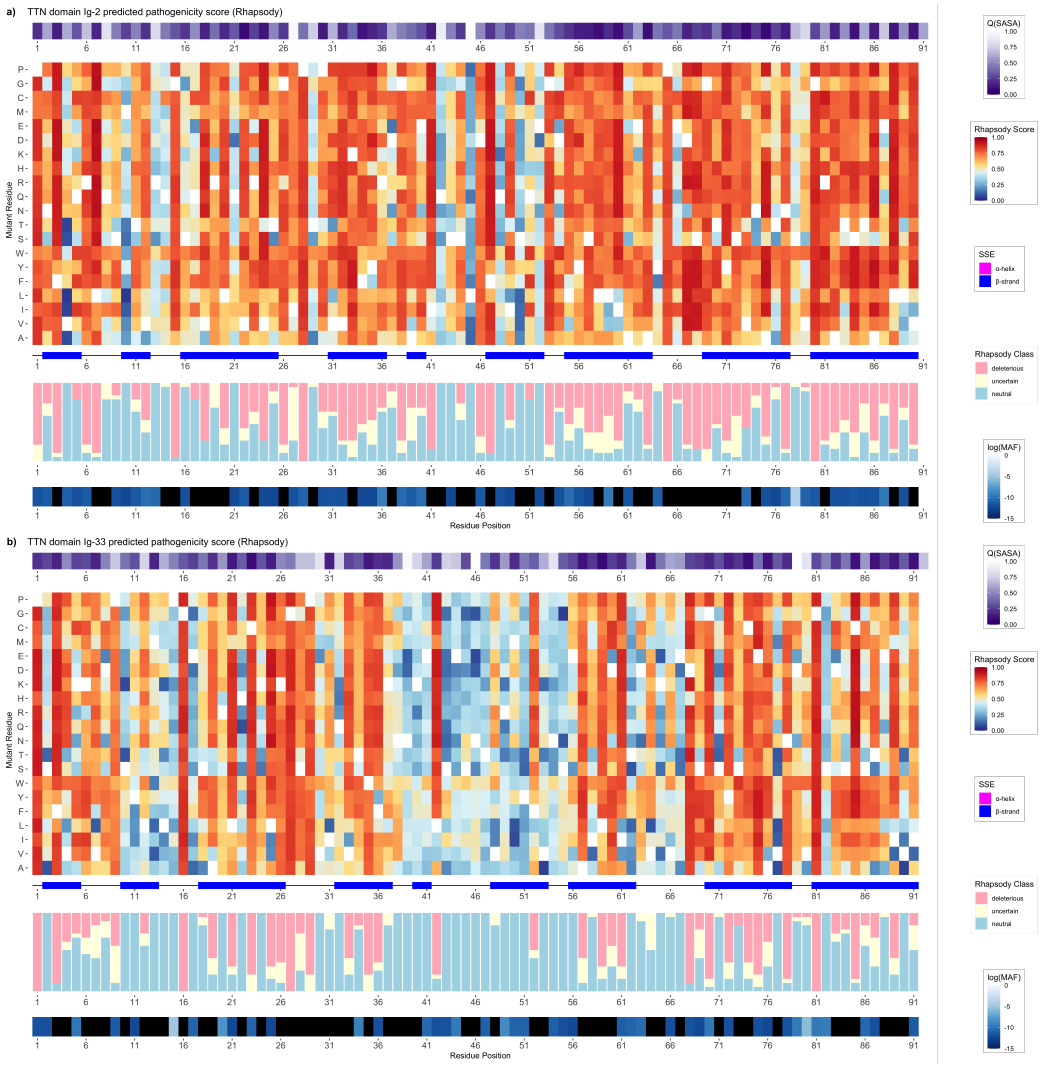


Figure 19: Saturation mutagenesis heat map of Rhapsody predictions for two disease-associated Ig domains

Prediction maps of in silico saturation mutagenesis pathogenicity predictions computed by Rhapsody for a) Ig-2 (Z-disc) and b) Ig-33 (I-band). In the main plot, the Rhapsody predictions for all possible single amino acid substitutions at all possible positions in the domain are illustrated by a single square, where the colour indicates the value. Rhapsody predicts a value from 0 (definite neutral) to 1 (definite damaging); in this plot, red-coloured values represent predictions between 0.5 and 1, and blue-coloured values represent predictions between 0 and 0.5. Wild-type residues are indicated by a white square. Above, the plots are annotated with the Q(SASA) of the wild-type residue (darker purple indicates more buried residues) and below, the predicted secondary structure elements (SSE) at that position (magenta for α-helix, blue for β-strand). Two supplementary plots below aid visualisation; the first gives a count of how many SAVs at that residue position are predicted to fall into each Rhapsody class: damaging (red) for Rhapsody > 0.74; uncertain (yellow) for values between 0.74 and 0.675; neutral (blue) for Rhapsody < 0.675. The second gives the log-transformed MAF of variants in gnomAD at that residue position; darker blue indicates that variants are rarer at that position, while black indicates that no variation is reported at that position.

used data to train variant impact predictors[145, 184], with structural features also included where information on protein structure is available[144]. Rhapsody incorporates not only these features, but also dynamics-based features calculated from network representations of protein structures; moreover, the speed of computation is such that the vast number of possible variants in folded titin domains may be annotated with a single-value predicted pathogenicity score in comparatively little time.

With this in mind, I investigated the patterns of Rhapsody predictions for all possible residue substitutions at all positions in the 167 Ig domains in the canonical N2AB titin, a total of 281,713 individual missense mutations. These were collected into heat maps displaying the distribution of damaging vs neutral predicted variants at each position in the domain. In addition, these plots were annotated with summary plots indicating the number of missense variants of each Rhapsody class at each position, the solvent accessibility at that residue position in the wild-type structure, the secondary structure elements at that position[189], and the MAF of population variants in gnomAD at that position (see Figure 19). This allowed for a representation of the positions in the domains most susceptible to mutation, as well as the residue substitutions predicted to be the most damaging when compared to others. Typically, residues most susceptible to damaging mutations were those that were most buried in the domain (low Q(SASA); darker colour in heat map) while those solvent-accessible residues, especially those outside of beta-strand secondary structure, tended to be more tolerant of most mutations.

These results tally with expectations based on the training of Rhapsody on these and similar features; however, despite the great similarity of these domains in their sequence, fold, and secondary structure, there were marked differences in some of the patterns of prediction between different Ig domains in different regions of titin, even where both domains were associated with disease. For example, the heat map for Ig-2 (Figure 19a) appears visually to have a preponderance of red-coloured "predicted pathogenic" across the whole of the protein, with most exceptions in unstructured loop regions; by contrast, the heat map for Ig-33 (Figure 19b) shows more positions dominated by blue-coloured "predicted neutral",

with the exception of the most buried core regions. Thus, while both of these domains have predicted key residues at the same structural positions intolerant of variation, and both have reported damaging SAVs (A178D in Ig-2, C5054R in Ig-33) , Ig-33 appears more tolerant of variation in all other areas, despite the 23% sequence identity between the two, and the identical I-set immunoglobulin domain fold.

To investigate this difference, all Ig domains in titin were aligned and their predictions combined into a single heat map, with the average variant impact (Rhapsody score) at each position across all 19 possible substitutions used as that position's tolerance to mutation (Figure 20). The 167 Ig domain sequences were aligned using T-COFFEE[190], and the aligned positions used to plot the heat map. The large insert sequence in the central loop region of Ig-112 was removed from the plot to improve visibility. The PROMALS3D[191] tool was used to align the structural positions of all experimentally solved X-ray crystal structures of titin Ig domains in the PDB, which were then used to derive an agreement for aligned positions in secondary structure, which was used to annotate the plot. This allows for the overall fold of the protein to be described in the aligned sequences.

In this overall plot, several points of interest may be observed. While there are differences in the distribution of predicted pathogenicity between Ig domains, particularly those in different regions of titin (red = Z-disc, blue = I-band, green = A-band, purple = M-band). However, the aligned positions most intolerant of mutation are remarkably consistent across the whole of titin, typically occurring within regions of secondary structure and alternating with more tolerant positions, reflecting the alternation of inward-facing core residues and outward-facing surface residues in the structure. This indicates that the multiple sequence alignment produced by T-COFFEE is reliable. One noticeable, visible difference is in the distribution of predictions in the centre of the I-band Ig domains, which follow the patterning previously described for Ig-33.

Interestingly, this region corresponds exactly to the Ig31-Ig82 region of titin's I-band; this region contributes to the extensibility of the titin protein *in vivo*, and most intriguingly,

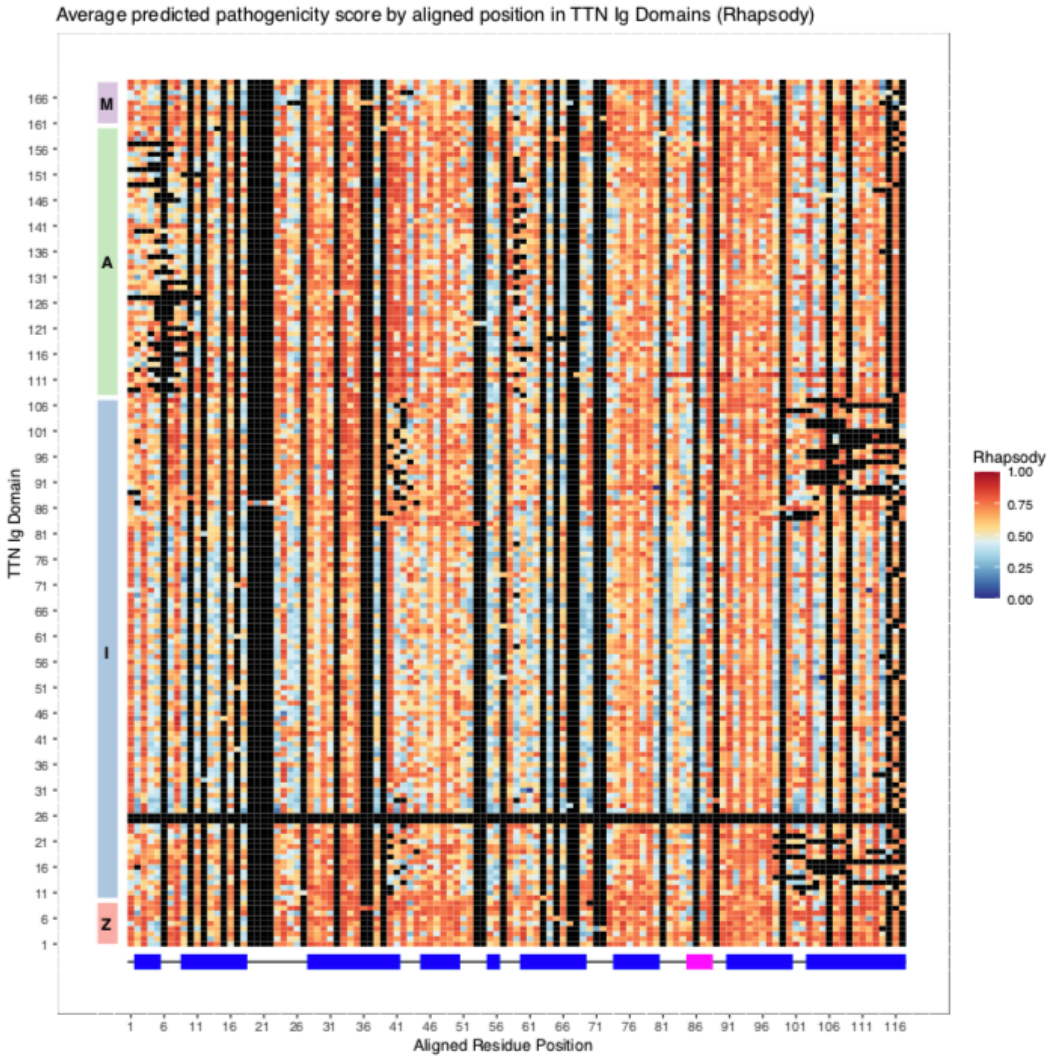


Figure 20: Aligned saturation mutagenesis heat map of average per-position Rhapsody predictions for all Ig domains in titin.

Combined saturation mutagenesis heat map of all predictions for Ig domains in titin at aligned positions. For each domain, protein sequences were aligned and residue positions were annotated with an aligned sequence position. A Rhapsody score for that position in that domain was then calculated as the average of all 19 substitution scores at that position; all Ig domains were then combined together to produce the final heat map. As before, Rhapsody predicts a value from 0 (definite neutral) to 1 (definite damaging); red-coloured values represent predictions between 0.5 and 1, and blue-coloured values represent predictions between 0 and 0.5. Absent predictions in the heat map, owing to gaps in the alignment, are depicted as black; the exception is the large horizontal gap representing Ig-25 and Ig-26. These domains are not in the canonical titin sequence and thus a prediction for Rhapsody could not be computed. The main plot is further annotated with aligned secondary structure elements (magenta for α-helix, blue for β-strand) and the location in the titin chain of that domain (red = Z-disc, blue = I-band, green = A-band, purple = M-band).

is spliced out from the major cardiac N2A isoform. Thus, the possibility arises that these domains are more tolerant to non-core mutations as, given their role in the elastic component of titin, these domains naturally possess a greater flexibility than Ig domains in other regions. While this hypothesis cannot be tested by Rhapsody, it is conceivable that future efforts to parse the pathogenicity of titin variants should take into account the position of the domain in the overall ultrastructure of titin, and re-emphasises that the same variation applied to the same structural position in two domains in different regions of titin will not necessarily have commensurate variant impacts.

2.7 Conclusion

Despite the importance of computational tools to interrogating the large genomic data sets produced from current and future disease populations, there remain the twin hurdles of accessing and then using the wealth of previous data and tools available for end users. I have demonstrated the continued application of TITINdb2 to the problem of wide-scale assessment of variant impact of missense variants in the giant protein titin, by providing a centralised resource with both sequence and structural data, a wide variety of annotations from different computational tools, and native visualisation tools. TITINdb2 also illustrates the current gaps in our collective knowledge, and the limitations of current bioinformatics approaches to consistently classify given variants; questions raised by the patterns of predictions in case studies and in *in silico* saturation mutagenesis approaches present questions that should form the foundation of future studies into disentangling the enigmas surrounding these small but deeply consequential protein variants.

Table 3: Experimentally solved Titin Domain structures in Protein Data Bank

PDB	Reported Domain	UniProt Domain	TITINdb2 Domain	Method	Resolution
2F8V	Titin-Nterm	Ig-like 1	Ig-1	X-Ray Diffraction	2.75A
		Ig-like 2	Ig-2		
2A38	Titin-Nterm	Ig-like 1	Ig-1	X-Ray Diffraction	2.00A
		Ig-like 2	Ig-2		
6FWX	Titin-Z1	Ig-like 1	Ig-1	X-Ray Diffraction	3.00A
	Titin-Z2	Ig-like 2	Ig-2		
6SDB	Titin-Z1	Ig-like 1	Ig-1	X-Ray Diffraction	2.80A
	Titin-Z2	Ig-like 2	Ig-2		
1YA5	Titin-Z1	Ig-like 1	Ig-1	X-Ray Diffraction	2.44A
	Titin-Z2	Ig-like 2	Ig-2		
6DL4	Titin-Z10	Ig-like 9	Ig-9	Solution NMR	
1G1C	Titin-I1	Ig-like 10	Ig-10	X-Ray Diffraction	2.10A
4QEG	Titin-I10	Ig-like 16	Ig-19	X-Ray Diffraction	2.00A
5JDJ	Titin-I10	Ig-like 16	Ig-19	X-Ray Diffraction	1.74A
5JDD	Titin-I9	N/A	Ig-18	X-Ray Diffraction	1.53A
	Titin-I10	Ig-like 16	Ig-19		
	Titin-I11	Ig-like 17	Ig-20		
5JDE	Titin-I9	N/A	Ig-18	X-Ray Diffraction	1.90A
	Titin-I10	Ig-like 16	Ig-19		
	Titin-I11	Ig-like 17	Ig-20		
2RIK	Titin I67	Ig-like 64	Ig-70	X-Ray Diffraction	1.60A
	Titin I68	Ig-like 65	Ig-71		
	Titin I69	Ig-like 66	Ig-72		
3B43	Titin I65	Ig-like 62	Ig-68	X-Ray Diffraction	3.30A
	Titin I66	Ig-like 63	Ig-69		
	Titin I67	Ig-like 64	Ig-70		
	Titin I68	Ig-like 65	Ig-71		
	Titin I69	Ig-like 66	Ig-72		
	Titin I70	Ig-like 67	Ig-73		
5JOE	Titin I81	N/A	Ig-84	X-Ray Diffraction	2.00A

7AHS	Titin-N2A Ig81	Ig-like 77	Ig-84	X-Ray Diffraction	2.04A
	Titin-N2A Ig82	Ig-like 78	Ig-85		
	Titin-N2A Ig83	Ig-like 79	Ig-86		
6YJ0	Titin I83	Ig-like 79	Ig-86	Solution NMR	
1TIT	Titin-I27	N/A	Ig-94	Solution NMR	
1TIU	Titin-I27	N/A	Ig-94	Solution NMR	
1WAA	Titin-I27	N/A	Ig-94	X-Ray Diffraction	1.80A
2RQ8	Titin-I27*	N/A	Ig-94	Solution NMR	
6I0Y	Titin-I27†	N/A	Ig-94	Electron Microscopy	3.20A
4O00	Titin-A3	FnIII 3	Fn3-3	X-Ray Diffraction	1.85A
8BXR	Titin I109	N/A	Ig-109	X-Ray Diffraction	2.70A
	Titin I110	FnIII 4	Fn3-4		
	Titin I111	FnIII 5	Fn3-5		
8BW6	Titin I110	FnIII 4	Fn3-4	X-Ray Diffraction	1.95A
8BVO	Titin I110	FnIII 4	Fn3-4	X-Ray Diffraction	2.55A
	Titin I111	FnIII 5	Fn3-5		
1BPV	Titin-A71	FnIII 62	Fn3-62	Solution NMR	
3LPW	Titin-A77	FnIII 66	Fn3-66	X-Ray Diffraction	1.65A
	Titin-A78	FnIII 67	Fn3-67		
3LCY	Titin-A164	Ig-like 140	Ig-156	X-Ray Diffraction	2.50A
	Titin-A165	N/A	Ig-157		
2J8H	Titin-A168	Ig-like 141	Ig-158	X-Ray Diffraction	1.99A
	Titin-A169	Ig-like 142	Ig-159		
2J8O	Titin-A168	Ig-like 141	Ig-158	X-Ray Diffraction	2.49A
	Titin-A169	Ig-like 142	Ig-159		
2ILL	Titin-A168	Ig-like 141	Ig-158	X-Ray Diffraction	2.20A
	Titin-A169	Ig-like 142	Ig-159		
2NZI	Titin-A168	Ig-like 141	Ig-158	X-Ray Diffraction	2.90A
	Titin-A169	Ig-like 142	Ig-159		
	Titin-A170	FnIII 132	Fn3-132		
4JNW	Titin Kinase	Protein kinase	Kinase-1	X-Ray Diffraction	2.06A
1TKI	Titin Kinase	Protein kinase	Kinase-1	X-Ray Diffraction	2.00A
6YGN	Titin-A168	FnIII 132	Fn3-132	X-Ray Diffraction	2.40A
	Titin Kinase	Protein kinase	Kinase-1		

	Titin-M1	Ig-like 143	Ig-160		
2BK8	Titin-M1	Ig-like 143	Ig-160	X-Ray Diffraction	1.69A
6HCI	Titin-M3	Ig-like 145	Ig-162	X-Ray Diffraction	2.12A
6H4L	Titin-M4	Ig-like 146	Ig-163	X-Ray Diffraction	1.60A
3QP3	Titin-M4	Ig-like 146	Ig-163	X-Ray Diffraction	2.00A
1NCT	Titin-M5	Ig-like 147	Ig-164	Solution NMR	
1NCU	Titin-M5	Ig-like 147	Ig-164	Solution NMR	
1TNN	N/A	Ig-like 147	Ig-164	Solution NMR	
1TNM	N/A	Ig-like 147	Ig-164	Solution NMR	
3PUC	Titin-M7	Ig-like 149	Ig-166	X-Ray Diffraction	0.96A
4C4K	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	1.95A
4UOW	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	3.30A
3Q5O	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	2.05A
2Y9R	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	1.90A
2WP3	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	1.48A
2WWK	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	1.70A
2WWM	Titin-M10	Ig-like 152	Ig-169	X-Ray Diffraction	2.30A
3KNB	Titin-Cterm	Ig-like 152	Ig-169	X-Ray Diffraction	1.40A

** Structure* 2RQ8 *is a mutant domain Ig-94, which includes variant Y14326P for analysis of domain destabilisation (see Yagawa* et al. *2010[192]).* † *Structure* 6I0Y *is a large TnaC-stalled ribosome complex to which Ig-94 is bound; Ig-94 represents chain* z *(*6I0Y_36*) in the structure.*

Table 4: Pathogenicity and stability predictors implemented in TITINdb2

Predictor Name	Source	Description	Availability	Reference
CADD	dbNSFPv4	Pathogenicity meta-predictor (sequence conservation, various other predictors)	Selected nsSNVs	Rentzsch *et al.* (2019)[193]
Condel	FannsDB	Pathogenicity meta-predictor (FATHMM, MutationAssessor)	Selected nsSNVs	González-Pérez & López-Bigas (2011)[173]
DUET	Calculation with supplied structures (web server)	Protein stability meta-predictor (mCSM, SDM)	Selected nsSNVs	Pires *et al.* (2014a)[175]
FATHMM	dbNSFPv4	Pathogenicity predictor (sequence homology, disease weighting)	Selected nsSNVs	Shihab *et al.* (2013)[194]
M-CAP	dbNSFPv4	Pathogenicity predictor (sequence homology, genetic constraints)	Selected nsSNVs	Jagadeesh *et al.* (2016)[195]
mCSM	Calculation with supplied structures (web server)	Protein stability predictor (graph-based structural signatures)	Saturation mutagenesis of SAVs (all domains)	Pires *et al.* (2014b)[174]
MetaSVM	dbNSFPv4	Pathogenicity meta-predictor (11 other predictors, allele frequency)	Selected nsSNVs	Dong, C. *et al.* (2015)[196]

MetaLR	dbNSFPv4	Pathogenicity meta-predictor (11 other predictors, allele frequency)	Selected nsSNVs	Dong, C. *et al.* (2015)[196]
MPC	dbNSFPv4	Pathogenicity predictor (sequence homology, genetic constraints)	Selected nsSNVs	Samocha *et al.* (2017)[197]
MutationAssessor	dbNSFPv4	Pathogenicity predictor (sequence homology)	Selected nsSNVs	Reva *et al.* (2011)[198]
MutationTaster	dbNSFPv4	Pathogenicity predictor (sequence homology)	Selected nsSNVs	Schwarz *et al.* (2014)[199]
MutPred	dbNSFPv4	Pathogenicity predictor (sequence homology, conservation, structural change)	Selected nsSNVs	Pejaver *et al.* (2020)[200]
MVP	dbNSFPv4	Pathogenicity predictor (sequence conservation, structural features, other predictors)	Selected nsSNVs	Qi *et al.* (2021)[185]
PolyPhen-2	Retrieval from web server	Pathogenicity predictor (sequence homology, structural annotation)	Saturation mutagenesis of SAVs (N2AB domains)	Adzhubei *et al.* (2013)[144]
PrimateAI	dbNSFPv4	Pathogenicity predictor (direct learning from sequence alignments)	Selected nsSNVs	Sundaram *et al.* (2018)[201]

PROVEAN	dbNSFPv4	Pathogenicity predictor (sequence homology)	Selected nsSNVs	Choi *et al.* (2012)[202]
REVEL	dbNSFPv4	Pathogenicity meta-predictor (13 other predictors)	Selected nsSNVs	Ioannidis *et al.* (2016)[184]
Rhapsody	Calculation with supplied structures (local pipeline)	Pathogenicity predictor (sequence conservation, structural features, coarse-grained dynamics)	Saturation mutagenesis of SAVs (N2AB domains)	Ponzoni *et al.* (2020)[176]
SIFT	dbNSFPv4	Pathogenicity predictor (sequence conservation)	Selected nsSNVs	Ng & Henikoff (2003)[145]

3 Analysis of Titin Missense Variants using Atomistic Molecular Dynamics

3.1 Synopsis

- In this chapter, I characterise the domain-level impact of selected missense variants in titin using molecular dynamics simulations, and analyse the trajectories produced.
- I discuss the extraction of features from high-dimensional, complex data that suggest a consistent distinction between titin domain systems and their variations in different categories (e.g. wild-type vs. destabilising mutant).
- I demonstrate the potential for such dynamical features to contribute meaningful information absent from static representations of variants in sequence or structure, which can be used in future classification methodologies.

3.2 Introduction

In the previous chapter, I discussed the annotation of missense variants of unknown significance with computational methods. The strength of these methods is in the speed with which large numbers of variants can be assessed, in a fraction of the time necessary to experimentally characterise the same number of variants; unfortunately, their weakness is that scores from bioinformatics tools do not always recapitulate experimentally determined parameters or clinical outcomes. Thus, no single tool or combination of tools may be confidently used to validate associations between variants and phenotypes; methods and targets for improving the quality of computational predictions continue to be actively pursued.

In TITINdb2, I considered predictors from many different sources as annotations, each of which results from different training strategies and methodologies to derive meaningful

outputs. In all variant impact predictors, the central challenge is to construct an *in silico* representation of the amino acid substitution, such that from a single input an accurate prediction may be produced. To elaborate, many popular variant impact predictors focus on a representation in protein primary structure[145]; in the protein amino acid sequence alone, or the DNA sequence that encodes it. In all of these approaches, the conservation of particular residues at specific sequence locations between homologous proteins is well-established as a strong indicator of that position's tolerance to mutation, and by extension the likely severity of consequences to a substitution. With ever-increasing libraries of sequence data allowing for deeper sequence alignments, and thus more accurate measures of evolutionary conservation, these methods continue to improve; currently, top-performing methods also make use of powerful deep learning architectures (e.g. MVP[185], DeepDDG[203]) to learn more complex mappings, or meta-prediction approaches (e.g. REVEL[184], Condel[173]) to incorporate features from multiple single-method predictors.

While the association between evolutionary conservation and tolerance of mutation is strong, it is not absolute; the amino acid sequence alone does not include all information necessary to consistently partition variants by their experimental or clinical consequences. Consequently, other approaches that have gained popularity in recent years seek to approach the challenge from a different angle, where rather than extending the quantity of data used, the variant impact is predicted using more diverse data types. In addition to their amino acid sequence, proteins also have secondary and tertiary structure; the way in which a protein folds in solution informs which residues are in contact with one another, whether the residues are buried or exposed to the solvent, and thus what contribution these residues make to the stability of the protein itself. Such structural features in predictors are not new (e.g. PolyPhen2[144]), but have long been limited by the information gap between available sequence and structural data.

A recent explosion of structural data has been driven by the introduction of AlphaFold2[170], which has demonstrated the power of large data sets and deep learning architectures by predicting the tertiary structures of protein monomers to a degree of accuracy that eclipses both

traditional template-based homology modelling and previous *ab initio* modelling techniques; in addition, whole-proteome predictions for model species have been made available to the scientific community[171]. While this demonstrates the principle that information about secondary and tertiary structure is implicit in the protein sequence, it also highlights the sheer complexity of mapping from one to the other; thus, the ability to derive the tertiary structure directly from sequence enables far more complex structural features to be used to classify variants. However, while such representations better approximate the state of the protein *in vivo*, they are still limited; the three-dimensional structure of proteins is maintained by interactions between their constituent atoms, which are constantly shifting in response to the myriad forces exerted by their neighbours in three-dimensional space, which in turn affects these interactions' strength.

In order to properly model the consequences of an alternate amino acid, and hence atoms, at a given position, the effects of a vast number of simultaneous interactions, which are themselves changing over time, must be considered. In short, proteins are best represented as four-dimensional, dynamic entities, and the information contained in the time evolution of a protein system may be germane to the improvement of variant impact prediction. This line of inquiry has led to the development of variant impact predictors (e.g. Rhapsody[176], LYRUS[204]) that incorporate these different levels of information; Rhapsody, which I have already described in Chapter 2, was in many ways the pioneer of these approaches, and demonstrated the ability of a combined classifier to predict correctly those variants that neither predictors trained solely on sequence features nor on dynamical features were able to. Thus, it can be said that the combination of these different features augment the shortcomings of the individual features; I have demonstrated the use of Rhapsody previously to annotate the domains in titin's canonical sequence. However, Rhapsody itself does not perfectly recapitulate variant impact; in order to improve the accuracy of predictions for titin, a protein composed of small, highly homologous domains, I elected to conduct my own investigation into the dynamical properties of titin domains.

Dynamical features embed the time evolution of a system undergoing a simulation; the granularity of the system dictates the precision of the output, but also the computational costs involved. Depending on the system in question, these can range from vast, internally frozen models at the mesoscopic level, to minute systems of atomic electronic structure that follow quantum mechanics[205]. The predictors described previously that incorporate dynamical features use coarse-grained modelling; groups of atoms are represented as single particles to simplify computation[176, 206]. I opted instead to focus on molecular dynamics-based approaches, alternatively "atomistic" dynamics, where all atoms are represented as particles and their interactions follow classical Newtonian mechanics[207]. This representation was chosen as the relatively small size of the titin domains allows for characterisation by higher-granularity methods in a reasonable time frame, with a commensurate increase in the data generated. My aim was to determine the extent to which these methods, and the features derived from them, could accurately characterise variants in titin domains.

3.3 Methods

3.3.1 Literature and Population Variant Data Sets

The initial selection of systems for simulation was comprised of those variants reported as pathogenic in the scientific literature (see Chapter 1), as well as an equal quantity of variants with the highest allele frequency in the population genetic database gnomAD[84]; these were selected as the known variants least likely to be injurious to health. The variants in this distribution also included several variants in the hotspot domains Fn3-119 and Ig-169 which, while rare in gnomAD, had been experimentally characterised or reported as non-pathogenic, as comparison variants to disease-associated variants in those domains; the variants A16466T (Fn3-7), V22193I (Fn3-49) and R27839Q (Fn3-90) were included with the same rationale (see Table 5, gnomAD MAF). Each variant was labelled in a binary classification of "Literature" or "Population" depending on their database of origin. After removing variants that were not located in folded domains, the final data set comprised a total

of 31 literature variants and 30 population variants (Table 5). These 61 variants were found in 24 different titin domains; therefore, the wild-type systems were simulated alongside the variant systems, for a total of 85 simulated systems.

3.3.2 Thermostability of Variants Data Set

The subsequent data set consisted of all single-domain titin variant and wild-type systems experimentally characterised for melting temperature (T_m) in the Gautel lab. Melting temperature of a protein system is defined as the temperature at which the equilibrium constant of unfolding K_u is equal to 1.0, or where half of the protein is in the unfolded state and half is in the folded state. Differential scanning fluorimetry (DSF) was used to define melting temperature in 53 wild-type and 66 variant systems, providing an indispensable resource to compare features from MD simulation with specific experimental data (Table 6). There was one mutant domain for which a corresponding wild-type domain was not characterised; this wild-type system was included in the simulation data set, for a total of 120 simulated systems.

3.3.3 Modelling of Domains without Experimental Structure

The majority of variants in the literature/population data set were not located in domains with solved experimental structures; therefore, template-based homology modelling was used to build models of these domains *in silico*. The domains of titin are well suited for this, as they are both numerous and highly homologous, so are well served by existing structural templates. A single reliable experimental template for each wild-type system was selected using protein BLAST[208] to search the Protein Data Bank (PDB)[209] for highly homologous structures, then determining the appropriate template based on sequence identity (as close to 100 as possible) and experimental resolution of the structure (as close to 0 as possible). The sequence of the wild-type domain and the template were aligned using

T-COFFEE[190], and the alignment used as input to SWISS-MODEL[210] to generate the template model (see Table 5). SWISS-MODEL was selected for its ease of use through a web server, as well as in-built quality metrics such as its native QMEAN[211], and additional summary statistics from MolProbity[182]. Having validated that the wild-type models were of good quality according to these parameters, the mutant systems were then generated using the Mutagenesis Wizard tool in PyMol[62] to replicate the residue substitution, followed by energy minimisation in GROMACS as part of the MD simulation pipeline.

For the thermostability data set, the data was acquired after the release of AlphaFoldDB[171] and the previously unavailable, highly accurate, DL-based models of previously unsolved proteins. Thus, the individual wild-type domain structures from the proteome predictions downloaded from AlphaFoldDB (see Chapter 2) were retrieved and the best-quality models for each domain were selected using the MolProbity score and Rama Z-score computed by MolProbity. The mutant systems were generated in the same manner as before, using PyMol and GROMACS.

3.3.4 Molecular Dynamics Simulations and Parameters

Principles of Molecular Dynamics

Atomistic molecular dynamics (MD) approaches aim to simulate the motion of discrete atoms in a protein system in solution, over a specific time frame. As previously mentioned, the movement and interactions between atoms is governed by classical mechanics; the electronic or sub-atomic configuration of atoms is not considered. Thus, from a single starting state, the change in atomic positions, velocities, and forces may be calculated; however, as these values change over time, an integration algorithm is necessary to produce the accurate, continuous trajectory. This algorithm integrates the equations of motion, simulating the motion of the atoms across a series of discrete time-steps (Δt). Many different software programs are available to compute these dynamical trajectories; for this work, I used the GROMACS

software package (version 2021.1)[212], a well-established method for MD simulation that can be run from the command line, with support for parallelisation and GPU acceleration, greatly increasing the efficiency of simulation.

For any given atom i, the relationship between the force (F_i), mass (m_i), and acceleration (a_i) with respect to that atom is given by Newton's second law:

$$F_i = m_i a_i = m_i \frac{\partial v_i}{\partial t} \tag{2}$$

For a known atomistic system in a given configuration, the mass and forces acting on the atom i are already known; thus, the acceleration, or change in velocity over time, can be calculated from these values. If the forces operating on atom i are constant, the velocity and position of atom i at time τ can be calculated by integration plus the initial velocity ($v_i(0)$) and position ($r_i(0)$) of i:

$$v_i(\tau) = v_i(0) + \int_0^\tau \frac{dv_i}{dt} dt \tag{3}$$

$$r_i(\tau) = r_i(0) + \int_0^\tau v_i(t) dt \tag{4}$$

As I have previously established, the forces operating on atom i change as its position changes, and thus the system atomic configuration changes. Thus, by progressing the simulation by Δt of fixed length, the position of atom i at time $t + \Delta t$ can be determined by derivation, using the information from previous time steps. This places an importance on the length of Δt: if too large, the trajectories will be inaccurate, and atoms may even collide with one another; if too small, the trajectory will take a long time to move.

GROMACS comes with an implementation of the Leapfrog Verlet algorithm, which is used as the integrator for all MD simulations. This algorithm is standard for MD simulations of proteins, as it is fast, stable, and time-reversible. The defining feature of Verlet algorithms is that velocity does not need to be explicitly calculated at each time step t; this decreases computational time required, and does not increase error significantly provided Δt is sufficiently small. "Leapfrog" Verlet is a modification to standard Verlet, which is more

memory-efficient; by calculating velocity from half time-steps before ($v_i(t - \frac{\Delta t}{2})$) and after ($v_i(t + \frac{\Delta t}{2})$) time step t, the position at the next time step ($r_i(t + \Delta t)$) may be calculated thus:

$$v_i\left(t + \frac{\Delta t}{2}\right) = v_i\left(t - \frac{\Delta t}{2}\right) + \frac{\Delta t}{m_i} F_i(t) \tag{5}$$

$$r_i(t + \Delta t) = r_i(t) + \Delta t v_i\left(t + \frac{\Delta t}{2}\right) \tag{6}$$

$$v_i(t + \Delta t) = \frac{v_i(t + \frac{\Delta t}{2}) + v_i(t - \frac{\Delta t}{2})}{2} \tag{7}$$

In summary, the MD simulation can be thought of as a sequence of discrete time steps, where forces, velocities, and atomic coordinates are updated in a stepwise, iterative manner: first, calculation of forces on all atoms using the system configuration; second, calculation of movement of atoms using the integrator algorithm; third, calculation of physical properties of the system, which are written to output files.

Force fields and Parameters

In the previous section, I described the calculation of forces operating on the atoms within the system of interest as part of the MD simulation. These forces are described by classical or "semi-empirical" force fields, which represent a collection of potential energy functions ($V(r)$). The gradient of the potential energy approximates the force:

$$F = -\nabla V(r) \tag{8}$$

The force field is constituted by individual potential terms that sum to produce the full potential energy V of a system. These can be partitioned into "bonded" (V_{bonded}) and "non-bonded" ($V_{nonbonded}$) potentials:

$$V(r) = V_{bonded}(r) + V_{nonbonded}(r) \tag{9}$$

The bonded terms describe interactions between atoms in the system that are covalently bonded to one another; thus, they describe intra-molecular interactions between atoms in

the protein only. These consist of covalent bond-stretching ($V_{stretch}$), angle-bending (V_{angle}), proper dihedrals ($V_{torsion}$), and improper dihedrals ($V_{improper}$):

$$V_{stretch}(r) = \sum_{bonds} \frac{1}{2} k_b \left(r_{ij} - r_0\right)^2 \tag{10}$$

$$V_{angle}(r) = \sum_{angles} k_\theta \left(\theta_{ijk} - \theta_0\right)^2 \tag{11}$$

$$V_{torsion}(r) = \sum_{bonds} \frac{V_1}{2}\left[1+\cos(\phi)\right] + \frac{V_2}{2}\left[1-2\cos(\phi)\right] + \frac{V_3}{2}\left[1+3\cos(\phi)\right] \tag{12}$$

$$V_{improper}(r) = \frac{1}{2} k_\phi \left(\phi_{ijkl} - \phi_0\right)^2 \tag{13}$$

$$V_{bonded}(r) = V_{stretch}(r) + V_{angle}(r) + V_{torsion}(r) + V_{improper}(r) \tag{14}$$

Non-bonded terms relate to interactions between atoms in the system that are not covalently bonded, thus both intra-molecular interactions within the protein and inter-molecular interactions between the atoms in both protein and solvent are considered; these consist of van der Waals (repulsion-dispersion term; V_{vdw}) and electrostatic (Coulomb term; V_q) interactions. These terms are computed as pair potentials based on a neighbour list; a cut-off radius (r_{cut}) of 1.2 nm is used to exclude atoms beyond this difference from the neighbour search, in order to simplify computation. This is appropriate for van der Waals' interactions, which rapidly fall off beyond this cut-off radius; dispersion and repulsion are modelled using the 12-6 Lennard-Jones potential V_{LJ}:

$$V_{LJ}(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{r_{ij}}\right) - \left(\frac{\sigma_{ij}^{6}}{r_{ij}}\right)\right] \quad \text{for} \quad r_{ij} < r_{cut} \tag{15}$$

$$V_{LJ}(r_{ij}) = 0 \quad \text{for} \quad r_{ij} \geq r_{cut} \tag{16}$$

For electrostatic interactions, a simple cut-off would introduce greater discontinuity, as these do not decay as rapidly; consequently, long-range electrostatic potentials beyond the cut-off radius are modelled using the Particle-mesh Ewald (PME) method[213, 214]. Briefly, short-range electrostatics within r_{cut} (V_{sr}) are calculated using Coulomb interactions, while long-range electrostatics (V_{lr}) are solved by representation of particle interactions on a mesh

grid in Fourier space:

$$V_q(r_{ij}) = V_{sr}(r_{ij}) + V_{lr}(r_{ij}) \tag{17}$$

$$V_{sr}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{\varepsilon_r r_{ij}} \text{ for } r_{ij} < r_{cut} \tag{18}$$

$$V_{lr}(r_{ij}) = \tilde{\Phi}_{lr}(k)|\tilde{\rho}(k)|^2 \text{ for } r_{ij} \geq r_{cut} \tag{19}$$

$$V_{nonbonded}(r) = V_{LJ}(r) + V_q(r) \tag{20}$$

The parameters used in the force field terms are weighted so as to accurately reproduce the experimental data used for fitting. These data, and the chemical species used, differ between available force fields; for example, CHARMM force fields are fitted to experimental enthalpies, whereas GROMOS and AMBER force fields are fitted to experimental free energies. For all MD simulations described herein, I used the AMBER ff14SB force field[215], a relatively recent improvement on previous AMBER force fields, alongside the Transferable Intermolecular Potential 3P (TIP3P) water model for solvent[216], using the atomic positions from the default Simple Point Charge (SPC-216) water in GROMACS.

The GROMACS port of the ff14SB force field was sourced from GROMACS User Contributions[217]. This port lacked certain improper dihedral values that were necessary for these proteins; thus, the following lines were added to the portable topology file for bonded interactions manually:

NA	CW	CC	CT	4	180.00	4.60240	2
NA	CV	CC	CT	4	180.00	4.60240	2

Simulation parameters and Workflow

Each protein system was simulated as a single protein chain within explicit solvent. The PDB files for each protein were cleaned, where necessary, to remove non-protein atoms, alternative atom positions, and anisotropic temperature factors. The proteins were then centred in a simulation box, to which TIP3P solvent molecules were added using the positions in the *spc*216.*gro* structure file. In order to eliminate edge effects at the borders of

the box, and to enable the use of PME electrostatics, periodic boundary conditions (PBCs) were used to model the system as the original simulation box surrounded by an effectively infinite series of identical periodic images. This necessitates a box shape with a lattice framework; for all simulations a rhombic dodecahedral box was used to minimise the size of the system while maintaining periodicity.

One issue that may arise from the use of PBCs is that the system will interact with its own periodic images, creating artifacts in simulation. Thus, the size of the each protein system specifies a minimum distance of 1.5 nm between the protein and the boundary, increasing computational cost but placing each periodic image beyond the cut-off for non-bonded interaction potentials. Furthermore, the net charge of the system must be neutral to prevent summation over all images to infinity; depending on the charge state of the system, GROMACS automatically neutralises the charge by replacing random solvent molecules with the corresponding counterions (Na^+ for positive, Cl^- for negative).

The systems were relaxed prior to simulation using the steepest descent energy minimisation method, running for 2000 steps with position restraints (force constant = $4000 kJmol^{-1}nm^{-2}$) and 10,000 steps without. Each time step covered 1 fs (10^{-15} s). Subsequently, the protein and solvent were equilibrated in first an NVT (constant number of particles, volume, temperature of system) and then an NPT (constant number of particles, pressure, temperature of system) ensemble. The NVT equilibration consisted of four runs at 50 K, 100 K, 200 K, and 300 K, each with decreasing position restraints (2000, 1000, 500, and 250 $kJmol^{-1}nm^{-2}$, respectively). The parallel LINear Constraint Solver (P-LINCS) algorithm[218] was used for position restraints. Each run consisted of 50,000 steps of 2 fs. The NPT equilibration consisted of four runs at 50 K, 100 K, 200 K, and 300 K, each with decreasing position restraints (2000, 1000, 500, and 250 $kJmol^{-1}nm^{-2}$, respectively). Each run consisted of 50,000 steps of 2 fs. For both NVT and NPT equilibrations, temperature coupling was used with the "velocity rescaling" thermostat[219]. For the NPT equilibration, pressure coupling was used with the Berendsen barostat[220]; an isotropic coupling of 1 bar, with an isothermal compressibility of $4.5 \times 10^{-5} bar^{-1}$.

Post-equilibration, the isothermal-isobaric NPT ensemble was used as starting coordinates for the full MD simulation. While simulation lengths varied depending on the analysis in question, all were between 50 and 400 ns in length. A typical 100 ns production run consisted of 50,000,000 time steps of 2 fs. Cartesian coordinates of atoms were written to the trajectory every 5000 steps, producing a 10001-frame trajectory for each 100 ns simulation. Energies were written to the energy output file every 1000 steps. Again, the "velocity rescaling" thermostat was used for temperature coupling, but here the Parrinello-Rahman barostat[221] was used for pressure coupling.

3.3.5 External computing resources

In order to run large numbers of MD simulations in parallel with GPU acceleration, I applied for and received computing time on several external high-performance computing (HPC) clusters. Production runs for initial simulations were performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (`www.csd3.cam.ac.uk`), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (EPSRC; capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (`www.dirac.ac.uk`). The majority of longer simulations were performed using the Joint Academic Data Science Endeavour (JADE2) HPC, using time granted via the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim (`http://hecbiosim.ac.uk`), supported by EPSRC (grant no. EP/R029407/1).

3.3.6 Feature Extraction

The recorded molecular dynamics trajectory is composed, at its core, of a large matrix of XYZ Cartesian coordinates for each frame in the trajectory, for all atoms in the system. In order to improve computational efficiency and reduce memory requirements, the atoms in

water molecules are removed, and to allow direct comparison between wild-type and mutant systems, the backbone alpha carbon atoms are extracted and the remainder discarded. Thus, for both wild-type and mutant systems, the number of atoms in the trajectory is equal to the number of residues in the protein, which for a single titin domain is on the order of 10^2 residues; however, this still produces a matrix, for a 100 ns simulation of a single domain, of approximate size 300×10001. This quantity of data precludes detailed comparison between species or systems; consequently, features are extracted from the data that reflect biologically meaningful dynamical properties. GROMACS allows users to natively perform trajectory analysis techniques using the structure, topology, and trajectory files generated as part of the MD simulation.

Root mean square deviation (RMSD) is a method for atom-wise comparison of multiple superimposed structures, and is commonly used to determine the degree to which the protein backbone deviates from its starting structure over the course of the simulation[222]. The RMSD at each frame may be described as the minimum distance, when superimposed, between all atoms in the starting topology and their corresponding atom in the that frame; the value reported is the square root of the mean of the square difference between Cartesian coordinates across all backbone alpha carbon atoms, in nanometres (nm). This gives a measure of the flexibility of the protein backbone over time. A second measure, root mean square fluctuation (RMSF), alternatively presents the displacement on a per-residue basis, using the average deviations over the entirety of the simulation. A key difference from RMSD is that RMSF calculates deviations from the average position of that residue across the entirety of the simulation, rather than from the starting position, thus presenting a more granular formulation of domain flexibility.

The radius of gyration (Rg) is another commonly used feature for trajectory analysis. This value reflects the compactness of the protein in that trajectory frame; the more stably folded the protein, the less this value is expected to change over time. As with RMSD, a root mean distance value across all atoms is computed each frame; here, the distance reflects the deviation between the atom and the centre of mass of the protein. Four other structural

properties were calculated per frame of the trajectory: the solvent-accessible surface area (SASA - calculated with double cubic lattice method[223] using van der Waals radii given in Bondi (1964)[224]); the secondary structure elements (SSE - calculated with DSSP[189]); the number of hydrogen bonds between residues in the protein; the number of contacts between atoms within 0.6 nm in the protein. Finally, the SASA was additionally computed on a per-residue basis.

Each of these eight properties were written by GROMACS to an XVG data file, which was read into R (version 4.0.4)[225] using the *readXVG*() function from the *Peptides* package[226]. The mean and standard deviation of each property was computed for each individual trajectory, as well as the difference between values for wild-type and mutant trajectories.

3.3.7 Principal Component Trajectory Analysis

As I have discussed, the raw trajectory data represents an extremely high-dimensional data set, which must be converted to a form suitable for further analysis. Extraction of individual features relevant to the dynamics of the system is one method to compare between trajectories; however, each individual feature incorporates only a small minority of the data available in the trajectory file. Thus, if the trajectory data is to be explored as a whole, it is necessary to explore methods of compressing the data set while preserving as much of the variance in the trajectory data as possible. These motivations form the basis of dimensionality reduction techniques, which are frequently employed both in data analysis and visualisation methodologies.

Principal component analysis (PCA)[227] is a commonly used algorithm for linear transformation of a larger number of variables in a data set to a smaller number of uncorrelated principal components; thus, it can also be used on trajectory data, by compressing the Cartesian coordinates of all of the backbone alpha carbon atoms in a single frame to a smaller

number of dimensions. In brief, the PCA implementation for reduction of n-dimensional data to k-dimensional data locates k vectors onto which the original data set is projected, such that the projection error is minimised. This allows for each principal component thus computed to incorporate a known quantity of the variance of the original data set.

As with the RMSD calculations described above, I conducted PCA on the alpha carbon backbone subset of atoms in the trajectory, so as to visualise both wild-type and mutant trajectories within the same principal component plot. To do so, the wild-type and mutant trajectories were directly concatenated in R, such that the first time step of the second trajectory followed immediately from the final time step of the first, and so on for all replicates. The points were then coloured according to their trajectory of origin; greyscale points were used for wild-type trajectories, while coloured points were used for mutant trajectories. 2D plots were generated for the first 2 principal components for each WT-mutant pair in each of the data sources, using the *Bio3D* R package[228].

3.3.8 Structural Alphabet Encoding and Mutual Information

A second method for dimensionality reduction for protein trajectory data relies on the encoding of the protein structure in a rotationally invariant language, rather than recording the position of each atom in Cartesian coordinates. Structural alphabet encodings are one means of achieving this; by describing each overlapping 4-residue fragment of the 3D protein structure by similarity to a finite set of pre-defined, representative conformations, the structure of a protein of sequence length n may be encoded as a 1-dimensional sequence of letters of length $n-3$, which in turn allows for the protein backbone conformation in coordinates to be reconstructed. By extending from individual structures to frames of a trajectory, the change in conformation over time may be observed in the change in encoded sequence across the trajectory. There are many different structural alphabets described in the literature; here, the R distribution of the *PDBencode* package is employed, which encodes a structure or trajectory using the structural alphabet M32K25[229].

The encoded trajectories are analysed using an implementation of the *AlloHubMat*[230] framework in Python (*AlloHubPy*[231]), which incorporates the structural alphabet dynamics analysis package *GSATools*[232]. To compute the similarity between different trajectories, the trajectory files are read into Python and encoded into the M32K25 structural alphabet format using the *TrajSAencode* package. The alphabet-encoded trajectories are split by a set block size; the normalised mutual information (MI) matrices of pairwise correlations in the encoded structure are calculated for all blocks in the trajectory. These blocks are themselves compared with all other blocks in the same trajectory, in replicate trajectories, and between wild-type and mutant trajectories, by computing pairwise overlap between eigensystems of MI blocks, and thus calculating a similarity score between 0 (identical) and 1 (fully independent). These scores are plotted as a heat map.

3.4 Simulation of Pathogenic vs Non-Pathogenic Variant Set

3.4.1 Features indicative of Pathogenicity

The 85 systems (24 wild-type, 61 variant) in the literature/population data set were simulated for 100 ns each, producing a trajectory of 10001 frames in length. Each simulation was replicated 2 further times, for a total of 300 ns of simulation time for each system, or 255 simulations in total. The features of each individual trajectory were computed using the methods described above. As I was interested in the difference in calculated features from the literature and population databases, the features calculated from the wild-type trajectories were used to compute additional properties of the mutant trajectories relative to the wild-type. The strength of the relationship between these individual dynamical features and the data class in which the variant was found were then investigated.

Across all features tested, there were several features for which the distinction between "Literature" and "Population" variants was significant, by comparison of population means with the non-parametric Mann-Whitney U-test[233] (see Figure 21). The strongest indicator

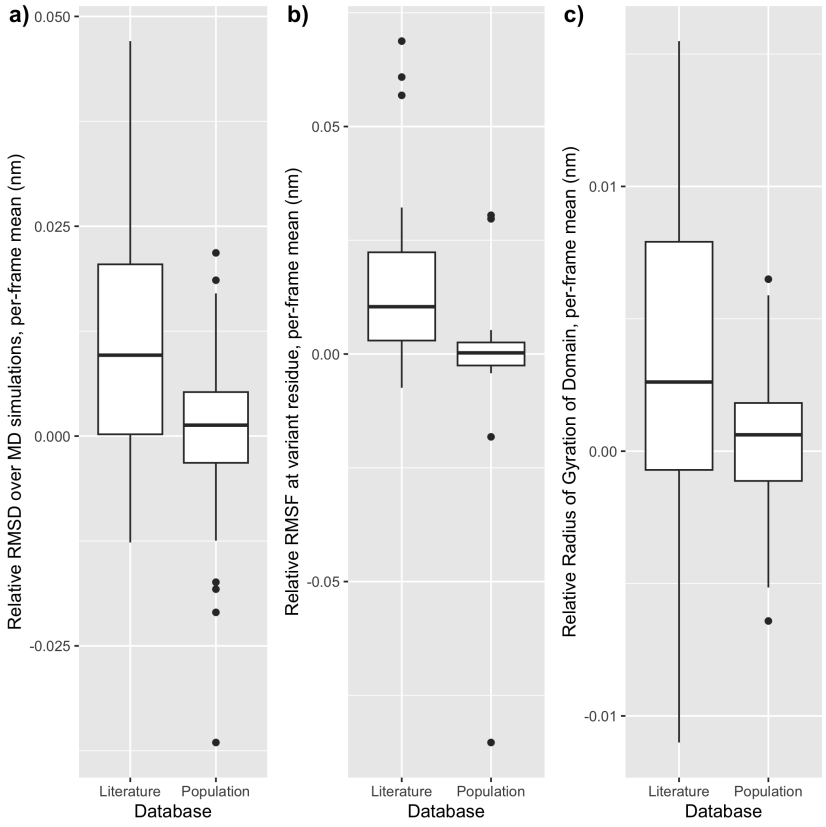


Figure 21: Distributions of dynamical features calculated from trajectories, versus class

*A comparison of selected dynamical features between the Literature and Population data sets. Relative values of features were calculated for **a)** root mean square deviation (RMSD), **b)** root mean square fluctuations (RMSF), and **c)** radius of gyration by obtaining the per-frame average value for each feature across all simulation frames and replicates, then subtraction of the mean for the mutant systems from the wild-type systems. In all three of the above cases, the difference between features from Literature and Population variants was determined to be significant, with RMSD (Mann-Whitney U-test, $p = 0.0005$) and RMSF ($p = 1.195\text{e}{-6}$) especially so. The difference in relative radius of gyration was lesser but still significant ($p = 0.0251$).*

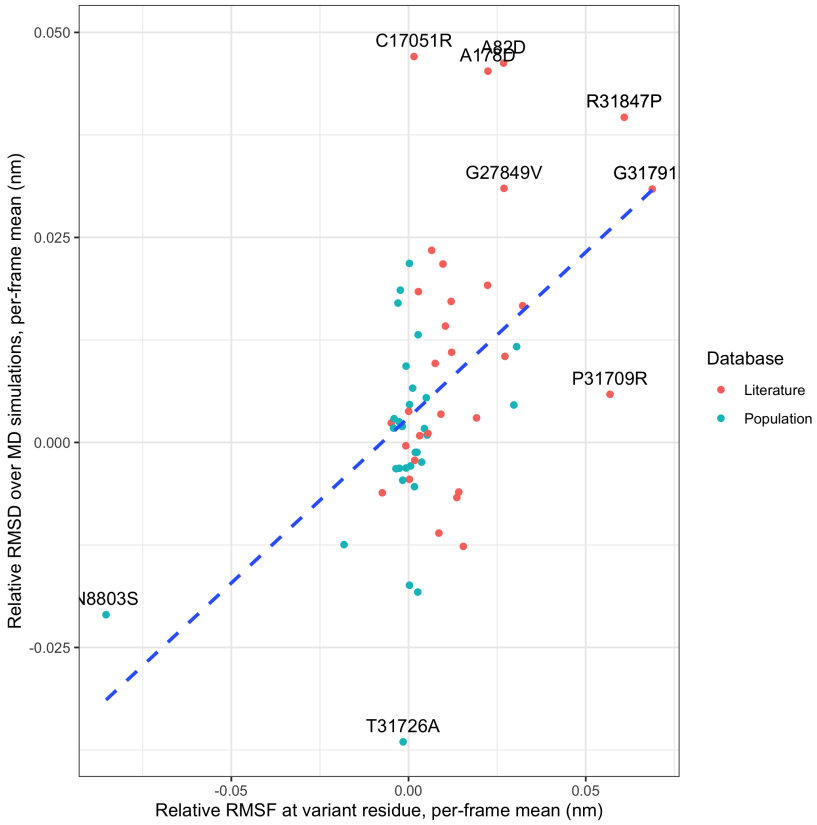


Figure 22: Relative RMSF and relative RMSD comparison of variant systems, by class

Plot of relative RMSD against relative RMSF for all systems in "Literature" and "Population" data sets. The two features are moderately positively correlated, with regression line given in blue (Spearman's rank correlation, $\rho = 0.608$, $p = 3.915\text{e}{-7}$). Variants with extreme values of either RMSD or RMSF are labelled.

of the class to which a variant belongs is given by the relative RMSF ($RMSF_{rel}$) between wild-type and mutant trajectories (Mann-Whitney U-test, $p = 1.195\text{e}{-6}$), where almost all relative values above 0.0 belong to "Literature" variants, and almost all below 0.0 are "Population" variants, suggesting that while more damaging variations tend to perturb the domain at that position *in silico*, most neutral variants in this data set do not, or even stabilise that position. By comparison, the relative RMSD ($RMSD_{rel}$) between wild-type and mutant trajectories is also significant, but less so ($p = 0.0005$). This is to be expected, as the RMSD includes the whole of the protein in its calculation, rather than only the region of interest; thus, the signal would be expected to be diminished compared to the RMSF, which only considers the mutated residue position.

While there is overlap between the distributions of "Literature" and "Population" variants for all features, this is to be expected; these features each represent only a single characterisation of the motions of the domains *in silico*, and the variants in the literature/population data set are diverse in their associated conditions, reported pathomechanisms, and inheritance patterns. Thus, it would be unlikely for all of them to exhibit the same aberrant dynamical properties compared to the wild-type. This being the case, I investigated whether or not the different features were sampling different properties of the simulation, and thus may in combination offer a more clear separation between classes.

As seen in Figure 22, when relative RMSF and RMSD are plotted together, there is some degree of information not shared between the two features; notably, the variants with an average relative RMSD below -0.015 nm are all "Population" variants, a distinction which is not captured by RMSF, despite RMSF having a stronger discriminatory power between "Literature" and "Population" variants overall. Variants such as C17051R and T31726A have values of average relative RMSD at opposite extremes, but very similar values for relative RMSF. However, the majority of the information that is included in these two features overlaps; a value with a high value for one feature is typically high in the other. This may be expressed as a moderately positively correlation (Spearman's rank correlation, $\rho = 0.608$, $p = 3.915\text{e}{-7}$). Thus, the overall increase in discriminatory power from combining these two features is less than the sum of the individual features.

Interestingly, while the dynamical features calculated from trajectories share information with one another, they share much less information with other common sequence- and structure-based features typically used for classification of damaging variants, and thus have the potential to meaningfully add to these vairant annotations. For example, the position-specific independent counts (PSIC)[234] is a feature used by PolyPhen-2[144] to represent the evolutionary conservation of a given residue at a specific position; by plotting the wild-type PSIC ($PSIC_{WT}$) of the variants in this data set against their respective $RMSF_{rel}$ from simulation (see Figure 23a), a much more clear separation of the data points is visible. A high conservation of the wild-type position and low $RMSF_{rel}$ is indicative of "Population" variants,

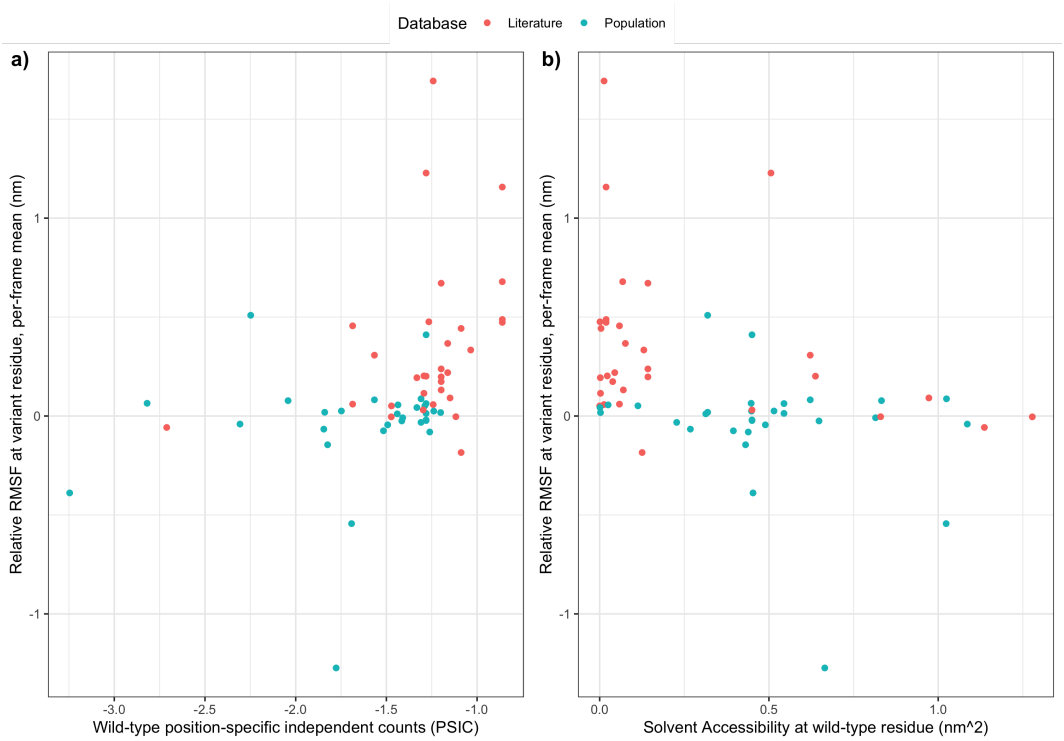


Figure 23: Relative RMSF of variant systems against wild-type PSIC and SASA, by class

Scatter plots of features associated with variants in the Literature (red) and Population (green) data sets. The average difference in RMSF (nm) at the mutant residue versus the wild-type residue at the same position across all trajectory frames (100ns) for these variants is plotted against a) the wild-type PSIC and b) the wild-type SASA (nm^2) at these residue positions.

whereas the reverse is the case for "Literature" variants (Spearman's rank correlation, $\rho = 0.352$, $p = 0.005$). Similarly, another oft-used feature for variant classification is the solvent-accessible surface area of the wild-type residue ($SASA_{WT}$); again, a high solvent accessibility and low $RMSF_{rel}$ is indicative of "Population" variants, and the opposite for "Literature" variants (Spearman's rank correlation, $\rho = -0.392$, $p = 0.002$, see also Figure 23b). This demonstrates how multiple features from different levels of data can add information lacking in the others.

3.4.2 PCA Cluster Analysis

With the observation that different features from trajectory analysis could, collectively, describe the separation between literature and population variants in greater detail than any one feature alone, I decided to use principal component analysis to capture the extent of backbone conformational differences between wild-type and mutant trajectories, and to determine the power of unsupervised dimensionality reduction techniques to contribute derived features in their own right.

Typically, a PCA is most useful in trajectory analysis for the purposes of visualisation; for such a high-dimensional data set as even the Cartesian coordinates of trajectories alone, absent velocities or energies, the projection of each frame to a single point on a two-dimensional grid allows for the observation of the extent to which the variation in this data set represents the difference between wild-type and mutant trajectories. A particularly striking example of this difference may be seen for the A178D variant in Ig-2, which is associated with dilated cardiomyopathy and has been extensively characterised both *in vitro* and *in vivo*. As seen in Figure 24a, the progression of the trajectories in conformational space from their starting configuration can be compared between wild-type (greyscale) and mutant (RGB) systems, by observing the frames in lighter (earlier) and darker (later) colours. Here, the wild-type trajectories are confined to a limited space represented as negative values in the first principal component (PC1); the mutant trajectories, while beginning in the same configuration, each progress over the course of the simulation to sample a new region of conformational space of positive values in PC1. There is little to no visible difference between wild-type and mutant trajectories in the second principal component (PC2). This may be compared to the population variant R27839Q (Fn3-90; Figure 24c), where all trajectories overlap within a more restricted region of conformational space.

The structural consequences of this variation in Ig-2 in projected dimensions may be visualised by superimposing the final frames of each of the six systems (see Figure 24b). Here, the tertiary structure of the domain is intact in both the wild-type and mutant domains;

however, the greatest discrepancy in the structures can be seen at and around the site of mutation (magenta, all systems). In the wild-type systems, the flexible loop that follows the mutation site (F-G loop) is relatively consistent in its conformation; however, in the mutant systems, the loop detaches, seemingly losing intramolecular contacts, as well as a breakdown of secondary structure on either side of the loop. Thus, the PC1/2 plot has captured the destabilisation of this region in the trajectory; by contrast, in the same representation for the Population variant R27839Q, both the wild-type and mutant systems exhibit a very similar final conformation (Figure 24d), demonstrating the relative lack of perturbation of the wild-type structure.

The visual inspection of the Ig-2 PC1/PC2 plot suggests that, in principle, a destabilising or otherwise structurally perturbing mutation should cause the variant trajectory to explore a different region of conformational space, whereas a neutral variant should overlap more with the space explored by the wild-type trajectory; in other words, a separation, as projected onto PC1/PC2, should be seen between wild-type space and mutant space, commensurate to the structural perturbation effected by the mutation. However, such a straightforward relationship is not realistic; I have already discussed the heterogeneity of pathomechanisms for the "Literature" variants in this data set, and how association of a variant with disease need not necessitate destabilisation of the parent domain, much less the altered flexibility of the backbone atoms. Nevertheless, across the literature/population data set, 21 of the 31 "Literature" variants demonstrated a clear separation in their wild-type and mutant trajectories, versus 8 of the 30 "Population" variants.

All simulations outlined above were repeated again for 100ns, thus giving a total of 6 trajectories of 100ns each for each of the wild-type and variant systems, to increase the coverage of conformational space accessed by the different systems. To determine whether the separation between wild-type and mutant trajectories could represent a meaningful feature to discriminate between literature and population variants, a simple clustering procedure was implemented to describe the separation numerically. Briefly, a centroid for each trajectory is described by the median value for each principal component; the distances between these

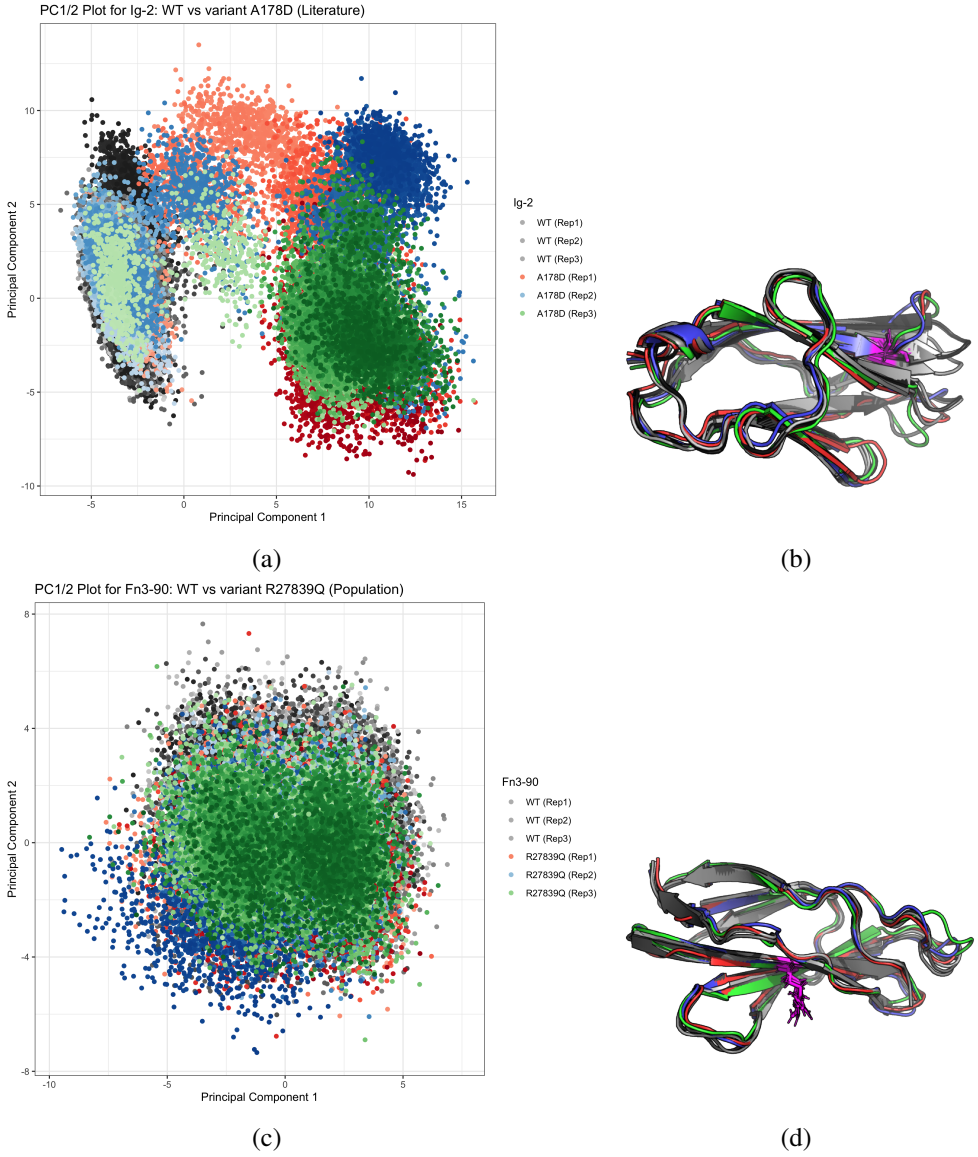


Figure 24: Visual representations of trajectory separation between typical Literature and Population variants

Visual representations of trajectory separation between six trajectories with wild-type (greyscale) and mutant (red, blue, green) residues for the disease-associated variant A178D (Ig-2) and the population variant R27839Q. Principal Component Analysis: a) PC1/PC2 plot of 3 mutant A178D and 3 wild-type trajectories of Ig-2; c) PC1/PC2 plot of 3 mutant R27839Q and 3 wild-type trajectories of Fn3-90. Each point in the plots represents 1 frame in each 100ns trajectory (10001 frames), where the Cartesian coordinates of all backbone α-carbon atoms in that frame are projected onto the first two principal components. Lighter colours indicate earlier frames while darker colours indicate later. Structural superimposition: b) Final frame positions of 3 mutant A178D and 3 wild-type trajectories of Ig-2; d) Final frame positions of 3 mutant R27839Q and 3 wild-type trajectories of Fn3-90. Site of mutation is represented in both wild-type and mutant structures (magenta, licorice).

centroids were then computed to determine the separation between two individual trajectories, as the sum of the distances for each individual dimension. Of note is that this distance metric computes those variants that substantially decrease the dynamical flexibility of the domain as well as those that increase it.

Initially, three different comparisons were made: the separation of wild-type trajectories of the same system from one another; the separation of mutant trajectories of the same system from one another; the separation of these mutant trajectories from their wild-type counterparts (see Figure 25). The distance between separate wild-type trajectories is low on average, which suggests that multiple replicates typically explore the same conformational space, as projected onto PC1/PC2. The distributions of these distances for both the literature and population variants are similar, suggesting that neither database is enriched for variants in inherently more flexible domains. There are two outliers for which this is not the case in wild-type domains where "Literature" variants are found (Ig-87, Fn3-20) and one for "Population" variants (Ig-23). Similarly, while the distance between separate mutant trajectories is on average greater than for the wild-type, the distributions of literature and population variants are again similar. This indicates, again, that the variant domains tend to explore the same space across replicates, but that there are some minor differences between individual trajectories.

Encouragingly, the greatest difference between literature and population distributions was observed for the separation between mutant and wild-type trajectories. Here, the literature variants tended to have a greater degree of separation between wild-type and mutant trajectories, as projected onto PC1/PC2, than the population variants; however, while the difference between the two distributions was significant (Mann-Whitney U-test, $p = 0.0149$), it was much less so than for single features such as relative RMSF. This indicates that the centroid distances derived from dimensionality reduction methods do not reliably distinguish between the literature and population variants. Interestingly, while the median separation distance for literature variants was higher than that for population variants, there were a number of ostensibly pathogenic systems which showed almost no difference from wild-type systems. Consequently, the "Literature" variants in the data set were isolated for closer inspection,

to determine whether any subsets of the reported pathogenic variants were pertinent to the variability in their trajectories.

Intriguingly, of the "Literature" variants found to be most similar in PC1/PC2 plots to wild-type, many of these were variants experimentally characterised as destabilising *in vitro*. Indeed, the most similar were composed of four variants from Ig-169, and one from Ig-19. The Ig-169 variants L35956P, H35946P, I35947N and T35915P are associated with tibial muscular dystrophy (TMD), and the Ig-19 variant T2896I, associated with arrhythmogenic right ventricular cardiomyopathy (ARVC). Interestingly, all of these variants are believed to relate to pathogenicity on the basis of disrupting interactions: L35956P and H35946P abolish the binding of *TTN* Ig-169 to the Ig-1 domain of its fellow giant protein, *OBSCN*, while T2896I is an interface-buried residue in the I-band, which has been suggested to destabilise the chain dynamics of titin when flanked by adjacent domains.

On closer inspection, it seemed to be the case that all variants in Ig-169, regardless of whether from the "Literature" or "Population" data sets, were very similar in terms of conformational space explored; the separation distance values for Ig-169 variants as well as the variants found in the other significant hotspot domain, Fn3-119, were plotted against one another (see Figure 26). Here, intriguingly, despite the observation that Fn3-119 variants are substantially more different to their wild-type domain in their backbone motions than Ig-169 variants, they too are not significantly different from the population variants found in the same system. This would seem to suggest that the majority of variance in the backbone conformations of these systems is a product of the domain as a whole, and masks the signal from the variant impact on the domain.

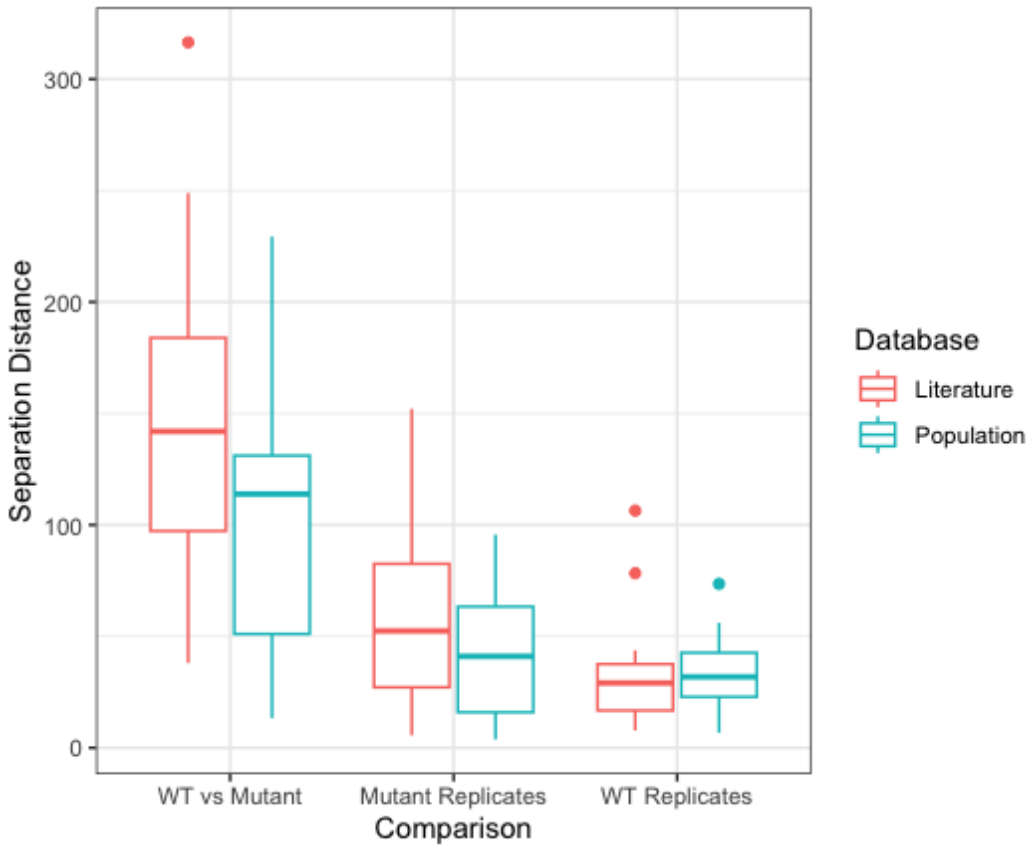


Figure 25: Separation Distances between WT and mutant trajectories for variants, by Database

Box-plots of differences between conformational space explored by mutant and wild-type trajectories as projected onto PC1/PC2. Separation distances between two trajectories are defined by the 2D distance between the two centroids for each trajectory in PC1/PC2 space, where centroids are calculated as the average position across all 10001 frames for each trajectory. The sum of pairwise distances between all 6 mutant trajectories ("Mutant Replicates") and all 6 wild-type trajectories ("WT Replicates") were computed for all variant systems in the Literature (red) and Population (green) data sets, followed by the sum of all pairwise mutant-wildtype distances ("WT vs Mutant"). These were then summed across all systems in these data sets and plotted as above.

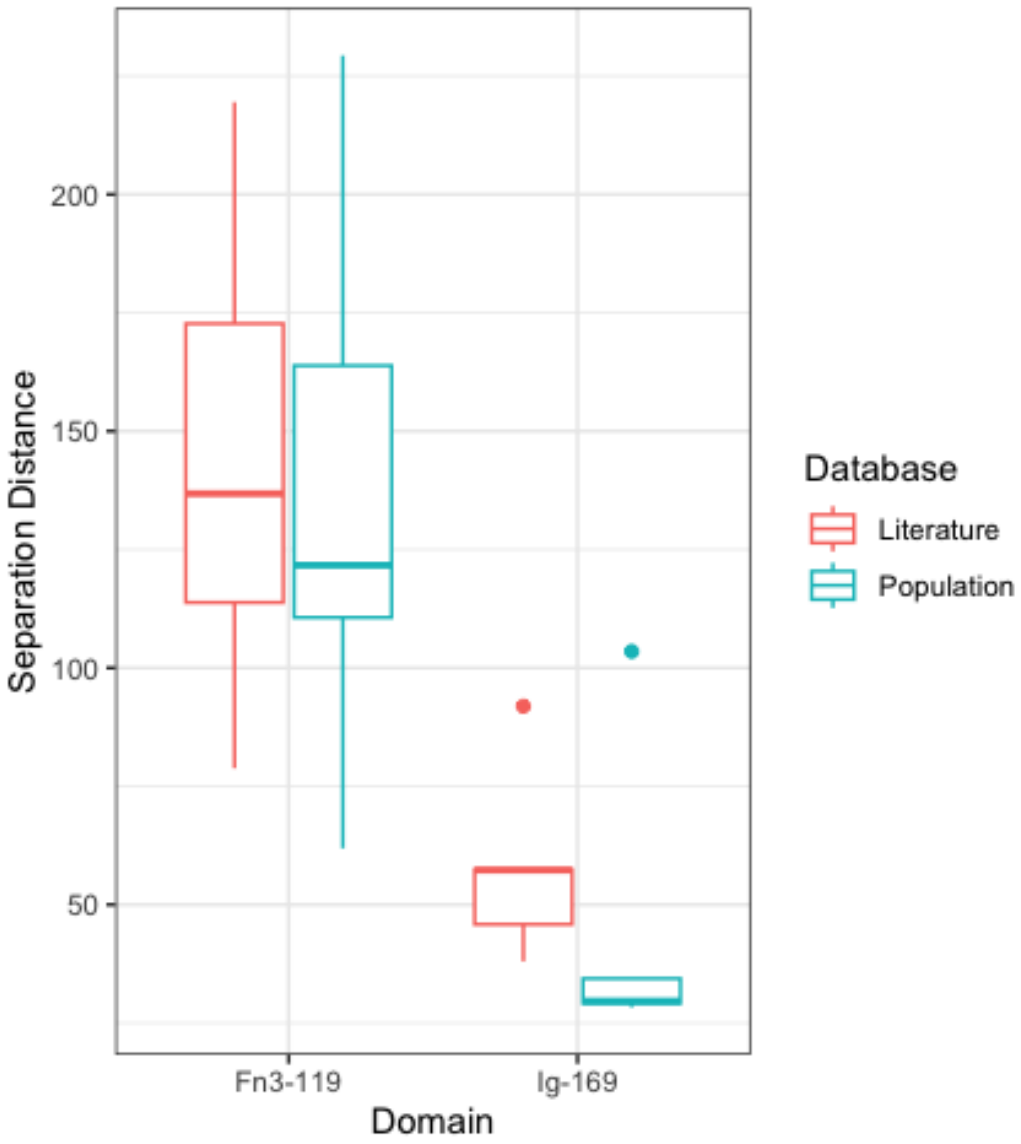


Figure 26: Separation Distances between WT and mutant trajectories for Fn3-119 and Ig-169 variants, by Database

Box-plots of differences between conformational space explored by mutant and wild-type trajectories as projected onto PC1/PC2. Separation distances between two trajectories are defined as before, by the 2D distance between the two centroids for each trajectory in PC1/PC2 space, where centroids are calculated as the average position across all 10001 frames for each trajectory. The sum of all pairwise mutant-wildtype distances ("WT vs Mutant") were calculated for all variants in the Literature and Population databases found in the hotspot domains Ig-169 and Fn3-119. These were then grouped by domain and source database, summed across all systems, and plotted as above.

3.5 Simulation of Thermostability Variant Set

3.5.1 Features indicative of Thermostability

In the previous section, I investigated the most common trends of MD simulation trajectories for variants in two curated data sets, each of which represented a notionally destabilising or stabilising mutation introduced to a naturally stable single titin domain. However, as the binary annotation of these variants as literature- or population-derived ignores potential discrepancies within these distributions, and does not constitute a continuous distribution, there is a limit to conclusions that may be drawn from their analysis. Thus, I elected to repeat the analysis done with another data set, this time of experimentally characterised systems with a single measured parameter, the melting temperature (T_m), which reflects the thermal stability of the domain. Thus, features derived from the simulation with a specific experimental value, which should be described by the change in dynamics of the protein, may be compared between domain systems.

The 120 systems (54 wild-type, 66 mutant) were simulated for 200 ns each, for 3 replicates, giving a total of 600 ns of simulation time, or 360 individual simulations. Again, features were extracted from the trajectories and average values across the whole trajectory were calculated. For each WT-mutant pair, the difference in experimental thermal stability (ΔT_m) was calculated, and all pairs were classed according to the direction and magnitude of stability change:

- "Weakly Stabilising" - $\Delta T_m \geq 0°C$ (n = 11)
- "Weakly Destabilising" - $0°C > \Delta T_m \geq -5°C$ (n = 33)
- "Moderately Destabilising" - $-5°C > \Delta T_m \geq -17°C$ (n = 19)
- "Strongly Destabilising" - $\Delta T_m \leq -17°C$ (n = 2)

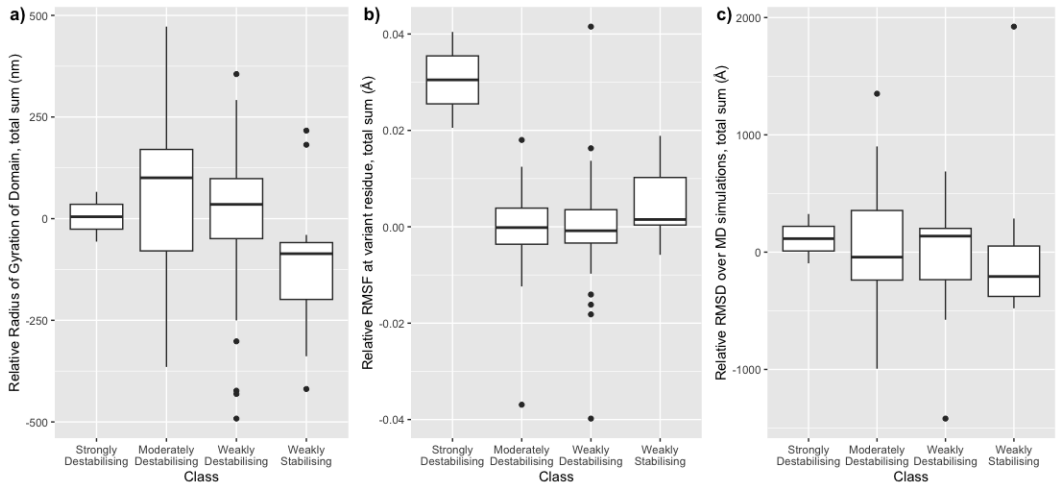


Figure 27: Feature differences between WT and mutant trajectories for variants, by Class

A comparison of selected dynamical features between variants in the thermostability data set in different stability change classes. Relative values of features were calculated for ***a)*** *radius of gyration (Rg),* ***b)*** *root mean square fluctuations (RMSF), and* ***c)*** *root mean square deviation (RMSD) by summation of the total values for each feature across all values and replicates, then subtraction of the total for the mutant systems from the wild-type systems. Here, when all destabilising variants are grouped as a single category, only the relative Rg is significantly different between destabilising and stabilising variants (Mann-Whitney U-test,* $p = 0.0415$*).*

As the majority of variants in the data set had a ΔT_m of $\pm 5^{\circ}C$, representing a very minor stability change, these were classed as either weakly destabilising or stabilising, depending on the direction of stability change. "Moderately" destabilising variants were those that reduced the melting temperature of the system by up to $17^{\circ}C$, a value which has been proposed[35] as indicative of involvement of the variant in titin-related pathology. Variants more destabilising than this value were considered "strongly" destabilising.

Unlike with the literature/population data set, the $RMSD_{rel}$ and $RMSF_{rel}$ features that were most descriptive for the literature/population data set are not so for the thermostability data set, with no significant differences between any of the groups (see Figure 27). Intriguingly, the relative radius of gyration, Rg_{rel}, is here the only feature for which the difference between "stabilising" and "destabilising" variants is significant, albeit only when all destabilising variants are combined into a single category (Mann-Whitney U-test, $p = 0.0415$). Indeed, of the 9 variants in the "Weakly Stabilising" category, only 2 have an overall positive

Rg_{rel} (mutant domain more compact than wild-type). This suggests a pattern within this data set where weakly stabilising variants tend to result in more compact mutant domains (negative Rg_{rel}) whereas destabilising variants result in less compact mutant domains (positive Rg_{rel}).

To compare these dynamical features to the explanatory power of other features, the differences in solvent accessibility between stabilising and destabilising variants were calculated. The solvent accessibility of the mutated residue is generally considered to be a reasonable proxy for the stability change associated with mutation, as more destabilising mutations would tend to be found in the core of the protein, where the introduction of hydrophilic or charged residues would result in a less stable configuration. In this data set, as with the Rg_{rel} feature, there was a significant difference between stabilising and destabilising variants for both the average difference between the solvent-accessible surface area of the mutated residue position in the wild-type and mutant simulations, $SASA_{diff}$ (Mann-Whitney U-test, $p = 0.0089$), and the SASA of the wild-type system alone, $SASA_{wt}$ ($p = 0.0109$). Interestingly, the significant difference between the $SASA_{diff}$ of stabilising and destabilising variants suggests that stabilising variants tend to occur to exposed residues that become less exposed on mutation to stabilising variants; however, this is not unique to stabilising variants, as several destabilising variants are less exposed at their residue position for the mutant domain compared to the wild-type.

As with the literature/population data set, plotting Rg_{rel} against $SASA_{diff}$ shows how these two features can complement one another in discriminating between stabilising and destabilising variants (see Figure 28a). While not as pronounced as for pathogenic variants, we can see how the majority of stabilising variants (blue) cluster with higher values of wild-type SASA combined with lower values of relative radius of gyration. However, even here there are many destabilising variants with similar values for these features as stabilising variants, and there is little to no separation between the weakly and strongly destabilising variants within these features alone. While the features themselves are not strongly correlated with one another (Spearman's rank correlation: Rg_{rel} with $SASA_{diff}$, $\rho = -0.439$, $p =$

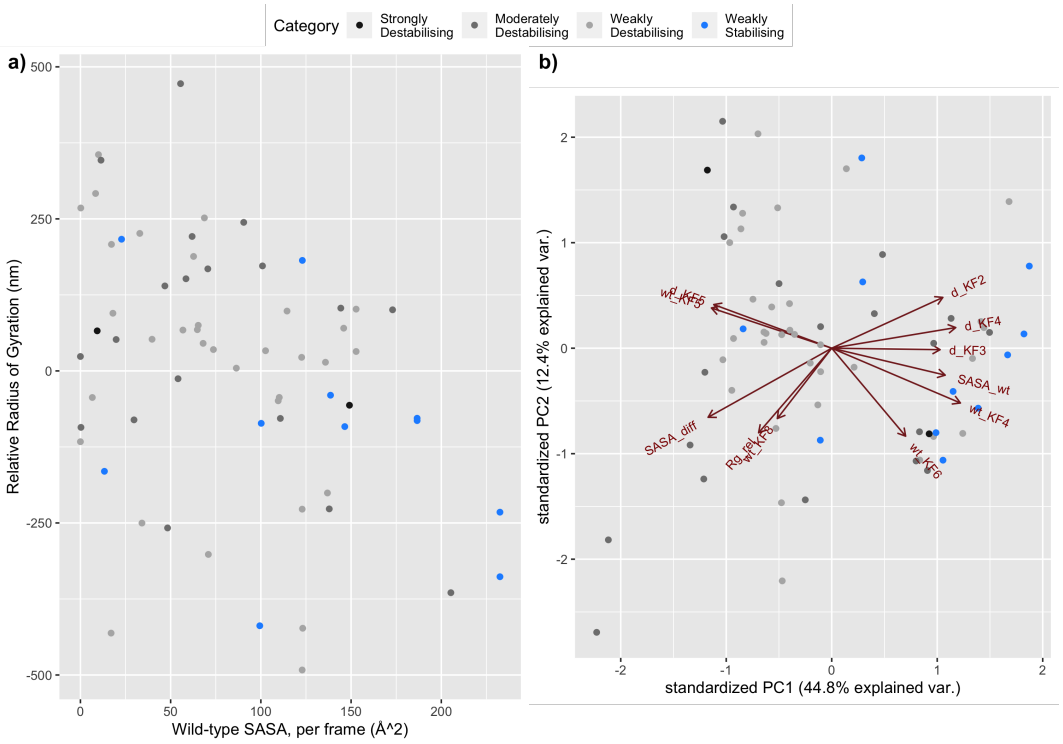


Figure 28: Separation of thermostability data set by descriptive features

Distributions of thermostability data set by dynamical and non-dynamical features correlated with change in melting temperature: a) Scatter plot of the sum total difference in radius of gyration (nm) of the mutant domain versus the wild-type domain across all trajectory frames (200 ns) against the wild-type SASA (Å^2) at the mutated residue position for all variants in the thermostability data set; b) PC1/PC2 biplot of variants in the thermostability data set as projected onto principal components 1 and 2 of features associated with significant difference between stabilising and destabilising variants. Principal component analysis was conducted on the feature set for features that were significantly different between stabilising and destabilising variants (Rg_{rel}, $SASA_{WT}$, $SASA_{Diff}$, $KF4_{WT}$, $KF5_{WT}$, $KF6_{WT}$, $KF8_{WT}$, $\Delta KF2$, $\Delta KF3$, $\Delta KF4$, $\Delta KF5$). Variants are coloured by Category of stability change versus wild-type: strongly destabilising (black); moderately destabilising (grey); weakly destabilising (light grey); stabilising (blue). The PC1/PC2 biplot was overlaid with the direction and magnitude of contributions of the different features to PC1 and PC2 (red).

0.0002; Rg_{rel} with $SASA_{wt}$, $\rho = -0.318$, $p = 0.0087$; $SASA_{wt}$ with $SASA_{diff}$, $\rho = -0.330$, $p = 0.006$), there is sufficient information missing from dynamical and structural data that thermostability cannot be reliably inferred from these features alone.

Thus far, explicit physico-chemical features associated with these residue substitutions have not been described. To investigate how such features compare to the aforementioned dynamical and structural properties in characterising the variant system, each variant was annotated with each of the 10 average Kidera factors[235] for the wild-type residue, and 10 further features representing the difference in these values between the wild-type and the mutant residue (ΔKF). These represent a formulation of the physical properties of each of the naturally-occurring amino acid residues, which together describe 86% of the 188 physical properties used to construct them. For this data set, four wild-type Kidera factors were significantly different between stabilising and destabilising variants: $KF4_{WT}$, relating to hydrophobicity (Mann-Whitney U-test, $p = 0.0297$); $KF5_{WT}$, relating to double-bend occurrence ($p = 0.0340$); $KF6_{WT}$, relating to average partial specific volume ($p = 0.0340$); $KF8_{WT}$, relating to α-region occurrence ($p = 0.0079$). In addition, four ΔKF features were significantly different: $\Delta KF2$, relating to the change in side-chain size ($p = 0.0356$); $\Delta KF3$, relating to extended structure preference ($p = 0.0273$); $\Delta KF4$ ($p = 0.0250$); $\Delta KF5$ ($p = 0.0261$).

That both $KF4_{WT}$ and $\Delta KF4$ would be significantly different between stabilising and destabilising variants is to be expected; in a beta-sandwich domain fold, as in immunoglobulin and fibronectin domains, the hydrophobic properties of core residues and the hydrophilic properties of surface residues contribute to the stability of the fold, such that any extreme value, or extreme change on residue substitution, would likely represent a decrease in thermal stability on mutation. Equally, one might expect changes in size or volume of the residue at a given position, or the secondary structure associated with that position, to be associated with the direction of stability change. In Figure 28b, a principal component analysis of the variant set describes almost 60% of the provided variance in only 2 principal components. Here, once again we see the majority of stabilising variants clustering in the region of positive

values of PC1; encouragingly, this clustering is more pronounced than for the previous plot of radius of gyration versus solvent accessibility, illustrating that this data set does contain signal that may discriminate between stabilising and destabilising variants. Even so, the important issue of the delineation of variants by differing degrees of destabilisation is still unresolved.

3.5.2 Case Study: Fn3-70 as a Model System

While certain features in the thermostability dataset appeared able to discriminate between stabilising and destabilising variants, especially in combination, no such separation was observed for the different categories of destabilising variants. As the magnitude of the destabilisation of the domain should greatly influence the pathogenicity of the variant, the inability of the selected features to distinguish between them merits further investigation. The domain Fn3-70 in this data set was identified as an ideal test system; in addition to the wild-type system, three different variants in this domain have been characterised in this data set. These constitute the "Strongly Destabilising" variant R7W ($\Delta T_m = -27.33°C$), the "Strongly Destabilising" variant D69H ($\Delta T_m = -12.40°C$), and the "Weakly Destabilising" variant Y27C ($\Delta T_m = -0.01°C$). Thus not only are there a representative variant from the same system in each category, but the "Weakly Destabilising" Y27C variant is almost indistinguishable within experimental parameters from the wild-type domain. Consequently, this variant is expected to behave, for features associated with thermal stability, very similarly to the wild-type system.

It is instructive, for sake of comparison, to visualise the features thus far described as compared to one another. In Figure 29, each of the normalised features previously described as significantly different between stabilising and destabilising variants is plotted for these three Fn3-70 variants; as can be seen, the annotation profiles between the three are very different. For the wild-type Kidera Factors (red), R7W has extreme values for each feature, while Y27C has commensurately extreme values for the same features, in the opposite direction.

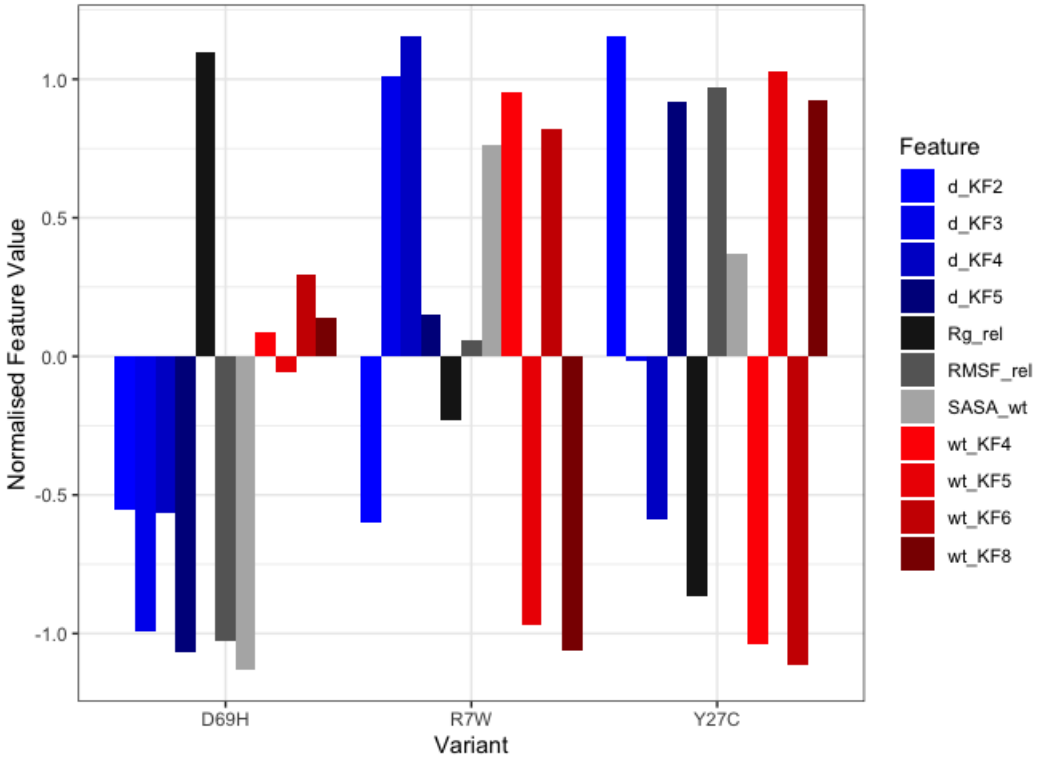


Figure 29: Fn3-70 variants with normalised features used for categorisation of thermostability variants

Comparison of normalised features relating to thermostability for three characterised Fn3-70 variants: D69H (moderately destabilising); R7W (strongly destabilising); Y27C (no change/weakly destabilising). Features were split into wild-type Kidera factors (reds; $KF4_{WT}$, $KF5_{WT}$, $KF6_{WT}$, $KF8_{WT}$), Δ Kidera factors (blues; $\Delta KF2$, $\Delta KF3$, $\Delta KF4$, $\Delta KF5$), and dynamical features (greyscale; Rg_{rel}, $RMSF_{rel}$, $SASA_{WT}$). Positive normalised feature values indicate those that were greater than the data set mean for that feature, while negative values indicate those that were lower than the mean.

By contrast, the values for D69H all occur between the two extremes for each feature. This reflects the opposing physico-chemical properties of arginine (R) and tyrosine (Y); where R is a charged, hydrophilic residue, Y is an aromatic, hydrophobic residue. Aspartate (D), by contrast, does not possess such extreme properties for the four features outlined. When the substitution is taken into consideration, the impact of the R7W substitution is intuitive; the hydrophilic R is replaced by hydrophobic tryptophan (W), a reversal of $KF4_{wt}$ giving rise to a high $\Delta KF4$.

This interplay between different features also serves to explain the difference in magnitude of destabilisation between R7W, D69H, and Y27C. By far the most buried residue position of the three is D69H, which has the lowest $SASA_{wt}$ (grey). This would seemingly run counter to the logic that the more buried a residue is, the more likely it is for variants at that position to destabilise the domain; however, the sequence properties of these features lend greater context to this observation. R7W, the most destabilising variant, is also the most solvent-exposed; the substitution of the native hydrophilic residue for a hydrophobic residue runs counter to the fold in which hydrophobic residues are buried. Y27C, which substitutes a hydrophobic residue for hydrophilic, is also an extreme substitution, but as it is also solvent-exposed, the introduction of a hydrophilic residue has little to no effect on the melting temperature. Finally, D69H also introduces a very hydrophilic histidine (H) residue, but as this position is almost entirely buried, the introduction of a hydrophilic residue is much less energetically favourable than it would be on the surface. Thus, the difference in magnitude of thermostability change for these three variants can, to some extent, be rationalised.

While this explanation is intuitive, it also illustrates the complexity of the relationships of different features to thermostability; the concept that the relationship between the magnitude of one feature and the thermostability annotation may reverse entirely depending on the value of another feature is difficult to represent with linear combinations of simple features. This is potentially another instance in which dynamical features could have application; as these are contingent on the physico-chemical and structural features defined in the starting topology, these complex relationships between static features may be represented in a single, dynamical feature. However, as can be seen in Figure 10, neither the Rg_{rel} nor the $RMSF_{rel}$ features reflect the relative differences in thermal stability; where the strongly destabilising R7W has smaller values, both D69H and Y27C have extreme values for both features.

To further investigate what information could be gleaned from these systems in MD simulation, the trajectories for each of the three replicates of these three variant simulations, as well as the three wild-type trajectories, were extended to 400 ns each, followed by a

PCA of the Cartesian coordinates of the backbone alpha carbon atoms superimposed atop the starting wild-type atomic configuration, in order to visualise the conformational space explored by each trajectory as projected onto the first two principal components; a plot of each trajectory as a separate window to aid visibility is shown in Figure 30. The trajectories move from the lighter to darker colours; as can be seen, almost all of the systems move from their starting configuration to explore a different region of conformational space, with the exception of the first Y27C replicate.

For the other trajectories, there are few observable differences in these plots that would mark them as being taken from a particular magnitude of stability change. All of the systems have at least one trajectory that moves from the positive to the negative values in principal component 1, which then explores either the conformational space in the centre or in the left-hand side of the plot. For principal component 2, almost all trajectories stay in the central region between -5 and +5; the main exception is the first replicate of the wild-type system, which moves towards exploration of the conformational space designated by more positive values of PC2. In addition, less than 30% of the variation in the data set is retained in PC1/2; it is unlikely, therefore, that any signal exists in the backbone conformation at such a low granularity to reliably provide information on thermal stability.

3.6 Discussion

3.6.1 Data from MD simulations augment sequence and structural annotation of variants

The time evolution of an atomistic system by simulation *in silico* may be likened to attempting to solve the question of how the sequence, structure, and myriad interactions of a given protein resolve into a constantly moving, changing, dynamic system. For a structural representation of a protein, the folded three-dimensional conformation taken in nature is a consequence of the properties and positions of all the atoms that are incorporated in the protein sequence;

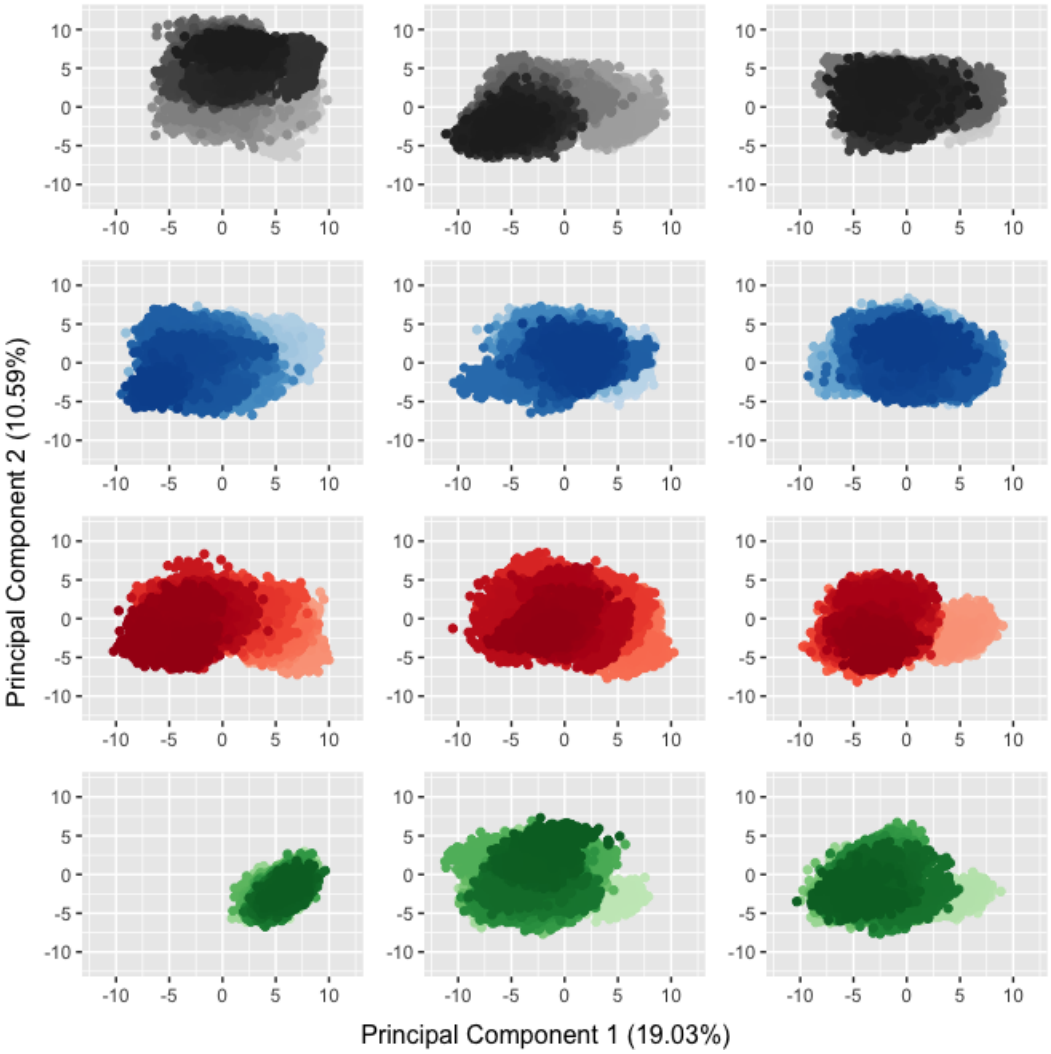


Figure 30: PCA of backbone coordinates of Fn3-70 variants versus WT

PC1/PC2 plot of 3 wild-type and 9 mutant trajectories of Fn3-70, split across 12 plots. Each point in the plots represents 1 frame in each 400ns trajectory (40001 frames), where the Cartesian coordinates of all backbone alpha carbon atoms in the atomic configuration in that frame are projected onto the first two principal components, thus describing the position of that configuration within conformational space. Represented are three replicates each of wild-type (black), variant D69H (blue, moderately destabilising), variant R7W (red, strongly destabilising), variant Y27C (green, no change/weakly destabilising). The colour of each trajectory becomes darker as it progresses (i.e. light colour = beginning of simulation, dark colour = end of simulation).

by categorising protein residues by these properties, it becomes possible to understand the general contribution of a single residue to the stability of the overall fold, and the likely contribution that another, physiologically different residue would make if substituted at that position. For example, a buried hydrophilic residue substituted in place of a hydrophobic residue would be expected to perturb the core of the protein; a wide range of such features are used by methods to predict structural damage (e.g. Missense3D[236]). However, to generalise the effect of mutation at the structural level risks eliminating potentially vital information; by explicit definition of the atoms and interactions involved, the mutant system may be more precisely described, and thus the overall effects of the rearrangement of inter-atomic contacts and forces simulated.

This explicit computation of the higher-level impacts of changes in the lower-level features that define them has become increasingly well studied in bioinformatics, most notably in the formulation of protein structure from sequence data; AlphaFold2 is the most outstanding example at the time of writing. Here, with the use of enormous sequence and structure data sets, and innovative deep learning architectures, a reliable mapping between a given protein sequence and a physiologically plausible, three-dimensional protein fold has been established. While this demonstrates that sequence data can implicitly encode the structure of the protein in question, it also emphasises the sheer complexity of the mapping from one to the other; similarly, the mapping of three-dimensional structural data to four-dimensional, temporal data is too complex for sequence or structural data to capture implicit dynamical properties without the use of prohibitively large data sets. Thus, MD simulations allow us to utilise this information for variant impact assessment, through mapping of structural features to dynamical properties.

In this chapter, I have investigated the use of data derived from MD simulations to describe the differences between wild-type and mutant titin domain systems. A substantial alteration in the sequence of the protein at a structurally important location would be expected to alter the energy landscape of the mutant system, changing the conformational space explored by the protein over the course of the simulation. A number of existing methods

have been employed to derive features which capture these changes; in the first investigation, where 31 single-domain variant titin systems reported in the literature as associated with various conditions were simulated and compared to 30 likely non-disruptive variants from population databases, the relative RMSF change between wild-type and mutant systems was significantly different between the literature and population variant data sets, while also capturing information that supplements the distinction between these groups made by sequence or structural features.

3.6.2 Variants in Ig-169 do not perturb the system *in silico* compared to wild-type

Not all of the literature variants simulated showed a marked difference in dynamics from the wild-type systems; of the 31 variants in the data set labelled as pathogenic, 10 variant trajectories did not demonstrate a clear separation from their respective wild-type trajectories in conformational space when projected onto their first and second principal components. This is to be expected, as not all disease pathomechanisms may be ascribed to the perturbation of the protein backbone, and as such would not be visible in the trajectory. Likewise, a neutral variant may perturb the backbone of a protein without any disease association, depending on the degree of structural damage and the tolerance of that domain to mutation. Even so, these exceptions to the general trend of damaging variants increasing fluctuation at the site of mutation are worthy of comment.

Variants in Ig-169 are long understood to be related to tibial muscular dystrophy; despite strong evidence in support of their pathogenicity, the variants L35956P, H35946P, T35915P, W35930R, and I35947N show dynamics very similar to the wild-type, and little difference from neutral variants in the same domain, within the simulation time performed. As previously mentioned, a proposed mechanism for the onset of TMD is the impact of the variant on the binding between Ig-169 and the Ig-1 domain of *OBSCN*; this offers the possibility that these variants show little difference in single-domain simulations owing to the specific pathomechanisms involved, which involve more biomolecular components. However, this

is in opposition to experimental evidence, which has characterised the autosomal dominant variants L35956P and H35946P as destabilising to the parent domain. L35956P is an especially peculiar case, as this is the most severely damaging missense variant known in the Ig-169 domain, characterised *in vitro* as producing severe misfolding and lowered expression of protein, in addition to abolition of *OBSCN*-Ig1 binding. In this analysis, however, it has the lowest variation in its mutant trajectories of any literature variant, as well as the lowest separation distance between wild-type and mutant trajectories. Notably, it has less effect on the dynamics of the Ig-169 domain, in terms of dynamical features calculated, than the three recessive variants T35915P, W35930R, and I35947N.

Even over 6 replica trajectories, there is little difference in the conformational space explored by the wild-type Ig-169 and any of its mutants *in silico*; thus, one possibility is that Ig-169 is more dynamically rigid than other domains, and would require more simulation time than other systems for meaningful differences to be observed. Another consideration is that these Ig-169 variants are associated with late-onset TMD; as patients only developed symptoms later in life, it may be inferred that there are other physiological considerations in play. It should also be observed that the introduction of proline in L35956P, T35915P, and H35946P would be expected to decrease the flexibility of the protein chain; thus, it would also follow that the backbones of these systems would be less mobile than wild-type domains *in silico*. As this is one of the key hotspot domains for titin-related pathology, this emphasises that dynamical features are not universally applicable; it is imperative that mutually informative features are found that can parse the differences between these different titin systems.

The sole reported damaging variant in Ig-19, T2896I, shares the issue of Ig-169 variants in that it has been experimentally characterised as damaging yet shows little difference in its dynamics *in silico* compared to wild-type. Like Ig-169, its proposed pathomechanisms involve interactions, here with respect to the chain dynamics at the domain interface between Ig-18 and Ig-19; the substitution at this position is thought to modify the chain dynamics in this region of the I-band. Indeed, this variant has been characterised previously by MD

simulation of the Ig18-Ig20 tandem [100], which indicated a difference between wild-type and mutant trajectories; a subsequent paper[114] further characterised all possible substitutions at this position, noting the increased inter-domain movement in the mutant system, and proposing the viscoelasticity of the titin chain as a feature of the disease mechanism. This suggests that, especially for interface residues in the elastic regions of titin, consideration of the chain dynamics of the protein may be necessary to observe the consequences of the variant fully.

3.6.3 Primitive dynamical features do not recapitulate domain thermostability alone

With the encouraging examples of positively identified examples of damaging titin variants, I decided to look more closely at a data set of variants that were more likely to be differentiated by their dynamical properties. As the melting temperature of the protein is associated with its thermal stability, I hypothesised that information contained in the dynamics of the domain, and the flexibility of the backbone, could be informative as to the tolerance of the protein fold to an increase of thermal energy in the system. However, this did not seem to be the case; of all of the dynamical features tested, only the relative radius of gyration showed a significant difference between stabilising and destabilising variants, and its application in a case study of Fn3-70 variants did not add information pertinent to the thermostability of those mutant domains.

The use of MD simulations and dynamical data to inform predictions of thermostability is not new; several authors have investigated the relationship of particular MD-derived features with experimentally determined thermostability measurements. Kleinjung *et al.* (2003)[237] demonstrated that conformational features could be used to recapitulate the thermal unfolding of apo-calmodulin domains using leap-dynamics methods; Zeiske *et al.* (2016)[238] reported that the temperature dependence (Λ) of the order parameter for backbone N-H bonds was correlated with *in vitro* thermostability measurements in four bacterial homologues of ribonuclease HI; Bekker *et al.* (2019)[239] likewise reported that the fraction of native

atomic contacts (Q) in simulations at 400 K correlated well with experimental T_m values in seven single-domain antibodies - which, given their similarity to the immunoglobulin and fibronectin domains of titin, makes this result particularly encouraging with respect to these systems. These methods were not considered in this analysis as they are typically more computationally intensive, involving multiple simulations of different temperatures, and are less amenable to large-scale saturation mutagenesis.

Consequently, while there remains potential for MD-related features to improve the prediction of experimental values such as T_m, it will require more sophisticated feature engineering beyond the standard characterisation of *in silico* trajectories, while simultaneously remaining sufficiently simple to allow for characterisation of as many variants as possible without sacrificing prediction accuracy. This would likely involve domain-specific exploration of features, rather than considering the protein as a whole, as the signal across whole-protein features is diminished for both the literature/population and the thermostability data sets here. It will also require more data points for the different categories of interest, sufficient to populate equal numbers of independent data in the different categories.

3.6.4 Future considerations relating to MD simulation-derived features

Despite the amenability of these systems to MD simulation, owing to their small size and the availability of reliable structural coverage, I have also encountered difficulties that must be addressed here. Both the Fn3 and Ig domain systems are small, beta-sandwich fold domain proteins; their initial structure is quite robust to the small chemical changes represented by missense variants, such that catastrophic perturbation of the domain tertiary structure in simulation is not observed. Thus, any insights about the impacts of mutation on the dynamics of the protein are derived from relatively small fluctuations, which must be distinguished from the fluctuations that are present in the native state. This means that the simulations undertaken must be both reliable and reproducible, and that they are sampling a sufficient quantity of the conformational ensemble of the system across the trajectory.

In order to improve the reliability of the simulations run, at least 3 simulations of each system were undertaken in all experiments, with a separate energy minimisation and equilibration protocol in each replicate, and thus slightly different starting coordinates. Encouragingly, in most systems a difference between trajectories of wild-type and mutant systems was seen to a greater degree than within categories, suggesting that these trajectories of the same system were sampling from similar distributions. However, performing more replica simulations decreased the length of time that each system could be simulated for, thus decreasing the sampling of conformational space across a given trajectory. I also found that the use of six replicates did not change the overall difference between wild-type and mutant systems compared to three trajectories, hence three replicates of 200ns were used in the investigation of the thermostability data set, as opposed to the six replicates of 100ns for the literature/population data set.

The sampling problem in MD simulation is well documented[205, 240]; no amount of simulation time will guarantee a full accounting of the possible energy landscape for a system, as the simulation proceeds in a stepwise fashion and must pass through intermediates to arrive at another possible native configuration. Thus, it is necessary to ensure that each simulation runs for sufficient time to capture a representative quantity of conformational space; however, this must be balanced against the time taken to generate each individual trajectory. One of the aims of this work was to determine if sufficient information could be captured in a short simulation, which in turn would allow for characterisation of large numbers of variants in a reasonable time frame; with the computational resources used in this work, a 100 ns trajectory would typically be generated in 24 hours. Thus, while characterisation of all variants in this data set, with simulations of 200 ns and multiple replicates, is possible with these resources, a larger-scale saturation mutagenesis of titin domains would not be feasible; enhanced sampling methodologies such as Replica Exchange MD (REMD)[241] and Gaussian Accelerated MD (GAMD)[242] could be used in future work to improve sampling within these smaller time frames.

Another issue that is pertinent to the analysis of data from trajectories is that of the dimensionality reduction approaches used to visualise these high-dimensional data sets. As previously mentioned, PCA is a standard method used to reduce the dimensionality of a data set while preserving the maximum variance present in the original; however, it is also a linear dimensionality reduction technique, meaning that non-linear relationships within the original data set will be poorly represented in the resultant low-dimensional data set. Many relationships within the motions of the alpha-carbon backbone of the domain, including those reflecting conformational change, are intrinsically non-linear, and thus better preserved through non-linear dimensionality reduction methodologies. Two non-linear techniques commonly applied for dimensionality reduction of MD simulations of biological macromolecules are t-distributed stochastic neighbour embedding (t-SNE)[243] and uniform manifold approximation and projection (UMAP)[244]; these typically produce better results than linear techniques such as PCA, but for a corresponding increase in the computational time required[245]. Future investigation of these trajectory data should consider the implementation of these methods to determine what additional insights may be gleaned from our simulations.

Finally, it must be stated that many of the dynamical features in this analysis are operating on very small quantities and within an equally small range of values; care must be taken here, as the calculated features are subject to the error within the force field used, which is not a perfect representation of real-world forces. While it is not possible to increase the accuracy of the simulation without substantial increases in computing time, the use of replica trajectories and average values for each system should minimise the risk of incorrect assessment.

3.7 Conclusion

In summary, I have demonstrated the robustness of these simulations to describe the motions of these systems across multiple replicates, and the results indicate in many cases that mutant systems move to occupy different energy minima to their corresponding wild-type

systems, through the differences in conformational ensembles generated over the course of the simulations. The features that are generated from these motions, particularly those that sample from the site of mutation itself (RMSF, SASA), can be shown to discriminate between reported pathogenic and common population variants, and augment the discriminatory power of standard sequence-based and evolutionary features often used to separate pathogenic from non-pathogenic variants. However, these same features show little correlation with the experimentally determined melting temperature of variant domains, or the difference between these domains and their respective wild-type domains, despite their ostensible link to domain stability. In combination with physico-chemical annotations, dynamical features can aid in distinguishing thermally stabilising from destabilising variants, but the task of differentiating destabilising variants by severity remains a challenge.

Table 5: Single amino acid variants in titin reported in the literature

Variant	Domain	Template / Structure	Domain Position	Database	gnomAD MAF
A82D	Ig-1	2A38	77	Literature	NA
A178D	Ig-2	2A38	75	Literature	NA
S1295L	Ig-5	1TLK	5	Population	9.60E-01
T2896I	Ig-19	5JDD	15	Literature	NA
V3261M	Ig-23	2A38	23	Population	8.44E-01
C5054R	Ig-33	2RIK	74	Literature	NA
D7462H	Ig-59	3B43	44	Population	1.77E-01
E8144A	Ig-66	2RIK	67	Population	1.75E-01
S8425N	Ig-69	3B43	66	Population	1.76E-01
N8803S	Ig-73	3B43	68	Population	1.78E-01
A13715E	Ig-87	3UTO	32	Literature	NA
N16133K	Fn3-4	3LPW	79	Literature	NA
A16466T	Fn3-7	4O00	17	Population	4.07E-06
W16471C	Fn3-7	4O00	22	Literature	3.88E-04
C17051R	Fn3-12	3LPW	5	Literature	4.03E-06
L18237P	Fn3-20	5FM4	72	Literature	6.43E-04
V22193I	Fn3-49	4O00	38	Population	8.10E-06
V22232E	Fn3-49	4O00	77	Literature	NA
V22359I	Fn3-50	3LPW	6	Population	2.40E-01
A22416P	Fn3-50	3LPW	63	Population	9.99E-01
R24947C	Fn3-69	4O00	40	Population	1.71E-01
R25480P	Fn3-73	4O00	76	Literature	NA
I27775V	Fn3-90	3LPW	10	Population	3.52E-01
R27839Q	Fn3-90	3LPW	74	Population	4.30E-04
G27849V	Fn3-90	3LPW	84	Literature	NA
I29396T	Fn3-102	4O00	62	Population	3.51E-01
P31709H	Fn3-119	3LPW	2	Literature	NA
P31709R	Fn3-119	3LPW	2	Literature	NA
C31712R	Fn3-119	3LPW	5	Literature	4.11E-06
C31712Y	Fn3-119	3LPW	5	Literature	NA

T31726A	Fn3-119	3LPW	19	Population	4.03E-06
W31729C	Fn3-119	3LPW	22	Literature	NA
W31729L	Fn3-119	3LPW	22	Literature	NA
W31729R	Fn3-119	3LPW	22	Literature	NA
P31732L	Fn3-119	3LPW	25	Literature	1.21E-05
V31745M	Fn3-119	3LPW	38	Population	3.57E-05
R31748C	Fn3-119	3LPW	41	Population	9.90E-04
R31748H	Fn3-119	3LPW	41	Population	1.21E-05
R31748S	Fn3-119	3LPW	41	Population	4.02E-06
S31766F	Fn3-119	3LPW	59	Population	3.39E-03
R31781L	Fn3-119	3LPW	74	Population	8.04E-06
R31781Q	Fn3-119	3LPW	74	Population	1.61E-05
A31784V	Fn3-119	3LPW	77	Literature	NA
V31785I	Fn3-119	3LPW	78	Population	4.02E-06
N31786K	Fn3-119	3LPW	79	Literature	NA
G31791D	Fn3-119	3LPW	84	Literature	NA
G31791R	Fn3-119	3LPW	84	Literature	NA
G31791V	Fn3-119	3LPW	84	Literature	NA
V31794L	Fn3-119	3LPW	87	Population	3.18E-05
I31799V	Fn3-119	3LPW	92	Population	6.44E-05
R31847P	Fn3-120	5FM4	41	Literature	NA
E35910K	Ig-169	3KNB	14	Population	1.21E-05
T35915P	Ig-169	3KNB	19	Literature	NA
A35917V	Ig-169	3KNB	21	Population	4.03E-06
W35930R	Ig-169	3KNB	34	Literature	NA
H35946P	Ig-169	3KNB	50	Literature	NA
I35947N	Ig-169	3KNB	51	Literature	4.01E-06
I35947T	Ig-169	3KNB	51	Population	4.01E-06
T35955N	Ig-169	3KNB	59	Population	1.20E-05
L35956P	Ig-169	3KNB	60	Literature	NA
I35986T	Ig-169	3KNB	90	Population	4.03E-06

Table 6: Experimentally characterised variant systems and their reported melting temperatures

Variant	Variant (IC)	Domain	$Tm_{mut}(^{\circ}C)$	$Tm_{WT}(^{\circ}C)$	$\Delta Tm(^{\circ}C)$
S86Y	S91Y	Ig-1	73.75	70.87	2.88
L37I	L17506I	Fn3-15	61.51	74.35	-12.84
H46R	H17515R	Fn3-15	62.13	74.35	-12.22
N43K	N17512K	Fn3-15	72.64	74.35	-1.71
N61S	N17530S	Fn3-15	73.07	74.35	-1.28
N16S	N19198S	Fn3-27	57.93	59.87	-1.94
Y37F	Y19219F	Fn3-27	57.13	59.87	-2.74
R46H	R19228H	Fn3-27	61.59	59.87	1.72
Q68P	Q19250P	Fn3-27	63.49	59.87	3.62
V88L	V19270L	Fn3-27	55.49	59.87	-4.38
L94F	L19276F	Fn3-27	56.43	59.87	-3.44
D26G	D20002G	Fn3-33	53.77	67.11	-13.34
E43K	E20019K	Fn3-33	66.05	67.11	-1.06
P90L	P20066L	Fn3-33	53.97	67.11	-13.14
P25T	P21683T	Fn3-45	60.37	77.94	-17.57
I34V	I21692V	Fn3-45	75.89	77.94	-2.05
S58L	S21716L	Fn3-45	73.64	77.94	-4.3
F74L	F21732L	Fn3-45	65.75	77.94	-12.19
Y93F	Y21751F	Fn3-45	76.87	77.94	-1.07
V38I	V22193I	Fn3-49	53.83	58.07	-4.24
R50K	R22205K	Fn3-49	59.37	58.07	1.30
D70N	D22225N	Fn3-49	57.61	58.07	-0.46
V80I	V22235I	Fn3-49	55.67	58.07	-2.40
R32W	R22482W	Fn3-51	64.29	73.77	-9.48
K46R	K22496R	Fn3-51	71.69	73.77	-2.08
R49Q	R22499Q	Fn3-51	62.14	73.77	-11.63
M51I	M22501I	Fn3-51	63.16	73.77	-10.61
S85N	S22535N	Fn3-51	67.67	73.77	-6.10
R42Q	R23767Q	Fn3-60	48.99	58.36	-9.37
R65C	R23790C	Fn3-60	58.67	58.36	0.31
R65H	R23790H	Fn3-60	59.12	58.36	0.76

V1L	V23826L	Fn3-61	54.00	55.60	-1.60
A78G	A23903G	Fn3-61	45.12	55.60	-10.48
I77T	I23902T	Fn3-61	51.70	55.60	-3.90
R7W	R25012W	Fn3-70	46.71	74.04	-27.33
Y27C	Y25032C	Fn3-70	74.03	74.04	-0.01
D69H	D25074H	Fn3-70	61.64	74.04	-12.40
T1M	T25601M	Fn3-74	54.21	56.31	-2.10
D63N	D25663N	Fn3-74	54.46	56.31	-1.85
I34T	I26023T	Fn3-77	57.65	61.58	-3.93
T49R	T26038R	Fn3-77	58.47	61.58	-3.11
S45G	S26330G	Fn3-79	52.17	50.8	1.37
A53S	A26338S	Fn3-79	50.52	50.8	-0.28
T52A	T26538A	Fn3-81	52.32	61.29	-8.97
R76Q	R26562Q	Fn3-81	51.49	61.29	-9.80
N81S	N26567S	Fn3-81	56.33	61.29	-4.96
G29V	G28280V	Fn3-94	53.69	62.21	-8.52
R40C	R28291C	Fn3-94	53.99	62.21	-8.22
R47Q	R28298Q	Fn3-94	55.98	62.21	-6.23
Q57P	Q28308P	Fn3-94	53.87	62.21	-8.34
R71C	R28322C	Fn3-94	58.64	62.21	-3.57
R71H	R28322H	Fn3-94	58.05	62.21	-4.16
A16T	A30239T	Fn3-108	54.18	55.39	-1.21
C53G	C30276G	Fn3-108	60.65	55.39	5.26
G6R	G30724R	Fn3-112	43.59	57.69	-14.10
R15H	R32513H	Fn3-125	55.36	57.36	-2.00
R40H	R32538H	Fn3-125	54.06	57.36	-3.30
A45T	A32543T	Fn3-125	55.59	57.36	-1.77
R60Q	R32558Q	Fn3-125	58.85	57.36	1.49
R89H	R32587H	Fn3-125	61.46	57.36	4.10
R89P	R32587P	Fn3-125	58.58	57.36	1.22
G29R	G33420R	Fn3-131	56.80	61.89	-5.09
R6K	R33694K	Fn3-132	53.94	57.23	-3.29
R50C	R33738C	Fn3-132	52.38	57.23	-4.85
R50H	R33738H	Fn3-132	52.81	57.23	-4.42

V61M	V33749M	Fn3-132	52.54	57.23	-4.69

4 Prediction of Variant Impacts of Titin SAVs using a combination of Sequence, Structural, and Dynamical Features

4.1 Synopsis

- In this chapter, I bring together the annotation and feature generation from all previous chapters to build classifiers for prediction of properties of titin variants relevant to clinical applications.
- I demonstrate that a reliable binary classifier can be constructed from simple features that gives a strong indication of an input titin variant's potential for pathogenicity; the use of deep mutational scanning data as a feature strongly improves prediction accuracy.
- I highlight the difficulty of predicting the thermostability of mutant proteins with the given feature set, and how the field might begin to address the discrepancies in prediction of pathogenicity and thermostability in proteins.

4.2 Introduction

With the rise of high-throughput methods of genome sequencing and causal variant discovery pipelines, the volume of variant data being produced and reported is ever increasing; thus, the importance of accurate, reliable annotation and characterisation of both novel and existing variants is increasing commensurately. Beyond the traditional, if simplistic ontological approach to variants by classification into disease-causing or neutral, there is a growing understanding of more complex associations between disease phenotypes and their associated variants, from such concepts as expressivity and penetrance of individual variants, to the potential combined effects of multiple mutations inherited compound heterozygous in *cis* or in *trans*.

In Chapter 1, I described and discussed the numerous methods in which reported disease-associated variants may be validated through the experimental determination of biophysical consequences; these range from the recapitulation of the disease phenotype in model organisms, to the assessment of features at the cellular or protein level, such as changes in stability, binding affinity, solubility, expression, and sub-cellular localisation of the mutant protein compared to the wild-type. These methods allow for comparison between different variants within the same protein concerning the magnitude of the variant impact; however, the more robust the validation method, typically the more time-consuming and expensive the method is. This limits the number of variants that can practically be characterised experimentally; thus, as discussed in Chapter 2, bioinformatics pipelines and tools are used to prioritise those variants from high-throughput data sets most likely to be related to the phenotype of interest, for the purpose of experimental validation.

State-of-the-art computational prediction tools have advanced in step with the quantity of data available for training, both for binary classifiers that aim to distinguish pathogenic from non-pathogenic variants, and for predictors that attempt to reproduce the measured properties of variant proteins as determined experimentally, such as changes to free energy or melting temperature of the protein. Machine learning approaches allow for complex relationships between features to be learned by parameter optimisation, with different architectures allowing for novel solutions to different problems. In addition, the increased availability of high-performance computing (HPC) resources has allowed for an increase in the complexity of calculated features; this includes the use of atomistic molecular dynamics (MD), as demonstrated in Chapter 3, to characterise whole variant domains over longer timescales. However, these methodologies remain secondary to experimental methods for variant characterisation, and especially for more complex properties, there is room for improvement in the accuracy of their predictions.

It has often been observed that when computational predictors of variant impacts are applied to novel data sets of mutations, particularly when those changes reflect multiple smaller differences in the same protein of interest, the performance is considerably worse

than reported benchmarking would suggest; a review by Marabotti *et al.* (2021)[246] of such comparisons for predictors of thermodynamic stability gives a good overview of these issues. For melting temperature specifically, Huang *et al.* (2020)[247] demonstrated that the predictions of standard computational predictors of domain stability change were poorly correlated with actual experimental measurements of melting temperature, using an in-house database comprised of melting temperature measurements for variants in β-glucosidase B (BglB), although some were able to give a reasonable binary prediction for soluble expression of the system *in vitro*. This is despite the good correlation of these data with experimental values in benchmarking studies. The authors suggest that over-representation of common protein families in curated training data, as well as for variants representing substantial changes in protein stability at the expense of more modest changes, which are considerably more common in single-residue substitutions.

This weakness in current computational methodologies highlights the necessity of improving general-purpose classifiers for novel variants, but also indicates a role for predictors of variant impacts in subsets of related proteins, or even single proteins, should there be sufficient information available on existing variants to inform the characterisation of others. Titin, the giant sarcomeric protein, is perhaps the clearest and most powerful example of such a protein - with 169 immunoglobulin I-set domains and 132 fibronectin type-III domains of identical folds, some of which are characterised extensively and some known only by homology, there is both a large quantity of information to learn from and a much larger quantity of data to apply it to. Moreover, with so many possible variants in titin, it would be impractical to separately characterise all of them *in vitro*; consequently, having a method to reliably characterise these variants *in silico*, in a fraction of the time required, would be of great benefit.

In Chapter 1, I undertook the collection and filtering of reported pathogenic missense variants in the scientific literature, producing a list of known variants with varying capacity for causing damage to health when inherited. In Chapter 2, I proceeded to annotate the known titin variants present in the general population and reported pathogenic variants with existing,

general computational predictors of variant impacts, summarising the characterisation of known titin variants with currently available, general bioinformatics tools. In Chapter 3, I generated new information on selected variants using atomistic molecular dynamics (MD) simulations of single domains, and demonstrated the application of MD-derived features for distinguishing between variants from these different source distributions. Thus, in this final Chapter I aim to synthesise these data together, to produce titin-specific classification architectures that can reliably reproduce those differences I have described.

4.3 Methods

4.3.1 Data Sets

As described in Chapter 3, I produced a set of features based on atomistic MD for variants found in pathogenic variants sourced from the literature, non-pathogenic variants sourced from population databases, and variants characterised in-house for altering the thermostability of the parent domain. From these distributions two non-overlapping data sets were derived, one of which contained variants annotated with either "pathogenic" or "non-pathogenic", and the other which contained variants annotated with both their absolute melting temperature (T_m) as calculated by differential scanning fluorimetry (DSF), and the change in T_m relative to the wild-type domain.

The pathogenic + non-pathogenic training data set (hereafter "pathogenicity") consisted of 61 variants, 31 of which were labelled "pathogenic" and 30 of which were labelled "non-pathogenic" (see Table 5). Because the size of the data set was relatively small, rather than initial separation into training and test data, classifier performance was calculated by use of leave-one-out cross-validation. For each individual variant in the data set, the data set of n variants was split into a test set (that data point alone, 1 variant) and a training set (all other data, $n-1$ variants), and a single prediction was computed; by iterating over all variants, the accuracy and other prediction metrics could be computed. Leave-one-out CV is generally

considered a robust measure of classifier prediction error where there is insufficient data for separate training, cross-validation, and test sets, and the size of the training data set is small enough to make it computationally feasible[248].

The training data set of domain melting temperatures (hereafter "thermostability") consisted of 65 variants, each labelled with a wild-type T_m, mutant T_m, and ΔT_m in degrees Celsius (see Table 6; this was reduced to 64 variants after exclusion of A23903G, which had issues with reading of its MD simulation features. As in Chapter 3, these were further categorised into "Weakly Stabilising" ($\Delta T_m \geq 0°C$, $n = 11$), "Weakly Destabilising" ($0°C > \Delta T_m \geq -5°C$, $n = 33$), "Moderately Destabilising" ($-5°C > \Delta T_m \geq -17°C$, $n = 19$), and "Strongly Destabilising" ($\Delta T_m \leq -17°C$, $n = 2$) for classification. Because of the imbalance between the different categories, however, I elected to re-assess the performance of the classifier as a binary problem, discriminating between neutral ("Weakly Stabilising" + "Weakly Destabilising", $n = 44$) and disrupting ("Moderately Destabilising" + "Strongly Destabilising", $n = 21$). To balance class sizes during model training, the Python package $imbalanced-learn$[249] was used to perform naive random over-sampling with replacement of the smaller "disrupting" class (see Machine Learning Architectures).

To further assess the abilities of these classifiers, a test set of variants with both reported pathogenicity/non-pathogenicity in the literature and reported measures of *in vitro* experimental characterisation was constructed. This gave us a data set of 20 variants, which were delineated into "damaging" ($n = 15$) or "not damaging" ($n = 5$) categories based on the experimentally determined T_m or solubility (see Table 7). As the Ig-169 variant I35947N is considered pathogenic but does not destabilise the domain, the test sets were split into two, with one in which I35947N is considered "damaging" ($n = 15$ (damaging); $n = 5$ (not damaging)) and one where it is considered "not damaging" ($n = 14$ (damaging); $n = 6$ (not damaging)). The former was then used to test classifiers trained on the pathogenicity data, and the latter to test classifiers trained on thermostability data. To preserve the split between training and test data, the pathogenicity classifiers were re-trained with the same feature sets on the remaining 41 variants in the training set. As the test set was enriched for damaging

Source	Variant	Method	WT	Mut	Pathogenic	Destabilising
[35]	A82D	DSF	71.2°*C*	34.0°*C*	1	1
[100]	T2896I	DSF	62.3°*C*	51.4°*C*	1	1
[35]	W16471C	DSF	75.2°*C*	43.4°*C*	1	1
[35]	L18237P	CD	80.4°*C*	63.5°*C*	1	1
[35]	V22232E	DSF	59.2°*C*	Unfolded	1	1
[35]	G27849V	DSF	61.5°*C*	Unfolded	1	1
[35]	R24947C	DSF	62.6°*C*	62.1°*C*	0	0
[35]	I27775V	DSF	61.5°*C*	63.5°*C*	0	0
[130]	P31709R	PAGE	Soluble	Insoluble	1	1
[130]	C31712R	PAGE	Soluble	Insoluble	1	1
[130]	W31729C	PAGE	Soluble	Insoluble	1	1
[130]	W31729R	PAGE	Soluble	Insoluble	1	1
[35]	P31732L	DSF	49.2°*C*	31.8°*C*	1	1
[130]	R31748C	PAGE	Soluble	Soluble	0	0
[130]	S31766F	PAGE	Soluble	Soluble	0	0
[130]	V31785I	PAGE	Soluble	Soluble	0	0
[35]	R31847P	DSF	51.7°*C*	Unfolded	1	1
[95]	H35946P	CD	58°*C*	37°*C*	1	1
[95]	I35947N	CD	58°*C*	53°*C*	1	0
[95]	L35956P	CD	58°*C*	Unfolded	1	1

Table 7: Variants in combined Test Set

Variants in the combined test set used to compare the performance of thermostability and pathogenicity classifiers on the same test data set. Experimental characterisations were taken from primary literature. Variants were labelled "1" if damaging or "0" if not damaging, for both of the pathogenicity and thermostability classes. Different methods were used for experimental characterisation: DSF = *Differential Scanning Fluorimetry (thermostability, °C);* CD = *Circular Dichroism (thermostability, °C);* PAGE = *Polyacrylamide Gel Electrophoresis (solubility).*

variants, classes in the reduced pathogenicity data set were again balanced using naive random over-sampling.

4.3.2 Machine Learning Architectures

Initial binary classification tasks were performed using an implementation of the Random Forest (RF) architecture[250]. RF algorithms consist of an ensemble of uncorrelated decision trees (minimising bias), where each individual decision tree is constructed by random sampling from the training data set and random selection of a subset of features. These

decision trees then split the training data into separate categories at nodes according to the features that produce resultant groups where the individuals within one group are as similar as possible to one another, and as different as possible to those outside of the group. The key advantage of RF architectures over single decision trees is that the ensemble of predictions will typically avoid errors made by individual trees (minimising variance), and thus increases the stability of prediction (likelihood of a correct prediction of any magnitude being made).

Feature importances given by the RF classifiers reflect the Gini impurity of these features[251], a measure of the quality of the data split at nodes across the ensemble of decision trees:

$$Gini(m) = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{21}$$

The Gini impurity, or Gini index, at node m is calculated from the proportion of class k observations in node m (p_{mk}), across all classes K. This results in an impurity value between 0 and 0.5, where 0 represents all observations in a single class, and 0.5 represents an equal split between all classes. For example, for a data set of 10 observations:

$$Gini(m) = \begin{cases} P_0 = 0; P_1 = 1 & 0*(1-0)+1*(1-1) = 0.00 \\ P_0 = 0.2; P_1 = 0.8 & 0.2*(1-0.2)+0.8*(1-0.8) = 0.32 \\ P_0 = 0.5; P_1 = 0.5 & 0.5*(1-0.5)+0.5*(1-0.5) = 0.50 \end{cases} \tag{22}$$

Thus, a given tree is grown by selecting those features that minimise the Gini impurity at each node; this is then used to compute feature importance by summing the nodes where that feature was selected, relative to the number of samples split.

Alternative supervised learning architectures were tested to compare against the performance of the RF architectures being used. Logistic Regression (LR)[252] represents binary classification in the form of a logistic function, where the function outputs a value between 0 and 1 representing the predicted probability of the test condition being true. The log-odds of the test condition being true are a linear combination of the input features, while the logistic function converts the log-odds to the probability, and a classification is made based on a cut-off value of 0.5. Parameters, or feature coefficients, in the classifier are optimised to minimise the cost function for the mapping of input features to output class; these reflect the weighting on specific features, and the expected change in the output for a given input. Support Vector Machines (SVM)[253] are algorithms similar to LR which use kernels to classify input data directly, rather than individual features; new features are computed based on the proximity of data points to landmark points in the data space using a similarity function.

All machine learning architectures were implemented in Python 3 using the $scikit-learn$ library[254]. RF classifiers were created using $sklearn.ensemble.RandomForestClassifier()$; LR classifiers were created using $sklearn.linear_model.LogisticRegression()$; SVM classifiers were created using $sklearn.svm.SVC()$. Naive random over-sampling with replacement during the training step for all architectures was performed with the $imbalanced-learn$ library[249], using $imblearn.over_sampling.RandomOverSampler()$.

The performance of all models was assessed using two metrics of accuracy, the F1 score and the Matthews correlation coefficient (MCC):

$$F1 = \frac{TP}{TP + 0.5 * (FP + FN)} \tag{23}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TN + FP) * (TP * FP) * (TN + FN)}} \tag{24}$$

Both methods are calculated from the confusion matrix of predictions for a given model: TP represents the sum of true positive predictions (predicted true, actual true); TN represents true negatives (predicted false, actual false); FP represents false positives (predicted true, actual false); FN represents false negatives (predicted false, actual true). The F1 score summarises both precision and recall for a prediction algorithm; the MCC, which is generally considered more informative than the F1 score or simple accuracy in binary classification[255], is also weighted by the accuracy of the model in predicting negative values.

4.3.3 Hyperparameter Tuning

Each of the machine learning architectures learns appropriate parameters for mapping between training and test set data; "hyperparameters", which affect the construction of the model itself, must be set and optimised prior to the training step. To search for the appropriate hyperparameters for each model, the *sklearn.model_selection* library was used; first, to narrow down the possible combinations of hyperparameters, $RandomisedSearchCV()$ was used to perform multiple runs of cross-validation on candidate models. 100 untrained candidate models were selected at random from all possible combinations of provided options for hyperparameters, each of which was evaluated using five-fold cross-validation (default recommended by $scikit-learn$), from which the model with the highest evaluation score was reported. Subsequently, hyperparameter options that were never selected were removed, and $GridSearchCV()$ was used to evaluate all possible combinations of the remaining hyperparameter options and report the best model, again with three runs of five-fold cross-validation.

RF classifiers were optimised for 6 hyperparameters:

- Number of trees in the random forest (*n_estimators*) - 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000

- Maximum number of features considered when looking for the optimal split of the data ($max_features$) - 'None' (no maximum), 'sqrt' (square root of total number of features)
- Maximum depth (number of levels) for each tree (max_depth) - 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110
- Minimum number of samples required to split an internal node ($min_samples_split$) - 2, 5, 10
- Minimum number of samples required at each leaf (terminal) node ($min_samples_leaf$) - 1, 2, 4
- Sampling of data from training set with or without replacement ($bootstrap$) - True, False

The hyperparameter *class_weight* was set to 'balanced' for all RF models considered. After random hyperparameter search, the $max_features$ hyperparameter was set as 'sqrt' and the $n_estimators$ hyperparameter was limited to between 200 and 1000 trees for the grid hyperparameter search for the pathogenicity data set. A 'base' hyperparameter set used for untuned RF models was defined with the following parameters: $n_estimators = 1000$, $max_features = 2$, $class_weight =' balanced'$. Default values for the classifier were as follows: $max_depth =' None'$, $min_samples_split = 2$, $min_samples_leaf = 1$, $bootstrap =' True'$.

LR classifiers were optimised for 3 hyperparameters:

- Inverse strength of regularisation used (C) - 100, 10, 1, 0.1, 0.01
- Type of regularisation used ($penalty$) - L2, None
- Maximum number of iterations for solver to converge (max_iter) - 100, 200, 300

SVM classifiers were optimised for 3 hyperparameters:

- Inverse strength of regularisation used (C) - 100, 10, 1, 0.1, 0.01
- Type of kernel used (*kernel*) - Linear, Polynomial, Radial Basis Function, Sigmoidal
- Kernel coefficient for non-linear kernels (*gamma*) - 'auto', 'scale'

Regularisation is a common method used to reduce the magnitude of features in the data set, as training LR or SVM classifiers on a large number of features may result in over-fitting of the model to the training data. The regularisation parameter C is the inverse of the strength of regularisation λ; thus, lower values of C increase strength of regularisation. All SVM classifiers used L2 regularisation. As there were far fewer possible combinations of hyperparameters for these architectures, a grid hyperparameter search was conducted using all models.

A 'base' hyperparameter set used for untuned LR models was defined with default parameters: $C = 1$, $penalty =' L2'$, $max_iter = 100$.

4.3.4 Feature Selection

Features used to construct the different classifier models were sub-divided into categories based on their provenance. Following previous similar work[206], which delineated features into sequence-dependent features (SEQ) and dynamics-derived features (DYN), I designated feature categories by shorthand abbreviations, as follows. A full list of features, their categorisation and abbreviations are given in Table 8.

SEQ features include both evolutionary features describing the conservation of sequence at the specific position at which the mutation occurs, using large-scale multiple sequence alignments of homologous sequences, and physico-chemical descriptors of the wild-type and variant residues. For the former, the position-specific independent counts (PSIC)[234] of the wild-type residue (WT_PSIC) and the difference in PSIC between

the wild-type and variant residues (*delta_PSIC*) were calculated using PolyPhen-2[144]. For the latter, each substitution was annotated with each of the ten Kidera factors (KFs) for the wild-type residues (wt_kf1, wt_kf2, wt_kf3, wt_kf4, wt_kf5, wt_kf6, wt_kf7, wt_kf8, wt_kf9, wt_kf10) and the difference in KF value between the wild-type and variant residues ($delta_kf1$, $delta_kf2$, $delta_kf3$, $delta_kf4$, $delta_kf5$, $delta_kf6$, $delta_kf7$, $delta_kf8$, $delta_kf9$, $delta_kf10$). Each KF represents a different physico-chemical feature or combination of features that collectively describe the properties of that residue in numerical form[235]; annotation of variants was performed using the $kideraFactors()$ function from the *Peptides* package in R[226].

Features from dynamical data were arranged into two categories, the first representing features reflecting the atomistic dynamics and mobility of the system (MD-MOB), and the second for structure-related features from the trajectory (MD-STR). The MD-MOB features were calculated as described in Chapter 3 and describe a subset of the dynamical motions of the atomistic system; three types of MD-MOB features were used as input for the classifier, these being the root mean square deviation (*RMSD*), the root mean square fluctuation at the mutated residue position ($RMSF$), and the radius of gyration (Rg). For each of these analyses, features were output corresponding to the absolute value of the mutant system across all frames (*abs*), the relative value of the mutant system versus the wild-type system (*rel*), and the standard deviation of values across all frames, both for absolute (*sd_abs*) and relative (*sd_rel*) properties.

The MD-STR features are all concerned with the solvent accessibility at the mutated residue; as MD trajectories were produced for each of the variant systems, solvent accessibilities were extracted as averages over the course of the simulation, for the wild-type system (*SASA_wt*), the mutant system (*SASA_mut*), and for the difference between the two ($SASA_diff$). From these same trajectories, the standard deviation in solvent accessibility across all frames was also calculated, again for the wild-type (*SASA_sd_wt*) and mutant (*SASA_sd_mut*) systems, and for the difference between them ($SASA_sd_diff$). Additionally, to investigate the contribution of these features in the absence of dynamical data, the

residue SASA for the wild-type (*st_SASA_wt*) and mutant (*st_SASA_mut*) were extracted from the post-equilibration, pre-simulation structures, as well as the difference between the two (st_SASA_diff). These were categorised as static structural features (ST-STR).

A final category of features was added after testing of initial classifiers, representing the physico-chemical change on mutation through amino acid substitution matrices (AASM). The first of these, Blocks Substitution Matrix 62 (BLOSUM62)[256], is a well-tested and commonly used substitution matrix for sequence alignments of proteins, based on substitution probabilities of amino acids in highly conserved proteins; BLOSUM62, so named as proteins with 62% or greater sequence identity are clustered as a single sequence, is the default matrix used for similarity search by protein BLAST (blastp)[208], and has also been used by pathogenicity prediction tools such as Rhapsody[176]. However, as the matrix is symmetric, it does not reflect residue substitutions with asymmetric substitution probabilities, such as proline; as such, I introduced another AASM which reflected these properties. DeMaSk[257] is a variant impact prediction tool which incorporates a directional AASM inferred from variant impact data in deep mutational scanning (DMS) data sets; I included the score from the AASM as an additional feature.

For each classifier, the model was first trained with base hyperparameter values using all possible features. The calculated feature importances were retrieved from the $feature_importances_$ parameter of the trained classifier; if features with little contribution to classification were removed, then the model was re-trained with this new feature set.

4.4 Prediction of Damaging / Neutral Titin Variant Impacts

Previous work on building a titin-specific RF classifier, building on the benefits of different feature types as described by Ponzoni *et al.*[176, 206] in their development of Rhapsody, was successful in producing a binary predictor of pathogenic vs. non-pathogenic variants that out-performed more general prediction algorithms. This predictor[258], trained on a data

set of 39 titin variants, found that the most descriptive features for their data set were SEQ features, including KF1, KF3, and ΔKF4, as well as PSIC, and that the most informative DYN features were contact-based. However, when I looked into the distributions of these features in a larger data set, which includes variants in both pathogenic and non-pathogenic categories not reported at the time of the previous work, the Kidera Factors, while still informative, seemed to separate the data less consistently than other features, especially *WT_PSIC* (SEQ), *SASA_WT* (MD-STR), and *RMSF_rel* (MD-MOB). Consequently, I constructed RF classifiers to train on the current titin variant data set.

To assess the baseline performance of this feature set for binary prediction of pathogenic vs. non-pathogenic titin variants, a first RF classifier was trained with all available SEQ, MD-STR, and MD-MOB features with the "base" set of hyperparameters (see Machine Learning Architectures). The performance of the classifier was tested using leave-one-out cross-validation (via $sklearn.model_selection.LeaveOneOut()$). This initial classifier successfully predicted the class of the target variant with $F1 = 0.78$ (24 TP, 7 FN, 26 TN, 4 FP). Interestingly, the top 6 features by importance in this classifier were almost without exception the same for all 61 leave-one-out iterations - in order of importance, these were *RMSF_rel*, *RMSF_mu_rel*, *SASA_rel*, *WT_PSIC*, *delta_PSIC*, and *d_KF4*.

RMSF_mu_rel, or the median per-trajectory replicate RMSF, is very similar to *RMSF_rel*, and was likely adversely impacting classifier performance by doubling the contributions of the MD-MOB features for which a median value had been calculated. When these features were removed and the classifier re-trained with these median features removed, the performance improved ($F1 = 0.87$; 26 TP, 5 FN, 27 TN, 3 FP). Repeat iterations of removal of features that did not contribute or hindered the classification reliability produced a feature set of, again, the six best features, which together performed as well as the previous classifier ($F1 = 0.87$; 26 TP, 5 FN, 27 TN, 3 FP), with optimal parameters for cross-validation as directed by parameter search ($n_estimators = 200$, $min_samples_split = 10$, $min_samples_leaf = 1$, $max_features =' sqrt'$, $max_depth = 20$, $bootstrap = True$).

Having assessed the capability of the given feature set to discriminate between damaging and tolerated variants, I looked more closely at the variants that the classifier did not classify correctly. Intriguingly, 3 of the 5 false negative predictions were made on AA-to-PRO substitutions - L18237P (Fn3-20)[35], H35946P (Ig-169)[150] and L35956P (Ig-169)[41] - while other pathogenic AA-to-PRO substitutions generally were predicted with lower probability of pathogenicity than other comparable variants. This is unexpected, as proline, the most rigid canonical amino acid, is not typically well tolerated as a substitution[259]. In Chapter 3, I discussed the difficulty with which the MD-MOB features could discriminate between pathogenic and non-pathogenic variants in Ig-169, a key hotspot domain; three of the five damaging substitutions in this domain are proline substitutions, and not only the wild-type Ig-169 but also both pathogenic and population variants were very stable in these simulations, with no indication that these variants were damaging.

Proline is a special case among the canonical amino acids; it is the only one for which the residue group is itself covalently bound to the backbone of the protein, thus introducing a conformational rigidity to the position that is not described by traditional physico-chemical annotations of the residue type. In addition, it does not follow the intuition of the classifier that the reduction of backbone flexibility in simulation (*RMSF_rel*) or that buried residues replaced with other hydrophobic residues (*SASA_wt*, d_KF4) increases the probability of a substitution being tolerated. Thus, I decided that features that more clearly described these special cases of mutation were a necessary addition to improve classifier performance.

Amino acid substitution matrices (AASMs) are commonly used for prediction or sequence similarity algorithms to define the physico-chemical variant impact upon substitution of a residue of one specific type with that of another in numerical form. However, the most commonly used such AASMs, such as BLOSUM and PAM, are symmetric, such that the value for an X-to-Y residue substitution is equivalent to a Y-to-X substitution. This does not reflect the consequences of proline substitution in terms of absolute variant impact, where mutations to proline are typically far less well tolerated than mutations from proline to any other residue. As such, I looked for substitution matrices that were more germane to the

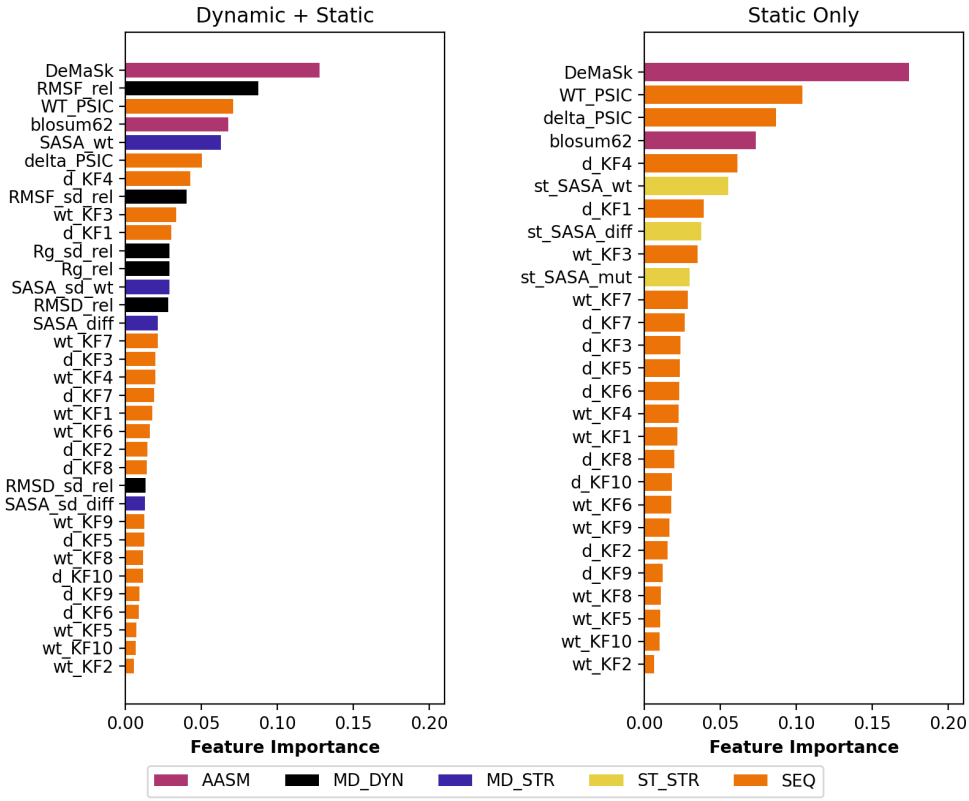


Figure 31: Calculated feature importances for Random Forest classifiers

Relative feature importances in the trained random forest classifiers for pathogenicity data incorporating both MD-derived and MD-agnostic features (left), and with MD-agnostic features only (right). Feature importances are calculated according to the Gini impurity of the feature across the whole ensemble of decision trees. Features are coloured according to their category: AASM (maroon); MD_DYN (black); MD_STR (dark purple); ST_STR (yellow); SEQ (orange).

question of variant impact in missense variants. DeMaSk is a relatively recent algorithm that predicts variant impact by combining a custom AASM with evolutionary conservation features; importantly, this AASM is derived from deep mutational scanning (DMS) data, so reflects the asymmetric nature of proline substitution.

The classifier pipeline was repeated using all SEQ, MD-STR, MD-MOB, and AASM features, and a further increase in model performance was seen, with all previously false negative proline substitutions bar H35946P predicted positive ($F1 = 0.9$; 27 TP, 4 FN, 28 TN, 2 FP); however, one non-pathogenic proline substitution, A22416P (Fn3-50) was incorrectly predicted positive. The hyperparameter set used was optimised via grid search ($n_estimators = 600$, $min_samples_split = 2$, $min_samples_leaf = 1$, $max_features =' sqrt'$, $max_depth = 10$, $bootstrap = True$). *DeMaSk* was judged the most important feature in the classifier, followed by *RMSF_rel*, WT_PSIC, $blosum62$, $delta_PSIC$, *SASA_wt*, and d_KF4 (see Figure 31).

In order to better understand what information was contributed by the different features and feature categories to the classifier, a second RF classifier was constructed with a starting feature set that excluded MD-MOB features entirely, and replaced MD-STR features with solvent accessibilities calculated from the static, pre-simulation structures. Surprisingly, after the same pipeline of hyperparameter tuning, this new classifier ($n_estimators = 400$, $min_samples_split = 2$, $min_samples_leaf = 1$, $max_features =' sqrt'$, $max_depth = 10$, $bootstrap = True$) marginally out-performed the classifier that included MD-MOB features ($F1 = 0.92$; 29 TP, 2 FN, 27 TN, 3 FP). In this "non-DYN" classifier, two more pathogenic variants were predicted correctly - H35946P (Ig-169) and A13715E (Ig-87). Here, *DeMaSk* was the most important feature by a substantial margin, followed by WT_PSIC, $delta_PSIC$, $blosum62$, d_KF4, and *st_SASA_wt* also contributing to the classifier (see Figure 31).

With this observation in mind, an untuned logistic regression classifier was constructed and the leave-one-out cross-validation repeated with multiple models trained on the most informative features of each of the categories. To elaborate, four classifiers were trained on

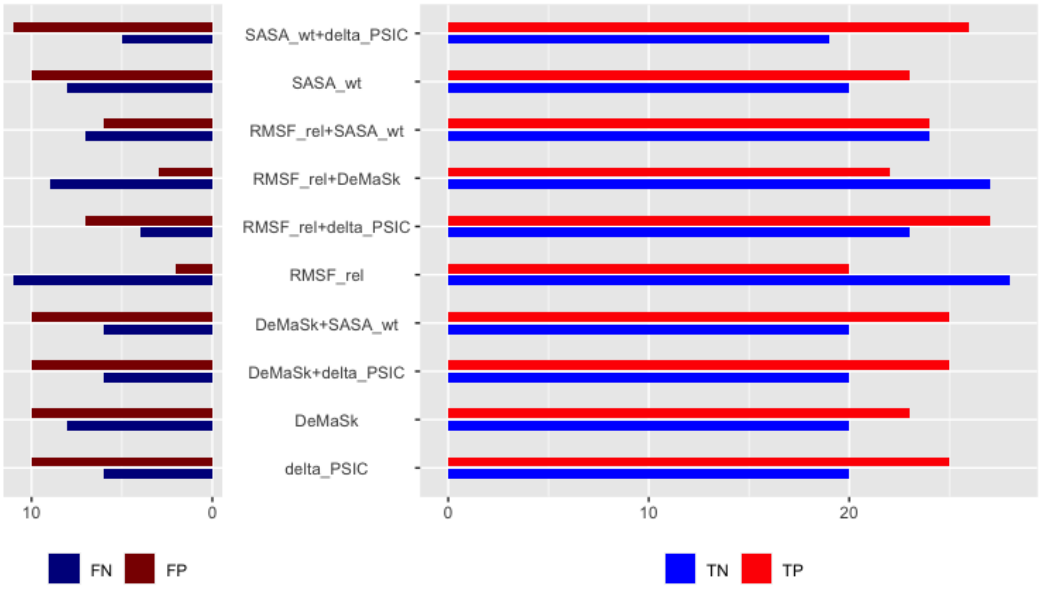


Figure 32: Distribution of predictions made by logistic regression classifiers trained on representative features

Distributions of prediction matrices for LR classifiers trained on 1-2 informative features for the pathogenicity data set. Counts for correct predictions are given in the plot to the right (TN, TP) while counts for incorrect predictions are given in the plot to the left (FN, FP). Feature(s) used in training each predictor are given in the y-axis.

each of *DeMaSk*, *RMSF_rel*, *delta_PSIC*, and *SASA_wt* individually, and then a further six classifiers on all binary combinations of features in this set. As can be seen in Figure 31, the performance of the three classifiers trained only on *DeMaSk*, *delta_PSIC*, and *SASA_wt*, respectively, are very similar, with a more accurate prediction of pathogenic variants than non-pathogenic; however, the classifier trained on *RMSF_rel* is more successful in predicting non-pathogenic variants ($TN = 28, FP = 2$) than pathogenic variants ($TP = 20, FN = 11$). When feature sets are combined, those that include *RMSF_rel* continue to have a higher accuracy of classification of non-pathogenic variants versus other classifiers, while improving their accuracy for pathogenic variants, whereas the combinations of the other three features do not appear to substantially alter the accuracy of their predictions relative to the classifiers using single features alone.

I had initially chosen to use the RF architecture for these machine learning models owing to its robustness in data sets with smaller numbers of samples and larger numbers of features and its successful use in similar, published methods[176]. To determine whether other oft-used architectures for machine learning could produce comparable results, two alternative architectures were constructed (both including and excluding MD features) and the leave-one-out cross-validation and hyperparameter tuning pipeline was repeated, using all features in the final random forest models. The LR model without MD features ($C = 1$, $max_iter = 100$, $penalty =' l2'$) performed almost as well as the best RF model ($F1 = 0.92$; 28 TP, 2 FN, 28 TN, 2 FP), but the LR model with MD features ($C = 0.01$, $max_iter = 100$, $penalty =' l2'$) suffered from diminished performance ($F1 = 0.82$; 23 TP, 8 FN, 28 TN, 2 FP). Both of the SVM models performed similarly, but slightly less well than the RF models; the SVM model with MD features ($C = 0.1$, $gamma =' scale'$, $kernel =' rbf'$) performed better at identifying damaging variants ($F1 = 0.89$; 28 TP, 3 FN, 26 TN, 4 FP), while the SVM without MD features ($C = 0.1$, $kernel =' linear'$) was better at identifying neutral variants ($F1 = 0.88$; 26 TP, 5 FN, 28 TN, 2 FP). This indicates that the signal in the data distinguishing pathogenic from non-pathogenic is picked up by different ML architectures, and also that, while the RF-based model performs better than other architectures overall, the difference in prediction accuracy is small (see Figure 33A).

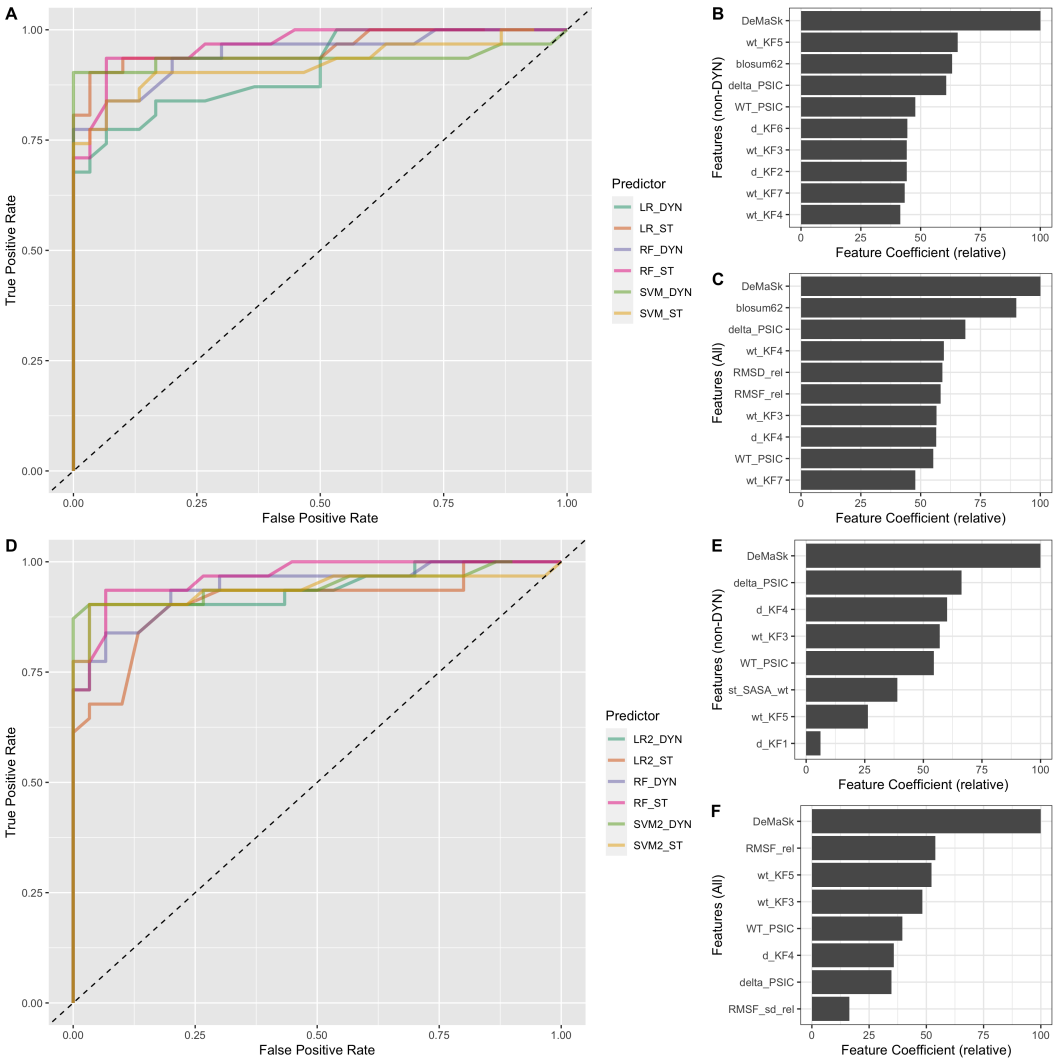


Figure 33: ROC curve for Titin variant classifiers with different architectures

A. *ROC curve for classifiers trained on all features (excluding absolute MD features) on the pathogenicity data set, based on leave-one-out cross-validation. Classifiers are identified by architecture (LR = logistic regression, RF = random forest, SVM = support vector machine) and by whether or not MD-derived features were included (DYN) or excluded (ST) from the feature sets.* ***B.*** *Most influential features by coefficient magnitude in training LR classifier without MD features.* ***C.*** *Most influential features by coefficient magnitude in training LR classifier with MD features.* ***D.*** *ROC curve for classifiers trained on reduced subset of features on the pathogenicity data set, based on leave-one-out cross-validation. Classifiers are identified by architecture (LR = logistic regression, RF = random forest) and by whether or not MD-derived features were included (DYN) or excluded (ST) from the feature sets.* ***E.*** *Most influential features by coefficient magnitude in training LR classifier with reduced feature set and without MD features.* ***F.*** *Most influential features by coefficient magnitude in training LR classifier with reduced feature set and with MD features.*

Curiously, the most heavily weighted features for the trained LR models are different to the features with the highest importances for the trained random forest methods (see Figure 33B, C). Notably, while *DeMaSk* remains the most informative feature by weight in both LR models with and without MD features, other heavily weighted features included wt_KF5 and wt_KF7, two features that had little to no importance in the RF models. Moreover, these feature coefficients are consistently highly weighted across all iterations of leave-one-out CV, and all features were scaled prior to model training. This suggests that there is redundancy in the feature set used for this classification, and that different combinations of weights on features can arrive at similarly good prediction results. Thus, it is conceivable that different starting points for training of LR models may arrive at more optimal final feature sets, and thus improve performance.

To investigate further how these different features influence the LR classifier, two additional feature sets were constructed and used to train new LR models; here, the starting feature set consisted only of those that were most informative in the original RF classifiers (*DeMaSk*, $RMSF_rel$, WT_PSIC, $delta_PSIC$, $SASA_wt$, d_KF4), then features highly weighted in the LR classifiers were added to gauge the change in performance ($blosum62$, wt_KF5, wt_KF7, wt_KF3, d_KF1, etc), followed by hyperparameter tuning. By seeding the initial feature set with those that our data exploration indicated were most strongly informative, we tested whether the LR and SVM classifiers could approach or exceed the performance of the random forest models. The revised LR classifier with MD-MOB features saw a substantial performance increase ($F1 = 0.93$; 28 TP, 3 FN, 29 TN, 1 FP), with a final feature set of 10 features (*DeMaSk*, $RMSF_rel$, WT_PSIC, $delta_PSIC$, $SASA_wt$, d_KF4, $RMSF_sd_rel$, wt_KF5, wt_KF3, d_KF1). However, the same process resulted in a decreased performance for the LR classifier without MD features ($F1 = 0.86$; 27 TP, 4 FN, 27 TN, 3 FP; 8 features; *DeMaSk*, WT_PSIC, $delta_PSIC$, st_SASA_wt, d_KF4, wt_KF5, wt_KF3, d_KF1).

Interestingly, both of these new feature sets resulted in a substantial improvement in predictor performance when they were used for SVM architectures. As seen in Figure 33 D, the performance of both the tuned SVM classifier with MD features ($C = 1.0$, $gamma =' scale'$,

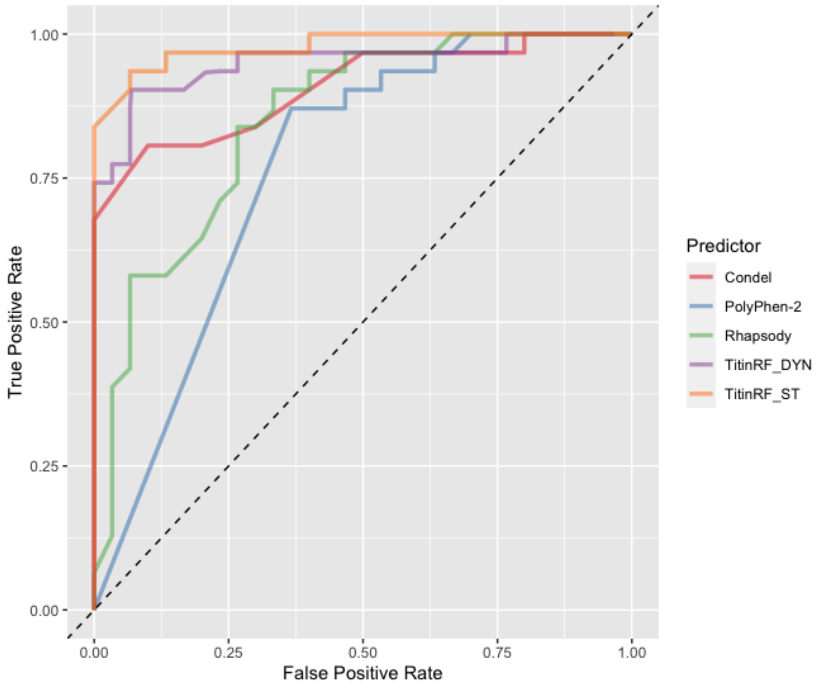


Figure 34: ROC curve for TitinRF classifiers and select pathogenicity classifiers in common use

ROC curve for random forest classifiers trained on the pathogenicity data set, based on leave-one-out cross-validation, against commonly used pathogenicity predictors (Condel, PolyPhen-2, Rhapsody). $TitinRF_{DYN}$ includes MD features, $TitinRF_{ST}$ excludes MD features.

$kernel =' rbf'$) and without MD features ($C = 0.1$, $kernel =' linear'$) were comparable to the RF architectures in performance: The $SVM2_{DYN}$ predictor had an F1 score of 0.93, with no incorrect predictions of population variants (27 TP, 4 FN, 30 TN, 0 FP), and the $SVM2_{ST}$ predictor had an F1 score of 0.92 (28 TP, 2 FN, 28 TN, 2 FP). However, for each of these models, slight changes to hyperparameter values could drastically affect leave-one-out accuracy; all of the different architectures could achieve very high prediction accuracy with different combinations of features and hyperparameters. This tells us that the features in the data set are ideally suited to discriminating between the pathogenic and non-pathogenic distributions, and work with several different architectures.

Finally, the performance of these predictors, both with ($TitinRF_{DYN}$) and without ($TitinRF_{ST}$) MD features, were compared against a selection of published and commonly used pathogenicity prediction algorithms. These comprised Rhapsody, the RF-based classifier that also incor-

porates dynamical features in prediction[176]; Condel, a sequence-based meta-predictor[173]; PolyPhen-2, a predictor that uses both sequence conservation and structural data, where available[144]. Predictions for each of the missense variants in this data set were retrieved and the F1 score and the Matthews correlation coefficient (MCC) were calculated for these predictors, using the classification thresholds typically employed for these classifiers to delineate predicted pathogenic from non-pathogenic ($Rhapsody \geq 0.7$, $Condel \geq 0.52$, $PolyPhen2 \geq 0.9$), and the number of true positive, true negative, false positive and false negative predictions for each:

	$TitinRF_{DYN}$	$TitinRF_{ST}$	$Rhapsody$	$Condel$	$PolyPhen2$
TP	27	29	26	25	29
TN	28	27	22	24	14
FP	4	2	8	6	16
FN	2	3	5	6	2
F1	0.90	0.92	0.80	0.81	0.76
MCC	0.81	0.84	0.58	0.61	0.46

The F1 score was similar for each of the published predictors, and higher for both $TitinRF_{DYN}$ and $TitinRF_{ST}$. The MCC further delineated the predictor performance: PolyPhen-2, with an MCC of 0.46 and a false positive rate exceeding its true negative rate, performed most poorly; Condel and Rhapsody performed similarly to one another, with only minor difference in MCC; both predictors were greatly out-performed by both the $TitinRF_{DYN}$ and $TitinRF_{ST}$ classifiers, which were likewise similar in terms of MCC. These differences are recapitulated in the ROC curves for each predictor (see Figure 34), which take into account all prediction thresholds; the curve for PolyPhen-2 has the lowest area under the curve, followed by Rhapsody, then Condel, and finally $TitinRF_{DYN}$ and $TitinRF_{ST}$ perform similarly well, with $TitinRF_{ST}$ slightly greater in terms of area under the curve.

4.5 Prediction of Titin Domain Stability Changes

The prediction of damaging variants, and thus the ability to quickly and reliably locate genetic causes for inherited diseases and conditions, has long been an aspiration in bioinformatics fields; however, by the very nature of the prediction task, the data sets involved are heterogeneous, with variants of many different pathologies and thus pathomechanisms collected together under a single label. Consequently, while the impetus for simple binary classification of damaging and non-damaging variants remains strong, there is now an increasing interest in using predictions of simpler metrics for characterisation of protein variants as input to pathogenicity prediction programs. For example, the melting temperature (T_m) of a mutant protein compared to its wild-type T_m is commonly characterised as an experimental derivation of thermostability, and a severe reduction in the thermostability of a protein is generally regarded as a strong indicator of pathogenicity. However, the prediction of the T_m of a protein or protein domain presents its own challenges; notably, existing predictors cannot reliably distinguish smaller changes in thermostability in the same protein, which include the majority of missense variants[247].

Titin presents an intriguing system for prediction of T_m and thus thermostability: it is a modular protein, and consists of many small, self-folding domains that can be considered separately from the remainder of the protein; these domains are, with the exception of the titin kinase domain, very similar in size and fold; finally, many variants that result in the destabilisation of their parent domain have been associated with different diseases, linking the thermostability and pathogenicity of titin variants. A moderately-sized data set of missense variants in individual titin domains is available, as are the constructs used for experimental validation, and the generated MD simulations of these systems and extracted features. Consequently, the lessons learned from the prediction of pathogenic and non-pathogenic variants were taken and applied to the problem of domain thermostability.

As a first step, I determined the effectiveness of the trained RF classifiers for pathogenicity variants in separating this thermostability data set. Both the RF classifier with and without

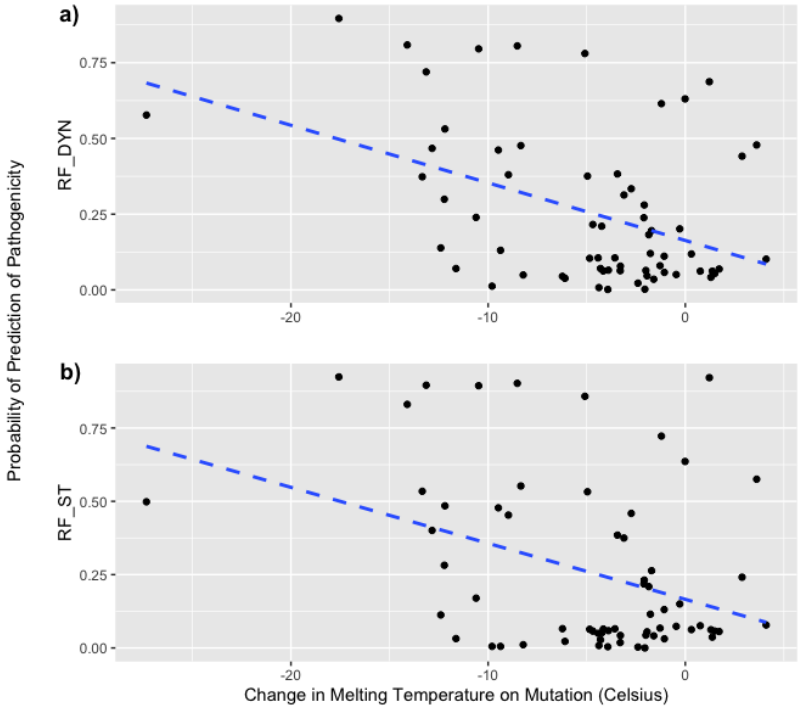


Figure 35: Plots of ΔT_m for variants in thermostability data set against their predicted pathogenicity

Scatter plots of the experimentally characterised melting temperature of titin variants in the Thermostability data set against the predicted probability of the variant's pathogenicity, calculated by a) the RF-based pathogenicity classifier including MD features, and b) the RF-based pathogenicity classifier excluding MD features. A linear regression slope calculated by the R package lm() is given in blue.

MD-MOB features were tested. Owing to the binary nature of the RF classifiers, the thermostability data was split into notional "pathogenic" and "non-pathogenic" labels based on the associated ΔT_m - variants in the "Weakly Stabilising" or "Weakly Destabilising" categories ($5^\circ C \geq \Delta T_m \geq -5^\circ C$) were denoted "non-pathogenic" ($n = 44$), while those in the "Moderately Destabilising" or "Strongly Destabilising" categories ($\Delta T_m < -5^\circ C$) were denoted "pathogenic" ($n = 21$). An issue with this split was that the two categories were very unequal; unfortunately, over half of the variants in the data set are categorised as "Weakly Destabilising" ($0^\circ C \geq \Delta T_m \geq -5^\circ C$), which precluded a clear and obvious split that would give balanced classes.

Both the RF classifiers showed a similar prediction output on the thermostability data (see Figure 35); there was a very weak but significant correlation between the estimated probability that the variant was pathogenic and the change in T_m in the mutant domain versus the wild-type (simple linear regression; $R^2 = 0.184, p = 3.57 \times 10^{-4}$ (RF_{DYN}); $R^2 = 0.139, p = 2.22 \times 10^{-3}$ (RF_{ST})). Both predictors performed most strongly in identifying the weakly T_m-altering variants as "non-pathogenic"; $TitinRF_{DYN}$ performed marginally better ($TN = 40$, $FP = 4$) than $TitinRF_{ST}$ ($TN = 39$, $FP = 5$). Both classifiers performed less well in identifying moderately destabilising variants as "pathogenic" ($TN = 7$, $FP = 14$); this may reflect the choice of cut-off boundary employed, as it may be argued that variants in the range of $-5°C \geq \Delta T_m \geq -15°C$ are not substantially damaging variants. If only variants in the "Strongly Destabilising" category ($\Delta T_m < -17°C$, $n = 2$), which are most likely to reflect pathogenic variants, are considered, P21683T is predicted by both classifiers to be the most likely pathogenic variant in the whole data set; however, the most destabilising variant R25012W is predicted non-pathogenic by $TitinRF_{ST}$, and pathogenic by $TitinRF_{DYN}$, albeit with a probability only just over the prediction threshold.

These results indicate that, while there is some correlation between pathogenicity and thermostability in titin variants, it is not the case that a reliable indication of decrease in melting temperature can be obtained through general pathogenicity assessment. This is expected, as a variant need not necessarily be destabilising to be causative of disease. However, this means that the features that are indicative of the pathogenicity label are not necessarily indicative of a thermostability label. Thus, the pipeline to construct RF, LR, and SVM classifiers was repeated to find features that would capture the information that could better represent the experimental measurements in question. Once again, to simplify the classification problem and allow for direct comparison with the previous results, the four ΔT_m classes were combined into two for binary prediction; "Weakly Stabilising" and "Weakly Destabilising" variants were denoted "non-destabilising" ($n = 44$), while those in the "Moderately Destabilising" or "Strongly Destabilising" categories were denoted "destabilising" ($n = 21$).

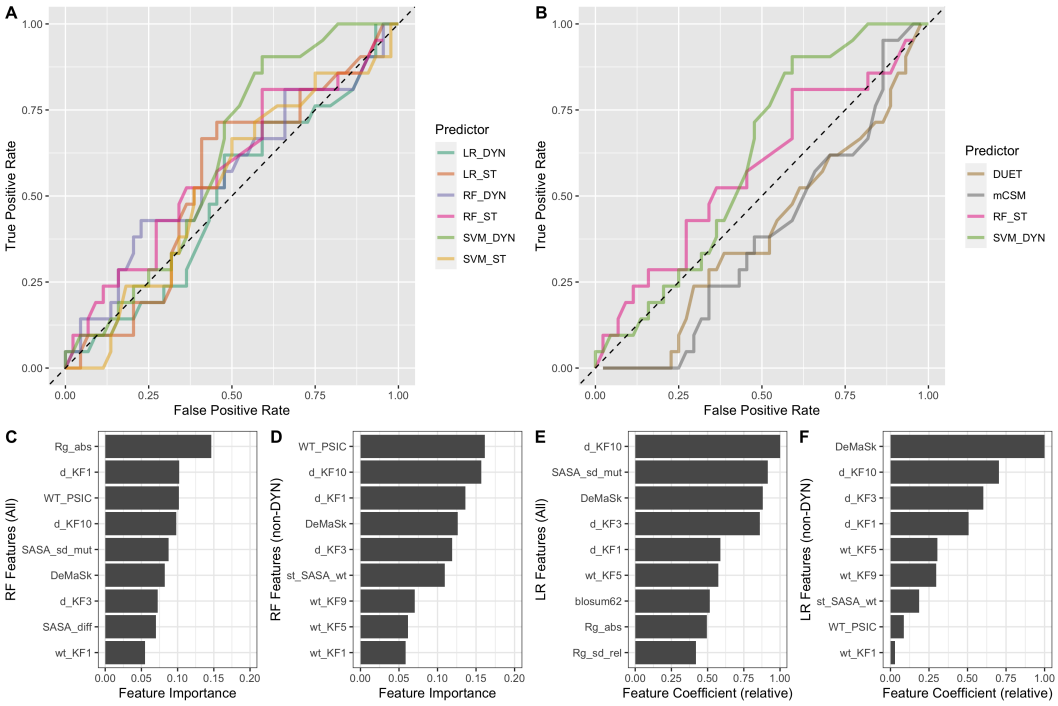


Figure 36: ROC curves for Titin thermostability classifiers with different architectures and feature sets

A. *ROC curve for classifiers trained on selected features on the thermostability data set, based on leave-one-out cross-validation. Classifiers are identified by architecture (LR = logistic regression, RF = random forest) and by whether or not MD-derived features were included (DYN) or excluded (ST) from the feature sets.* ***B.*** *ROC curve for Random forest classifiers with MD-derived features included (DYN) and excluded (ST) trained on the thermostability data set, based on leave-one-out cross-validation, as well as classifiers used for saturation mutagenesis prediction of thermostability in TITINdb2 (mCSM, DUET).* ***C.*** *Most important features by Gini impurity in training RF classifier with MD features.* ***D.*** *Most important features by Gini impurity in training RF classifier without MD features.* ***E.*** *Most influential features by coefficient magnitude in training LR classifier with MD features.* ***F.*** *Most influential features by coefficient magnitude in training LR classifier without MD features.*

As can be seen in Figure 36A, the performance of these classifiers, after feature selection and hyperparameter tuning, remained poor; there was no substantial improvement for any combination of architecture or feature categories. Despite some degree of improvement versus the pathogenicity classifiers, the distribution of predictions made is largely the same, with a stronger recognition of the "non-destabilising" class (RF_{DYN}: TN = 40, FP = 4; RF_{ST}: TN = 35, FP = 9; LR_{DYN}: TN = 32, FP = 12; LR_{ST}: TN = 36, FP = 8; SVM_{DYN}: TN = 40, FP = 4; SVM_{ST}: TN = 32, FP = 12) than of the "destabilising" class (RF_{DYN} TP = 9, FN = 12; RF_{ST}: TP = 11, FN = 10; LR_{DYN}: TP = 14, FN = 7; LR_{ST}: TP = 13, FN = 8; SVM_{DYN}: TP = 10, FN = 11; SVM_{ST}: TP = 14, FN = 7). This suggests that both the pathogenicity and thermostability-based classifiers suffer from the same issue when attempting to describe the separation between highly destabilising and relatively non-destabilising variants, in that the features collected cannot reliably identify the former, despite the observed high performance in identifying pathogenic variants. However, these predictors did compare favourably to predictors of domain stability change used to annotate variants in TITINdb2, with similar computation times; both mCSM and DUET, when predicted stability change ($\Delta\Delta G, kcalmol^{-1}$) was transformed to relative probability of the variant being destabilising, performed poorly at discriminating between the two classes (see Figure 36B). While thermodynamic stability in $\Delta\Delta G$ is not equivalent to melting temperature, the two properties are highly correlated[247], making them suitable for comparison.

While these results confirm the difficulty with which computational approaches must approximate experimental characterisations of domain properties, it is curious to note that the SVM predictor with MD features (SVM_DYN; see Figure 36A, light green) has an improved ratio of predictions compared to the other models at higher prediction thresholds. While this does not compare to the performance of the pathogenicity classifiers, it does suggest that the SVM architecture is at least partially learning a mapping eluded by the other architectures; as the SVM is kernel-based, and is thus able to learn non-linear functions, it is possible that there is information within the dynamical features derived from MD that is more informative by its interaction with other features. In the previous chapter, we discussed the interaction between the solvent accessibility and the change in residue hydrophobicity for destabilising

variants (see "Case Study: Fn3-70 as a Model System"); more such non-linear relationships between different features may be in play here.

Interestingly, despite the similarity of the distributions of the predictions made by both the pathogenicity-based (hereafter "*TTN_path*") and thermostability-based (hereafter "*TTN_therm*") classifiers, and the majority of variants classified incorrectly in the one being classified the same in the other, the features used to construct the different *TTN_therm* classifiers are distinct from those used to construct the *TTN_path* classifiers (see Figure 36B-D). For example, all of the *TTN_therm* classifiers are strongly influenced by ΔKF10, a feature that made little to no difference to the performance of the *TTN_path* classifiers and was excluded; conversely, *DeMaSk*, which was by some way the most important feature in all *TTN_path* classifiers, is only the most influential in one of the four *TTN_therm* classifiers, the LR model without MD-MOB features. The radius of gyration, *Rg*, which was highlighted in Chapter 3 as one of the few MD-MOB features descriptive of the domain thermostability, is the most prominent feature in the RF model with MD-MOB features, and is also a contributing feature to the LR model with MD-MOB features to a lesser extent. In short, these results suggest that a similar mapping between inputs and labels has been established based on training on these two data sets, using different features in *TTN_path* and *TTN_therm* classifiers.

To investigate this point further, I elected to see how well the classification architectures compared to one another by testing on a single, separate data set. First, all variants in the "pathogenicity" data set that had both been reported in the literature as pathogenic or as non-pathogenic variants, and which had been experimentally characterised as both a single mutant domain system and a wild-type system were extracted. This gave a new test set of 20 variants (see Data Sets), which were used to test the *TTN_path* classifiers (trained on the 41 remaining variants in the pathogenicity training set), and the latter to test the *TTN_therm* classifiers (trained on the 64 variants in the thermostability data set). While the small size of the test set makes it hard to reliably gauge the performance of these classifiers, all but two of the classifiers performed similarly well:

	PRY	*PRN*	*PLY*	*PLN*	*TRY*	*TRN*	*TLN*
Training Data	Path	Path	Path	Path	Therm	Therm	Therm
Architecture	RF	RF	LR	LR	RF	RF	LR
MD Features	Y	N	Y	N	Y	N	N
TP	13	13	14	11	11	5	13
TN	3	3	3	3	3	3	5
FP	2	2	2	2	4	2	2
FN	2	2	1	2	3	10	1
F1	0.867	0.867	0.903	0.828	0.759	0.455	0.897
MCC	0.467	0.467	0.577	0.378	0.224	NA	0.671

To simplify the naming of the classifiers, the models were identified with a 3-letter code: the first letter indicates the training set used ("P" for the reduced pathogenicity data set, "T" for the thermostability data set); the second indicates the machine learning architecture used ("R" for the random forest, "L" for logistic regression); the third indicates whether or not MD-MOB and MD-STR features were included in the feature selection step ("Y" indicates their presence, "N" indicates their absence). As an example, the random forest classifier trained on thermostability data without MD features would be identified as *TitinTRN* or *TRN*.

The LR classifier trained on thermostability data with MD features (*TLY*) was excluded from the analysis due to predicting all inputs as "damaging". As can be seen in the table above, most of the classifiers performed to a similar degree in identifying "damaging" and "non-damaging" variants; the exception, *TRN*, performed poorly overall, with even a negative MCC. This can be seen in a ROC curve plotted for each of these classifiers (see Figure 37); the large gaps between lines in this plot are somewhat misleading, as they reflect the proportions of the test set in their magnitude, but are instructive for comparing the different classifiers. The best-performing classifiers are the LR-based models *PLN*, *PLY* and *TLN*, with the RF-based classifiers *PRY* and *PRN*, both with and without MD features, performing only slightly less well. In terms of F1 and MCC, the RF-based thermostability-

trained classifier *TRY* shows a small decline in performance compared to the other models, but with a similar distribution of predictions.

While it is tempting to declare *TLN* the best-performing model on the basis of its high F1 and MCC scores, it must be reiterated that the test set is by necessity an extreme distribution of variants, dominated by strongly destabilising and damaging variants from a minority of select domains. The two variants that it incorrectly classes as non-damaging are A82D (Ig-1) and H35946P (Ig-169), both of which are found in immunoglobulin domains; as there is only one non-fibronectin domain variant in the thermostability data set, it makes sense that it would likely perform less well on a test set that also had an equal distribution of Ig and Fn3 domain variants. In the test set, only 5 domains in the data set are Ig domain variants, while 15 are Fn3 variants, thus increasing its performance. Similarly, the decrease in the size of the pathogenicity training data set by almost a third has impacted the performance of the *TTN_path* classifiers, as seen by now incorrect predictions of variants that were previously predicted correctly, such as the AA-to-PRO variants L18237P and H35946P. More than this, however, the numbers of incorrectly predicted variants are simply too small to detach from artefact.

With this in mind, the fact that thermostability-trained variants can successfully and accurately predict variants reported pathogenic on account of thermostability-related characterisation, despite poor performance in the thermostability cross-validation, indicates both that a reliable thermostability predictor would be a useful input to future pathogenicity predictors, and also, that despite these efforts, the thermostability predictor continues to perform better on extreme data than on the more realistic distributions of small stability changes associated with missense variants. While there is a signal that can be identified by these machine learning architectures that is indicative of these pathogenicity changes, to improve predictions of thermostability will require wholly different feature sets, despite the presumed overlap between these two characteristics of these variants.

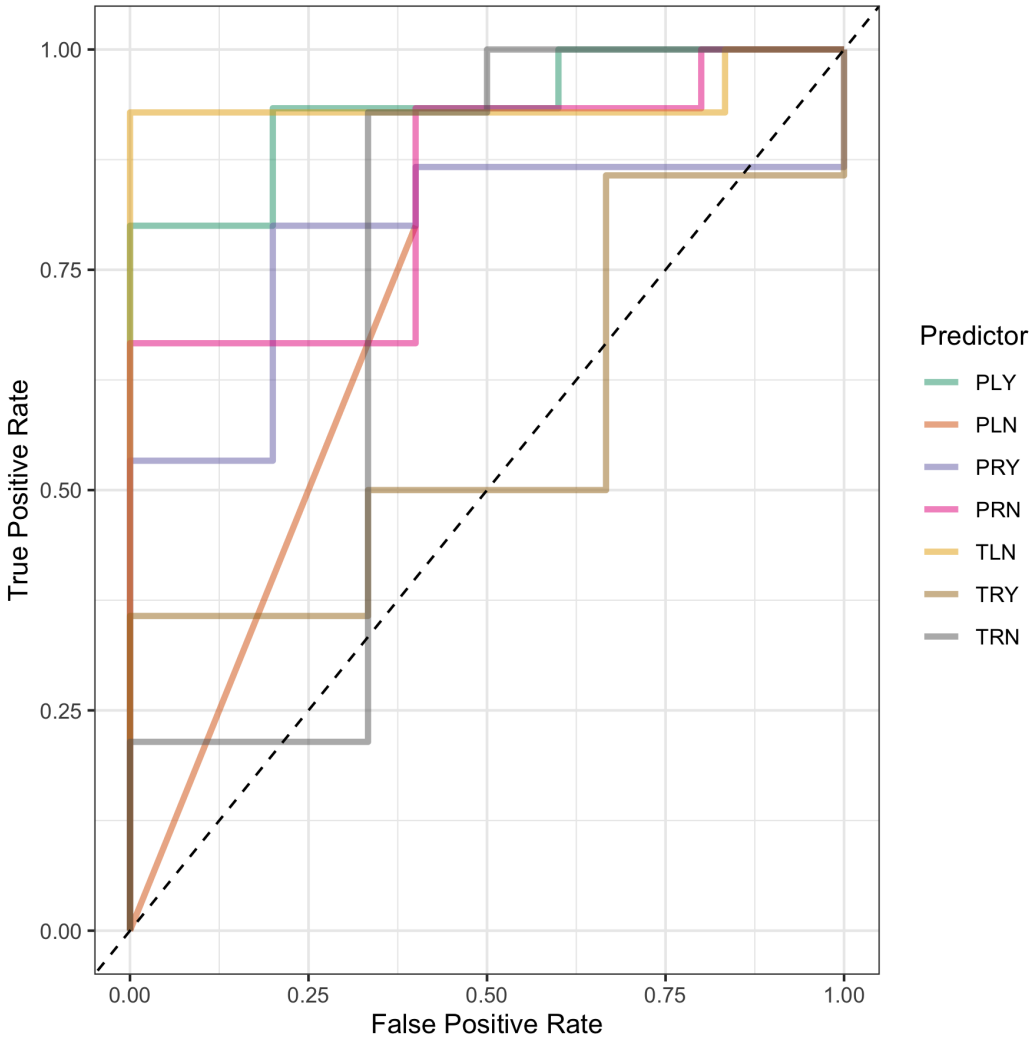


Figure 37: ROC curves for Titin classifiers with different architectures and training data, on test set with known pathogenicity and thermostability

ROC curve for classifiers tested on the combined thermostability + pathogenicity annotation set. Classifiers are identified by a three-letter code describing the training set (P = pathogenicity, T = thermostability), the model architecture (L = logistic regression, R = random forest), and by the presence or absence of MD-derived features in the feature sets (Y = included, N = excluded).

4.6 Discussion

4.6.1 The Importance of Features in Different Prediction Contexts

An emerging focus of prediction tools seeks to leverage information that was previously impractical or impossible to obtain through classification algorithms adept at navigating large, complex, heterogeneous feature sets; such approaches have previously led to the development of Rhapsody, a RF-based classifier that employs sequence, structural and dynamical data, and the creation of a simple titin-specific classifier that demonstrated the potential of these methodologies[258]. However, while said classifier performed equally well compared to my titin-based classifiers in predicting damaging and neutral variants, with a near-equivalent F1 score and MCC, it curiously differs greatly from these classifiers in terms of the features of greatest importance in constructing the RF. To elaborate, the previous titin classifier (hereafter $TitinRF_{init}$) was most influenced by sequence-based features, including ΔPSIC and the Kidera factors 1 and 3, whereas $TitinRF_{DYN}$ and $TitinRF_{ST}$ were most influenced by the DeMaSk substitution matrix, with the wild-type solvent accessibility, the wild-type and ΔPSIC also strongly contributing; for $TitinRF_{DYN}$, the relative RMSF at the mutated residue was more informative than any feature bar DeMaSk.

Such a difference appears unexpected, when the performance of all classifiers is so similar; the differences appear to reflect the discrepancies in the training sets used. In $TitinRF_{init}$, the training set consisted of 39 variants (19 damaging, 20 neutral); for $TitinRF_{DYN}$ and $TitinRF_{ST}$, many more variants were added to this data set from reports of disease-associated variants published subsequent to $TitinRF_{init}$ and further releases of population genome and exome data, up to 61 variants (31 damaging, 30 neutral). This increase in the training set size brings with it greater clarity on the relationship of different features to the broader span of titin mutations; wild-type Kidera factor 3, which as a feature is strongly predictive of damaging variants in $TitinRF_{init}$, while still contributing especially to $TitinRF_{ST}$, is secondary to features reflecting solvent accessibility and sequence conservation.

Interestingly, this is also observed for alternative versions of $TitinRF_{init}$; when the RF architecture was trained on variants only found in domains with solved crystal structures (15 variants) the most important features predominantly comprised those relating to the shortest paths of communication in network representations of protein structure, with the importance of sequence-based features greatly diminished.

These differences emphasise the importance of the representative data set for machine learning; biases present in smaller data sets lead to spurious relationships between features and labels. This applies equally to this data set of 61 variants; a relatively small proportion of the enormous quantity of possible variants in all titin domains, and certainly not sampling all possible circumstances that might result in pathogenicity. Moreover, each of the variants in the data set is, in some respect, an extreme; the most reliably annotated disease-causing variants, with effects so significant that they stand apart from the enormous mutational background of titin, versus the most reliably annotated neutral variants, which may be several orders of magnitude more common than most titin variants. Compounding this issue is the over-representation in the data set of variants in hotspot domains, especially Fn3-119 and Ig-169; while the similarity between these and other domains of the same fold in titin makes them well suited to inform a specific classifier, they are not absolutely representative of all domains of that fold. For example, in Chapter 3, I discussed how both the wild-type Ig-169 and damaging variant domains are remarkably similar and restricted in their dynamics compared to almost all other damaging variants in this data set, which showed much more variation.

While these issues described do not impact upon the current performance of these classifiers in cross-validation, they are a useful window through which to view the discrepancy between features most descriptive for the pathogenic vs non-pathogenic data set, and those for the thermal stability data set. Despite a degree of overlap between pathogenic and strongly destabilising titin variants, there are substantial differences between the most descriptive features for the two data sets; where there is a balance in feature importance between descriptors from different categories for the pathogenicity predictor, the thermostability

predictor is most influenced by different features, most notably the Δ for Kidera factor 10, wild-type PSIC, and radius of gyration. However, even with these features, the predictor does not perform well, even on a relatively simple binary classification.

The question of why prediction of thermostability is a more difficult problem than that of pathogenicity is one that merits further elaboration. In general terms, attempts by different groups to predict change in thermostability on mutation have been hampered by issues including the inherent experimental error, differences in experimental methodologies used, and a lower quantity of valid data compared to that available for pathogenicity prediction[247]. However, in this work these concerns were mitigated by the use of a small data set with consistent experimental design, of approximately the same size as the data set used for training the pathogenicity predictors, yet there was still a substantial margin in the performance of all thermostability predictors versus the pathogenicity predictors. Part of this may be ascribed to the feature sets used in this work, which were developed based on previous application to pathogenicity prediction but not specifically to thermostability prediction; however, it is also the case that the very question of thermostability change represents challenges not present for the pathogenicity prediction case. Pathogenicity is treated as a binary prediction problem, using distributions from two extremes, those variants definitively associated with pathogenicity and those least likely to be; while thermostability was here formulated in the same binary framework, the majority of variants represented only a minor change to the stability of the system, making it far harder to draw a decision boundary within the feature space employed. This is compounded by the discrepancy in the availability of destabilising versus stabilising variants, making it hard to draw a categorical boundary that reflects the variant impact on stability. Consequently, more work needs to be done to achieve a degree of accuracy in thermostability prediction comparable to pathogenicity prediction; one possibility is that future approaches split the problem into two or more constituent prediction targets, such as the prediction of the wild-type Tm and then the subsequent prediction of thermostability change.

Part of the explanation for this discrepancy may be laid on the sampling of variants for the thermostability data set, which is mostly comprised of A-band variants in fibronectin type-III domains with a single exception. It may be argued that this distribution, with strongly and moderately stability-altering variants in the minority, is more reflective of the reality of titin missense variants; a majority of small genetic changes in a given domain will have commensurately minor changes to experimentally derived parameters. Thus, the ability to successfully distinguish between weakly stabilising and destabilising variants, and not just to separate the most extreme examples of variation from the opposing extreme, will be key to future improvement in investigation of titin-related pathology. With many mutations present in the titin protein of any given individual, rapidly assessing not only those which are most damaging, but also the cumulative effect of many mutations on different domains, is an undertaking for which an alternative to experimental characterisation is needed.

4.6.2 The Contribution of Dynamical Features to Classification of Titin Variants

By training RF machine learning models on available damaging and neutral titin variants, I was able to construct two predictors which both performed very well in leave-one-out cross-validation - one which included dynamical features calculated from MD simulations of whole domain systems ($TitinRF_{DYN}$), and one without ($TitinRF_{noDYN}$). However, the performance of $TitinRF_{noDYN}$, equal to if not better than $TitinRF_{DYN}$ in cross-validation, begs the question of what is contributed to the classifier by the computationally expensive dynamical features. Indeed, when this question was explored with $TitinRF_{init}$, while it was found that atomistic dynamical features captured information absent from coarse-grained or network-based representations of proteins, their classifiers likewise performed similarly both with and without any dynamical features, whereas predictors with only dynamical features fell short.

With this in mind, an indication of the value of these simulations to classification may be gleaned from a comparison of the feature importances found in $TitinRF_{init}$ and those in these

classifiers. In $TitinRF_{init}$, the contributions of properties such as RMSF, which are strongly associated with pathogenicity in these classifiers, represent very little information; this is surprising, given the overlap between the two training data sets. This may reflect a difference in the way that values were calculated; it is notable that, in this data set, the two dynamical properties that most reflected the pathogenicity of the variant, the RMSF and the solvent accessibility, were the two properties that concerned only the residue position in question, rather than the entirety of the domain. This lends credence to the idea that, in this trajectory analysis, too much information is lost in features where the domain is considered as a whole; more focus on regions of interest within the domain, such as in RMSF- and SASA-related features, should be prioritised in future classification efforts.

In Chapter 3, I discussed how the effect of the A178D variant in Ig-2 could be observed in the breakdown of secondary structure in the simulation in areas in contact with that residue position; while there is still substantial perturbation around the mutant residue position itself, there are also other locations in the domain that experience altered dynamics, which are excluded by residue-specific features like this implementation of RMSF, and drowned by background noise in whole-domain features like RMSD. Thus, there is need for an intermediate that can meaningfully retrieve this information, yet doing so is difficult; as each variant will affect its parent domain in a different way, so too will the consequent dynamics be different. In short, the next step in feature extraction from dynamical data will be to connect elements of contacts, secondary structure, and other native features of the domain with atomic motions, and synthesise metrics that can accurately describe perturbation pertinent to the location of substitution.

Returning specifically to features included in these classifiers, a second question is that of why relative RMSF, which was more descriptive than any other non-substitution matrix feature, seemingly does not affect the prediction made when removed from the classifier. One possibility is that, as discussed above, the implicit information contained in the dynamical features selected is limited to the domain as a whole and the residue position, such that structure- and sequence-based features, as well as substitution-specific

features, obviate the information that would be included in the dynamical features; in other words, that while dynamical features contain information not found in these other features individually, it mostly overlaps when all of them are put together. Another possibility reflects the performance of the LR classifiers trained on only the most descriptive features of each source category; the RMSF-based classifier performed much better than those trained on DeMaSk, wild-type SASA and ΔPSIC in identifying neutral variants, but worse at identifying damaging variants. As the training set is largely composed of definitively pathogenic and neutral variants, this capacity to increase the specificity of the classifier for a slight decrease in sensitivity may be masked; the addition of RMSF to the classifier may make it more reliable on the broader set of titin variants, despite its slightly worse overall performance on extreme values.

This understanding of the redundancy in the feature set also offers an insight into the performance of the different logistic regression classifiers trained on the pathogenicity data set. In our original LR classifier, the most heavily weighted features were very different from those that were calculated as most important for the random forest classifiers. These produced a reasonable albeit less accurate classifier compared to random forest; however, when we used the most important features, based on the random forest feature importances and our own intuition from previous feature exploration, as a seed feature set to build new logistic regression classifiers, performance was much improved to a point where classification accuracy was comparable to the random forest architectures. Thus, the training of the LR classifiers may be likened to an optimisation problem covering a methodological landscape; it is possible for these classifiers to become stuck in local optima, and the resultant models to suffer in performance. Moreover, slightly correlated variables, such as those extracted from MD trajectories, may be under-estimated in importance purely by considering the feature coefficients, as if these variables were correlated, the weights of these two features together are diminished, even if they offer useful, separate information.

4.6.3 Consideration of Current and Future Machine Learning Strategies

For these prediction tasks, I primarily employed random forests as the standard architecture for machine learning; as the data set would necessarily include a large number of different features, with a limited number of samples available to make up the training data, it is necessary to select an architecture that is robust against over-fitting and maximises the capacity to make a correct prediction of any magnitude. I also compared the results obtained from random forest architectures to classifiers constructed from logistic regression and support vector machine architectures. For all of these different methodologies, while there were minor differences in performance, all models confirmed the presence of a strong signal in the features selected that was indicative of the pathogenicity label, although not indicative of domain thermostability. The prediction of thermostability has long been understood to be a difficult problem for machine learning approaches; between inherent experimental error, and discrepancies in experimental procedures or conditions between different studies, the theoretical maximum Pearson correlation between predicted and true values for thermostability predictors has been estimated at around 0.8[260]. In this work, rather than large quantities of data from available databases, a relatively small number of characterised titin variants with consistent experimental techniques and conditions were used to minimise the uncertainty of the data; however, there is still a limit to the prediction accuracy that can be reasonably expected.

Each of the classification architectures that were tested possess attributes that make them more or less suited to the prediction task at hand[261]. RF architectures, which are composed of an ensemble of branching decision trees, have the potential to learn mappings between input and output not afforded to the others, but are more opaque as a methodology and cannot extrapolate to values beyond the training set. LR architectures are examples of linear classifiers, which predict the class of a given input by linear combination of features and learned weights or coefficients, but cannot learn non-linear relationships between features. SVM architectures can learn non-linear relationships, but prediction probabilities and feature importances must be calculated independently and often inaccurately, and the small size

of the data set renders the use of non-linear kernels difficult. This leads to the question of whether or not a more suitable architecture exists for this problem.

As discussed in Chapter 3, the relationship between hydrophobicity and solvent accessibility with respect to thermostability is non-linear; a substitution of a hydrophilic residue in a buried position will be destabilising, as will a hydrophobic residue in an exposed position, but each feature must be at the opposite extreme to the other. Consequently, an architecture better placed to learn such relationships may be superior, and we have seen some evidence of increased performance for SVM architectures versus RF and LR; however, this is not the only consideration that must be balanced. The trade-off between RF and simpler linear architectures such as LR can be grossly simplified to one of accuracy versus interpretability; random forests, which tend to perform better than linear models on complex data, are less easy to interpret, and feature importances are less reliable than the parameter weights given by a linear model. The same concerns are pertinent to the discussion of other non-linear classifiers; methodologies that are generally considered more accurate than random forests, such as gradient-boosted trees or neural networks, are commensurately less interpretable and thus less informative on relevant biological questions. Similarly, as SVM architectures do not compute feature importances or weights, they are less informative than RF and LR methods. Finally, while I have not discussed regression methods for prediction of numerical differences in thermostability, this is another area where tree-based methods will perform less well as they cannot extrapolate outside of the training data.

As the quantity of data available for prediction of pathogenicity and thermostability increases, methodologies designed to leverage that data will become more attractive compared to the feature-based architectures I have described. Great success has been achieved in recent years in the field of protein structure prediction through AlphaFold2[170] and RoseTTAFold[262], among others; rather than learning from limited numbers of samples with highly curated feature sets, these predictors learn from enormous quantities of simple data using deep neural network architectures. Such methods will be of increasing importance for other prediction tasks in the future, but will require many intermediate steps before

realising this potential. For example, for prediction of protein thermostability, there is a sizeable quantity of data on all proteins in databases such as ProThermDB[263] and Thermo-MutDB[264]. However, these are not always properly curated; protein structures may not be readily available for all of the data; there are differences in experimental procedures and conditions between data (e.g. inconsistent pH); in many cases, there is only data on absolute thermostability or relative to wild-type systems, but not both. Once curated for reliable data points, the quantity of data available is insufficient for reliable deep learning methodologies. As thermostability data continues to be added and validated, and deep learning architectures become yet more intricate, this will doubtless change.

4.7 Conclusion

The predictors I have constructed demonstrate that the ability to reliably predict the pathogenicity of variants in titin has improved as more variants associated with different conditions have been identified; however, the ability to predict the thermostability of these variants has not progressed quite so much. Despite these two conceptions of the "damage" a missense variant can do to its parent protein or protein domain very much overlapping in standard thought, my analysis suggests that there is little to no overlap in features from these data sets that can reliably discern both properties. This is reflective of both a paradigm of feature design in which features tend to follow the very clear and obvious examples of damaging variants that are predominant in available training data sets, rather than the many more subtle changes associated with variants throughout nature, and also the limitations of the quantity of data available from which to draw conclusions. Thus, there is a need for more precise features that can accurately map from these smaller changes in domains to the numerical property assigned; while dynamical features are not necessary for accurate prediction of binary pathogenicity, they must be explored further in order to find the answers to more complex problems for both pathogenicity and thermostability.

Table 8: Features used in Classifier Development

Category	**Feature**	**Abbreviated Name**
SEQ features	Position-specific Independent Counts (PSIC) for wild-type residue at mutation position	WT_PSIC
	Difference in PSIC between wild-type and mutant residue at mutation position	$delta_PSIC$
	Preference for α-helix or bend structure for wild-type residue (Kidera Factor 1)	wt_KF1
	Bulk of side-chain for wild-type residue (Kidera Factor 2)	wt_KF2
	Preference for β- and extended structure for wild-type residue (Kidera Factor 3)	wt_KF3
	Hydrophobicity of wild-type residue (Kidera Factor 4)	wt_KF4
	Frequency of occurrence in double-bend regions of wild-type residue (Kidera Factor 5)	wt_KF5
	Average value of partial specific volume of wild-type residue (Kidera Factor 6)	wt_KF6
	Preference for flat extended structure of wild-type residue (Kidera Factor 7)	wt_KF7
	Frequency of occurrence in α-regions of wild-type residue (Kidera Factor 8)	wt_KF8
	pK-C of wild-type residue (Kidera Factor 9)	wt_KF9
	Hydrophobicity in β-structure surrounding wild-type residue (Kidera Factor 10)	wt_KF10
	Difference in KF1 between wild-type and mutant residue	d_KF1
	Difference in KF2 between wild-type and mutant residue	d_KF2
	Difference in KF3 between wild-type and mutant residue	d_KF3
	Difference in KF4 between wild-type and mutant residue	d_KF4
	Difference in KF5 between wild-type and mutant residue	d_KF5
	Difference in KF6 between wild-type and mutant residue	d_KF6
	Difference in KF7 between wild-type and mutant residue	d_KF7
	Difference in KF8 between wild-type and mutant residue	d_KF8
	Difference in KF9 between wild-type and mutant residue	d_KF9
	Difference in KF10 between wild-type and mutant residue	d_KF10

MD-MOB features	Root Mean Square Deviation (RMSD) of protein backbone of mutant domain system from starting structure, sum across all frames (*nm*)	*RMSD_abs*
	Difference in RMSD between wild-type and mutant systems	*RMSD_rel*
	Standard deviation of per-frame RMSD of mutant domain system	*RMSD_sd_abs*
	Difference in standard deviation of per-frame RMSD between wild-type and mutant systems	*RMSD_sd_rel*
	Root Mean Square Fluctuation (RMSF) of protein backbone at mutated residue position of mutant domain system versus average position, sum across all frames (*nm*)	*RMSF_abs*
	Difference in RMSF between wild-type and mutant systems	*RMSF_rel*
	Standard deviation of per-trajectory RMSF of mutant domain system	*RMSF_sd_abs*
	Difference in standard deviation of per-trajectory RMSF between wild-type and mutant systems	*RMSF_sd_rel*
	Radius of Gyration (Rg) of mutant domain system, sum across all frames (*nm*)	*Rg_abs*
	Difference in Rg between wild-type and mutant systems	*Rg_rel*
	Standard deviation of per-frame Rg of mutant domain system	*Rg_sd_abs*
	Difference in standard deviation of per-frame Rg between wild-type and mutant systems	*Rg_sd_rel*
MD-STR features	Solvent-Accessible Surface Area (SASA) of the residue at the mutated position in the wild-type system, average across all frames (nm^2)	*SASA_wt*
	SASA of the residue at the mutated position in the mutant system, average across all frames	*SASA_mut*
	Difference in SASA at the mutated position between wild-type and mutant systems	$SASA_diff$

	Standard deviation of per-frame SASA of the residue at the mutated position in the wild-type system	*SASA_sd_wt*
	Standard deviation of per-frame SASA of the residue at the mutated position in the mutant system	*SASA_sd_mut*
	Difference in standard deviation of per-frame SASA at the mutated position between wild-type and mutant systems	$SASA_sd_diff$
ST-STR features	Solvent-Accessible Surface Area (SASA) of the residue at the mutated position in the wild-type structure, post-equilibration (nm^2)	st_SASA_wt
	SASA of the residue at the mutated position in the mutant structure, post-equilibration	st_SASA_mut
	Difference in SASA at the mutated position between wild-type and mutant structures	st_SASA_diff
AASM features	Blocks Substitution Matrix 62 (BLOSUM62) score for residue substitution between wild-type and mutant	$blosum62$
	DeMaSk substitution matrix score for residue substitution between wild-type and mutant	*DeMaSk*

Discussion and Conclusions

The course of this work can be said to represent a pipeline leading to the annotation of variant impacts in a single protein; from the first Chapter, in which the variants associated with diseases are identified and curated, to the correlation of these variants with proposed sequence, structural, and dynamical features, and finally to the presentation of models that best map the input features to the annotation in question, I have sought to answer the question of what properties of single missense variants can reliably account for observed variant impact in titin. While it is not possible to infer global conclusions from the limited quantity of data available, these results present a blueprint for the investigation of other clinically relevant proteins, and hopefully the foundation upon which future applications, in titin or other proteins, may be built. In the following section, I discuss some of the considerations and potential avenues for these future directions.

When considering a single protein, especially one with many possible variants, it is not only the annotation of specific variants that must be optimised, but the ability of the landscape of computational evaluation to match the mutational landscape; in other words, for the assessment of impacts for different variants to permit quantitative and qualitative comparison between different variants in the same protein. In this work, I have taken steps in this direction; identifying the different features that have been, or could be, used to discriminate between variants in different functional classes, is a prerequisite to more meaningful variant annotation. Consequently, in Chapter 2, I investigated the performance and contributions of existing variant annotation tools, which predominantly rely on training using sequence or, more rarely, structural data, while in Chapter 3, I extended my analysis to features derived from MD simulations, not used in any of the previously described classifiers. This investigation of the coherence of these features with the actual, observed impacts of these variants led us to the development of my own predictor applications and models, with the proposition that protein-specific classification architectures could overcome the issues of

annotation of variants currently experienced by existing, general pathogenicity (see [265, 266] and thermostability (see [247]) predictors.

A curious finding of these predictor applications seemed to be that, while these predictors of pathogenicity performed very well compared to predictors of thermostability when cross-validated against variants from the same distributions, I surprisingly found that variants for which I had both pathogenicity and thermostability annotations were equally well-classified by models trained on either data set. In Chapter 4, I discussed some potential reasons why this would be the case despite the poor performance of the thermostability predictor in distinguishing moderately-to-strongly destabilising variants from others, including the composition of the training and test sets, the extreme values of the test set, and the redundancy of various feature combinations. However, one point that must be addressed is that of the very labels of "pathogenic" and "destabilising" themselves. While they are treated as simple binary problems to facilitate the evaluation of these methods in this text, the two properties are both multi-faceted, and their relationship to one another is not as simple as it first appears.

The pathogenicity and thermostability of variants have often been treated as equivalent; the rationale here is that a highly destabilising variant is almost certain to be damaging, in some fashion, to a protein or protein domain. However, as has previously been noted, the relationship does not hold when reversed; a variant associated with a particular disease need not necessarily be destabilising to the protein fold, for example if it affects a protein-protein interaction. Moreover, it is conceivable that a stabilising variant may also be pathogenic, if for example the reduction in flexibility of a titin domain affects unfolding and re-folding rates; the folding of titin domains has been proposed as having functional relevance to the muscle, though these claims have little support[267–270]. Along similar lines, these results showed that while there was overlap in predictions for variants that were specifically characterised as pathogenic on the basis of destabilisation of the parent domain, the features which were most indicative of pathogenicity-associated variants did not translate to a meaningful indication of decreased thermostability where there was no associated pathogenicity. This suggests that, rather than viewing thermostability as an input to the problem of pathogenicity, these two

parameters of variants should be treated as wholly separate problems in future computational work.

I have already described how the problem of pathogenicity extends far beyond the comparatively simple question of whether a variant is damaging or not; two damaging variants may cause entirely different conditions, localised to different sub-cellular or bodily locations, via different pathomechanisms or pathways, with differing severity, expressivity, or penetrance of the associated condition, yet are treated as the same class for the purposes of machine learning approaches. With this in mind, the ability to predict a property such as the thermostability change on mutation is appealing; a concrete numerical value may be applied to the property, which enables cross-comparison of different variants and prediction without the requirement of disease association for variants in the training set. However, as previous work has demonstrated and I have recapitulated here, the prediction of thermostability is fraught with difficulties; while most pathogenic variants are extreme examples that are clear outliers from the typical population variant, the ideal thermostability predictor should be able to delineate between these very minor changes, especially with respect to stabilising versus destabilising variants. However, most characterised missense variants in titin effect only minor changes in stability, which cannot be reliably inferred from the features that capture more dramatic changes.

The domain that has been most extensively studied in titin in terms of the methodologies used to ascertain stability is likely Ig-94, often referred to as "I27" in the literature, which has become a model system for the stability of Ig domains in titin, particularly those under mechanical stress in the I-band. This has included mutant domains, both with disease-associated variants[97] and with core destabilising variants[271]. Intriguingly, several of these papers note that characterisations of the stability of the domain can vary depending on different factors; for example, Ig-94 containing the missense variant Y14326P appears to have different stability depending on the unfolding pathway undertaken[192], with a decrease in thermal stability but a mechanical stability (as measured by single-molecule force spectroscopy) comparable to the wild-type domain[272]. Further discrepancies between

stability measurements in Ig-94 resulting from factors such as the composition of inter-domain linkers[273] and the temperature dependence of mechanical stability have been reported[274], alongside the effect of post-translational modifications such as Ca^{2+} binding on individual domain stability in other systems[122, 275].

While these considerations do not necessarily impact upon the thermostability predictors trained in this work, as these models specifically seek to predict the effect on melting temperatures of titin domains, based on measurements using the same methodology and protocol throughout, it does highlight the potential for given characterisations of variant stability impacts to be misleading in certain contexts. In general, there is a consensus that unfolding rates from different methodologies should correlate[99]; however, even if a single stability measurement is taken as definitive of the variant impact on stability, one must still beware of treating all modules in titin as identical, despite their high similarity to one another. For example, the wild-type domain Fn3-119, a hotspot for destabilising mutations, has a low melting temperature of around $49°C$[35]; this is around $10°C$ lower than the melting temperature for the mutant Fn3-20 domain with the destabilising variant L18237P ($63.5°C$), which destabilises its parent domain by around $17°C$. This is important to bear in mind, as a successful predictor of thermostability must not only be able to ascertain the impact of a given variant, but also to estimate the thermal stability of the wild-type domain. A potential next step may be to explicitly inform these models of wild-type stability, such that the specific features related to the change in stability on mutation can be better determined. Interestingly, rather than domains in different regions of titin having different characteristic stabilities, the gradient of stability appears to be distributed evenly across the length of the protein[276].

The selection of features for these machine learning models, both for thermostability and for pathogenicity, bears further examination. In this work, I selected derived properties of both static and dynamic representations of proteins that could contribute information relevant to these variant annotations, and assessed how well they could delineate between different annotation classes by combining them within a machine learning framework. My analysis

showed that the features were, in keeping with previous work, well suited to the prediction of pathogenicity, but poorly suited to the prediction of thermostability, or the construction of that question thus assembled. As I have mentioned, there are several reasons that can explain this discrepancy, including the differences in distributions of training data, the preponderance of MD-derived features averaging over the whole protein domain, and the relative simplicity of extracted features from MD simulation - with little focus on sub-groups of atoms in the system in lieu of treating the system as a whole. What I have yet to discuss is how these feature selection protocols reflect the purpose behind these methodologies. To elaborate, there is a distinction that can be drawn between explanatory models, which seek to identify those features which can rationalise the observations made, and predictive models, which seek to find the optimal mapping between input and output values.

Through these models, I have demonstrated that a number of features can contribute meaningful information concerning the damage that can be done to a titin domain by a given variant; however, this does not imply that these features, which may be highly predictive of a given outcome, are thus explanatory. For example, the evolutionary conservation of residue positions and frequency of substitution, as described by PSIC, is highly correlated with the damage dealt by that substitution; however, here the lack of the given residue substitution in nature is symptomatic of the negative selection under which that substitution is placed, but not the cause. In this work, I have attempted to explore both aspects of model design; in Chapter 3, I looked into how MD-derived features could describe the difference between wild-type and mutant systems, but also at what those changes constituted in actual terms, such as for the flexibility of the protein backbone or the integrity of the fold. In Chapter 4, I looked into how these features could contribute to predictive models of pathogenicity and thermostability based on their explanatory power. In future work, the MD-derived features should be tailored to encode a greater signal for classification purposes in predictive models; however, this will require a more nuanced view of the explanatory power of these features, such that the relevant subsets of molecular motions may be identified and explored. Thus, these problems should be considered separately, and indeed equally prioritised.

A question that was raised during the construction of the predictive models was that of why these models seemingly struggled, in the absence of DeMaSk or other substitution matrix-derived features, to identify substitutions to proline as pathogenic, when existing predictors of pathogenicity such as Condel and Rhapsody generally predict these substitutions correctly, using much the same features and information. We have already discussed the rationale for the introduction of AASM features as representing a property of specific residues, such as proline and glycine, that have no equivalent in other residues; the steric limitations introduced by proline substitution are not encoded in the sequence-based features used, as this is not a property shared by most residue types. The simplest reason for why these older predictors might predict these proline variants successfully, without the use of these specific features, is that most of these variants are older and very well-known; these variants may be included in their training data or in data used to construct subsidiary features, artificially inflating their performance on these variants. Interestingly, one of the most recent proline substitution variants, L18237P, which was described in 2021[35] and thus postdates the training of these classifiers, is misclassified by both Condel and Rhapsody. It should also be mentioned, however, that informative dynamical features such as the relative RMSF are more similar to wild-type for proline substitutions, owing to the reduced flexibility at that position; this may also contribute to the difficulty in classifiers incorporating dynamical features alone to correctly classify these variants.

A problem that will continue to prove challenging to overcome in any continuation of this work, and indeed in the wider field of computational biology, is that of sufficient sampling. For MD-derived features, sampling of conformational space over the course of a trajectory is necessarily limited by the length of the simulation, and the position of the initial configuration within the energy landscape; in a short simulation such as the 100 ns or 200 ns production runs, the majority of sampling will be in the vicinity of the starting position, without substantial changes to the overall backbone configuration of the protein. This is sufficient to produce a detectable signal associated with pathogenicity for the Literature + Population data set for mutated residue-specific features, but limits the useful information retrieved from a given quantity of computing time. Methods of enhanced sampling, such

as Replica Exchange MD (REMD)[241] and Gaussian Accelerated MD (GAMD)[242], have recently gained popularity as a means of sampling beyond the limits of conventional MD; REMD additionally offers a means of generalising multiple simulations at different temperatures simultaneously, which could be of benefit to deriving stability-associated features.

A second issue with sampling reflects the quantity of data points available; despite the enormous size of titin and the number of possible variants in its folded domains, only a tiny fraction of the most divergent variants are known to be disease-causing or disease-associated. Such associations can only be made when the variant is observed in affected populations, and with sufficient evidence to support its classification; similarly, the thermostability of a mutant protein domain is a labour- and time-intensive parameter to measure, limiting the speed with which new data becomes available. While these small data set sizes allow for more computationally expensive features to be calculated in a practical time frame, as per the MD-derived features, they also limit the applicability of the conclusions drawn from these models. Enormous sets of simple data have made great advances in prediction of protein structure[170, 262] and large-scale variant annotation[277, 278] possible, through the continued improvement of deep learning architectures that drive the principles of machine learning to their logical extreme. While it is not yet possible for such methods to be applied to smaller data sets reflecting single proteins, there are many approaches being developed that are beginning to bridge this gap; for example, variational auto-encoders (VAEs), which encode high-dimensional data into latent space and can generate outputs from that same space, can be used to augment smaller data sets with generated values. They have been successfully employed, in combination with transfer learning and convolutional neural networks, to investigate the underlying conformational space for different proteins[279, 280].

These considerations lay the groundwork for future investigation of this and similar problems. Although great strides have been made in the fields of computational biology and its application to applied sciences and medicine, the analysis of individual proteins and comparison of smaller genetic variations to one another is plagued by the shortage of

protein-specific data. Titin was and remains at the forefront of these efforts, as a model system best placed to finally break through these walls; future work will no doubt extend the lessons learned from it to other, similar proteins such as obscurin and myomesin. A deeper understanding of the consequences of protein damage, more than that of absolute dysfunction, is integral to the accurate annotation of missense variants, and I have demonstrated how methods that can capture that minute signal, such as molecular dynamics, will be needed more than ever as the field progresses. However, as with these predictors, these methods cannot prevail alone; it will take the combined efforts of scientists in clinical discovery, experimental characterisation, data curation, and computational analysis, for these efforts to truly bear fruit.

References

1. Maruyama, K. Connectin, an Elastic Protein from Myofibrils. *The Journal of Biochemistry* **80,** 405–407 (Aug. 1, 1976).
2. Labeit, S. *et al.* A regular pattern of two types of 100-residue motif in the sequence of titin. *Nature* **345,** 273–276 (May 1990).
3. Labeit, S., Gautel, M., Lakey, A. & Trinick, J. Towards a molecular understanding of titin. *The EMBO Journal* **11,** 1711–1716 (May 1, 1992).
4. Granzier, H. L. & Labeit, S. The Giant Protein Titin. *Circulation Research* **94,** 284–295 (Feb. 20, 2004).
5. LeWinter, M. M. & Granzier, H. Cardiac Titin - A Multifunctional Giant. *Circulation* **121,** 2137–2145 (May 18, 2010).
6. Loescher, C. M., Hobbach, A. J. & Linke, W. A. Titin (TTN): from molecule to modifications, mechanics, and medical significance. *Cardiovascular Research* **118,** 2903–2918 (Oct. 1, 2022).
7. Wang, Z. *et al.* The molecular basis for sarcomere organization in vertebrate skeletal muscle. *Cell* **184,** 2135–2150.e13 (Apr. 15, 2021).
8. Herzog, W., Leonard, T., Joumaa, V., DuVall, M. & Panchangam, A. The three filament model of skeletal muscle stability and force production. *Molecular & cellular biomechanics : MCB* **9,** 175–91 (Sept. 1, 2012).
9. Lindstedt, S. & Nishikawa, K. Huxleys' Missing Filament: Form and Function of Titin in Vertebrate Striated Muscle. *Annual Review of Physiology* **79,** 145–166 (Feb. 10, 2017).
10. Herzog, W. The multiple roles of titin in muscle contraction and force production. *Biophysical Reviews* **10,** 1187–1199 (Aug. 1, 2018).
11. Tonino, P. *et al.* The giant protein titin regulates the length of the striated muscle thick filament. *Nature Communications* **8,** 1041 (Oct. 19, 2017).

12. Luis, N. M. & Schnorrer, F. Mechanobiology of muscle and myofibril morphogenesis. *Cells & Development. Quantitative Cell and Developmental Biology* **168,** 203760 (Dec. 1, 2021).

13. Linke, W. A., Stockmeier, M. R., Ivemeyer, M., Hosser, H. & Mundel, P. Characterizing titin's I-band Ig domain region as an entropic spring. *Journal of Cell Science* **111,** 1567–1574 (June 1, 1998).

14. Linke, W. A. *et al.* I-Band Titin in Cardiac Muscle Is a Three-Element Molecular Spring and Is Critical for Maintaining Thin Filament Structure. *The Journal of Cell Biology* **146,** 631–644 (Aug. 9, 1999).

15. Anderson, B. R. & Granzier, H. L. Titin-based tension in the cardiac sarcomere: Molecular origin and physiological adaptations. *Progress in Biophysics and Molecular Biology. SI: Beating Heart* **110,** 204–217 (Oct. 1, 2012).

16. Sheetz, M. Mechanobiology in cardiac mechanics. *Biophysical Reviews* **13,** 583–585 (Oct. 1, 2021).

17. Solís, C. & Russell, B. Striated muscle proteins are regulated both by mechanical deformation and by chemical post-translational modification. *Biophysical Reviews* **13,** 679–695 (Oct. 1, 2021).

18. Lieber, R. L. & Binder-Markey, B. I. Biochemical and structural basis of the passive mechanical properties of whole skeletal muscle. *The Journal of Physiology* **599,** 3809–3823 (2021).

19. Crocini, C. & Gotthardt, M. Cardiac sarcomere mechanics in health and disease. *Biophysical Reviews* **13,** 637–652 (Oct. 1, 2021).

20. Van der Pijl, R. J. *et al.* The titin N2B and N2A regions: biomechanical and metabolic signaling hubs in cross-striated muscles. *Biophysical Reviews* **13,** 653–677 (Oct. 1, 2021).

21. Steinberg, S. F. Oxidative Stress and Sarcomeric Proteins. *Circulation research* **112,** 393–405 (Jan. 18, 2013).

22. Chauveau, C., Rowell, J. & Ferreiro, A. A Rising Titan: TTN Review and Mutation Update. *Human Mutation* **35,** 1046–1059 (2014).

23. LeWinter, M. M. & Granzier, H. L. Titin Is a Major Human Disease Gene. *Circulation* **127,** 938–944 (Feb. 26, 2013).

24. Eldemire, R., Tharp, C. A., Taylor, M. R., Sbaizero, O. & Mestroni, L. The Sarcomeric Spring Protein Titin: Biophysical Properties, Molecular Mechanisms, and Genetic Mutations Associated with Heart Failure and Cardiomyopathy. *Current Cardiology Reports* **23,** 121 (July 16, 2021).

25. Herman, D. S. *et al.* Truncations of Titin Causing Dilated Cardiomyopathy. *New England Journal of Medicine* **366,** 619–628 (Feb. 16, 2012).

26. Fomin, A. *et al.* Truncated titin proteins and titin haploinsufficiency are targets for functional recovery in human cardiomyopathy due to TTN mutations. *Science Translational Medicine* **13,** eabd3079 (Nov. 3, 2021).

27. Bailey, A. J. *et al.* Role of titin in vertebrate striated muscle. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **357,** 199–206 (Feb. 28, 2002).

28. Lange, S., Pinotsis, N., Agarkova, I. & Ehler, E. The M-band: The underestimated part of the sarcomere. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. Cardiomyocyte biology: new pathways of differentiation and regeneration* **1867,** 118440 (Mar. 1, 2020).

29. Wang, J.-H. The sequence signature of an Ig-fold. *Protein & Cell* **4,** 569–572 (Aug. 2013).

30. Laddach, A., Gautel, M. & Fraternali, F. TITINdb—a computational tool to assess titin's role as a disease gene. *Bioinformatics* **33,** 3482–3485 (Nov. 1, 2017).

31. Kellermayer, D., Smith, J. E. & Granzier, H. Novex-3, the tiny titin of muscle. *Biophysical Reviews* **9,** 201–206 (Apr. 7, 2017).

32. Lahmers, S., Wu, Y., Call, D. R., Labeit, S. & Granzier, H. Developmental Control of Titin Isoform Expression and Passive Stiffness in Fetal and Neonatal Myocardium. *Circulation Research* **94,** 505–513 (Mar. 5, 2004).

33. Zou, J. *et al.* An internal promoter underlies the difference in disease severity between N- and C-terminal truncation mutations of Titin in zebrafish. *eLife* **4** (ed Dietz, H. C.) e09406 (Oct. 16, 2015).

34. Savarese, M. *et al.* Improved Criteria for the Classification of Titin Variants in Inherited Skeletal Myopathies. *Journal of Neuromuscular Diseases* **7,** 153–166 (Jan. 1, 2020).

35. Rees, M. *et al.* Making sense of missense variants in TTN-related congenital myopathies. *Acta Neuropathologica* **141,** 431–453 (Mar. 1, 2021).

36. Oates, E. C. *et al.* Congenital Titinopathy: Comprehensive characterization and pathogenic insights. *Annals of Neurology* **83,** 1105–1124 (June 2018).

37. Huang, K. *et al.* Clinicopathological features of titinopathy from a Chinese neuromuscular center. *Neuropathology* **41,** 349–356 (2021).

38. Tubridy, N., Fontaine, B. & Eymard, B. Congenital myopathies and congenital muscular dystrophies. *Current Opinion in Neurology* **14,** 575–582 (Oct. 2001).

39. Shieh, P. B. Muscular Dystrophies and Other Genetic Myopathies. *Neurologic Clinics. Clinical Neurogenetics* **31,** 1009–1029 (Nov. 1, 2013).

40. Udd, B. & Hackman, P. in *GeneReviews®* (eds Adam, M. P. *et al.*) (University of Washington, Seattle, Seattle (WA), 1993).

41. Hackman, P. *et al.* Tibial Muscular Dystrophy Is a Titinopathy Caused by Mutations in TTN, the Gene Encoding the Giant Skeletal-Muscle Protein Titin. *American Journal of Human Genetics* **71,** 492–500 (Sept. 2002).

42. Hackman, P. *et al.* Truncating mutations in C-terminal titin may cause more severe tibial muscular dystrophy (TMD). *Neuromuscular disorders: NMD* **18,** 922–928 (Dec. 2008).

43. Pfeffer, G. & Chinnery, P. F. in *GeneReviews®* (eds Adam, M. P. *et al.*) (University of Washington, Seattle, Seattle (WA), 1993).

44. Palmio, J. *et al.* Expanding the importance of HMERF titinopathy: new mutations and clinical aspects. *Journal of Neurology* **266,** 680–690 (Mar. 2019).

45. Pfeffer, G. *et al.* Titin Founder Mutation is a Common cause of Myofibrillar Myopathy with Early Respiratory Failure. *Journal of neurology, neurosurgery, and psychiatry* **85,** 331–338 (Mar. 1, 2014).

46. Uruha, A. *et al.* Necklace cytoplasmic bodies in hereditary myopathy with early respiratory failure. *Journal of Neurology, Neurosurgery & Psychiatry* **86,** 483–489 (May 1, 2015).

47. Ceyhan-Birsoy, O. *et al.* Recessive truncating titin gene, TTN, mutations presenting as centronuclear myopathy. *Neurology* **81,** 1205–1214 (Oct. 1, 2013).

48. Chauveau, C. *et al.* Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Human Molecular Genetics* **23,** 980–991 (Feb. 15, 2014).

49. Cid, R. D. *et al.* A new titinopathy: Childhood-juvenile onset Emery-Dreifuss–like phenotype without cardiomyopathy. *Neurology* **85,** 2126–2135 (Dec. 15, 2015).

50. Jordan, E. *et al.* Evidence-Based Assessment of Genes in Dilated Cardiomyopathy. *Circulation* **144,** 7–19 (July 6, 2021).

51. Mazzarotto, F. *et al.* Reevaluating the Genetic Contribution of Monogenic Dilated Cardiomyopathy. *Circulation* **141,** 387–398 (Feb. 4, 2020).

52. Owens, A. T. & Day, S. M. Reappraising Genes for Dilated Cardiomyopathy: Stepping Back to Move Forward. *Circulation* **144,** 20–22 (July 6, 2021).

53. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine : official journal of the American College of Medical Genetics* **17,** 405–424 (May 2015).

54. Elliott, P. *et al.* Classification of the cardiomyopathies: a position statement from the European Society Of Cardiology Working Group on Myocardial and Pericardial Diseases. *European Heart Journal* **29,** 270–276 (Jan. 2008).

55. Maron, B. J. *et al.* Contemporary definitions and classification of the cardiomyopathies: an American Heart Association Scientific Statement from the Council on Clinical Cardiology, Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functional Genomics and Translational Biology Interdisciplinary Working Groups; and Council on Epidemiology and Prevention. *Circulation* **113,** 1807–1816 (Apr. 11, 2006).

56. Reuter, C. M., Dries, A. M. & Parikh, V. N. Arrhythmogenic Cardiomyopathy: Mechanisms, Genetics, and Their Clinical Implications. *Current Cardiovascular Risk Reports* **15,** 7 (Mar. 26, 2021).

57. Corrado, D. *et al.* Spectrum of Clinicopathologic Manifestations of Arrhythmogenic Right Ventricular Cardiomyopathy/Dysplasia: A Multicenter Study. *Journal of the American College of Cardiology* **30,** 1512–1520 (Nov. 15, 1997).

58. Sen-Chowdhry, S. *et al.* Left-Dominant Arrhythmogenic Cardiomyopathy: An Under-Recognized Clinical Entity. *Journal of the American College of Cardiology* **52,** 2175–2187 (Dec. 16, 2008).

59. Towbin, J. A., Lorts, A. & Jefferies, J. L. Left ventricular non-compaction cardiomyopathy. *The Lancet* **386,** 813–825 (Aug. 22, 2015).

60. Arbustini, E. *et al.* The MOGE(S) Classification for a Phenotype–Genotype Nomenclature of Cardiomyopathy: Endorsed by the World Heart Federation. *Journal of the American College of Cardiology* **62,** 2046–2072 (Dec. 3, 2013).

61. Westphal, J. G. *et al.* The MOGE(S) classification for cardiomyopathies: current status and future outlook. *Heart Failure Reviews* **22,** 743–752 (Nov. 1, 2017).

62. Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8* Nov. 2015.

63. Hulo, N. *et al.* The 20 years of PROSITE. *Nucleic Acids Research* **36,** D245–D249 (Database issue Jan. 2008).

64. Akinrinade, O. *et al.* Relevance of Titin Missense and Non-Frameshifting Insertions/Deletions Variants in Dilated Cardiomyopathy. *Scientific Reports* **9,** 4093 (Mar. 11, 2019).

65. Huynh, K. Truncated titin proteins in the pathophysiology of DCM. *Nature Reviews Cardiology* **19,** 6–6 (Jan. 2022).

66. McAfee, Q. *et al.* Truncated titin proteins in dilated cardiomyopathy. *Science Translational Medicine* **13,** eabd7287 (Nov. 3, 2021).

67. Dewi, I. P. & Nugroho, J. Genetic Polymorphism in Peripartum Cardiomyopathy. *Gynecology Obstetrics & Reproductive Medicine,* 11–5 (Mar. 31, 2021).

68. Iorgoveanu, C., Zaghloul, A. & Ashwath, M. Peripartum cardiomyopathy: a review. *Heart Failure Reviews* **26,** 1287–1296 (Nov. 1, 2021).

69. Roberts, A. M. *et al.* Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Science translational medicine* **7,** 270ra6 (Jan. 14, 2015).

70. Rich, K. A. *et al.* Novel heterozygous truncating titin variants affecting the A-band are associated with cardiomyopathy and myopathy/muscular dystrophy. *Molecular Genetics & Genomic Medicine* **8** (Aug. 20, 2020).

71. Jiang, H. *et al.* Functional analysis of a gene-edited mouse model to gain insights into the disease mechanisms of a titin missense variant. *Basic Research in Cardiology* **116,** 14 (Feb. 26, 2021).

72. Gigli, M. *et al.* A Review of the Giant Protein Titin in Clinical Molecular Diagnostics of Cardiomyopathies. *Frontiers in Cardiovascular Medicine* **3** (2016).

73. Campuzano, O. *et al.* Rare Titin (TTN) Variants in Diseases Associated with Sudden Cardiac Death. *International Journal of Molecular Sciences* **16,** 25773–25787 (Oct. 27, 2015).

74. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine* **375,** 655–665 (Aug. 18, 2016).

75. Haggerty, C. M. *et al.* Genomics-First Evaluation of Heart Disease Associated With Titin-Truncating Variants. *Circulation* **140,** 42–54 (July 2, 2019).

76. Nicolao, P. *et al.* Autosomal dominant myopathy with proximal weakness and early respiratory muscle involvement maps to chromosome 2q. *American Journal of Human Genetics* **64,** 788–792 (Mar. 1999).

77. Lange, S. *et al.* The Kinase Domain of Titin Controls Muscle Gene Expression and Protein Turnover. *Science* **308,** 1599–1603 (June 10, 2005).

78. Hedberg, C., Melberg, A., Dahlbom, K. & Oldfors, A. Hereditary myopathy with early respiratory failure is caused by mutations in the titin FN3 119 domain. *Brain* **137,** e270 (Apr. 1, 2014).

79. Weintraub, R. G., Semsarian, C. & Macdonald, P. Dilated cardiomyopathy. *The Lancet* **390,** 400–414 (July 22, 2017).

80. Evilä, A. *et al.* Atypical phenotypes in titinopathies explained by second titin mutations. *Annals of Neurology* **75,** 230–240 (2014).

81. Olivotto, I. *et al.* Obesity and its association to phenotype and clinical course in hypertrophic cardiomyopathy. *Journal of the American College of Cardiology* **62,** 449–457 (July 30, 2013).

82. Reineck, E. *et al.* Physical activity and other health behaviors in adults with hypertrophic cardiomyopathy. *The American Journal of Cardiology* **111,** 1034–1039 (Apr. 1, 2013).

83. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586,** 749–756 (Oct. 2020).

84. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (May 2020).

85. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (Aug. 2016).

86. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research* **45,** D840–D845 (Database issue Jan. 4, 2017).

87. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (Oct. 2015).

88. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38,** e164–e164 (Sept. 1, 2010).

89. Begay, R. L. *et al.* Role of Titin Missense Variants in Dilated Cardiomyopathy. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease* **4** (Nov. 13, 2015).

90. Fields, S. & Song, O.-k. A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246 (July 1989).

91. Satoh, M. *et al.* Structural Analysis of the Titin Gene in Hypertrophic Cardiomyopathy: Identification of a Novel Disease Gene. *Biochemical and Biophysical Research Communications* **262,** 411–417 (Aug. 1999).

92. Itoh-Satoh, M. *et al.* Titin Mutations as the Molecular Basis for Dilated Cardiomyopathy. *Biochemical and Biophysical Research Communications* **291,** 385–393 (Feb. 2002).

93. Matsumoto, Y. *et al.* Functional analysis of titin/connectin N2-B mutations found in cardiomyopathy. *Journal of Muscle Research and Cell Motility* **26,** 367–374 (June 29, 2006).

94. Arimura, T. *et al.* Cardiac Ankyrin Repeat Protein Gene (ANKRD1) Mutations in Hypertrophic Cardiomyopathy. *Journal of the American College of Cardiology* **54,** 334–342 (July 2009).

95. Rudloff, M. W., Woosley, A. N. & Wright, N. T. Biophysical characterization of naturally occurring titin M10 mutations. *Protein Science: A Publication of the Protein Society* **24,** 946–955 (June 2015).

96. Hastings, R. *et al.* Combination of Whole Genome Sequencing, Linkage, and Functional Studies Implicates a Missense Mutation in Titin as a Cause of Autosomal Dominant Cardiomyopathy With Features of Left Ventricular Noncompaction. *Circulation. Cardiovascular Genetics* **9,** 426–435 (Oct. 2016).

97. Zuo, J., Zhan, D., Xia, J. & Li, H. Single-Molecule Force Spectroscopy Studies of Missense Titin Mutations That Are Likely Causing Cardiomyopathy. *Langmuir* **37,** 12128–12137 (Oct. 19, 2021).

98. Anderson, B. R., Bogomolovas, J., Labeit, S. & Granzier, H. Single Molecule Force Spectroscopy on Titin Implicates Immunoglobulin Domain Stability as a Cardiac Disease Mechanism*. *Journal of Biological Chemistry* **288,** 5303–5315 (Feb. 22, 2013).

99. Kelly, C. & Gage, M. J. Protein Unfolding: Denaturant vs. Force. *Biomedicines* **9,** 1395 (Oct. 5, 2021).

100. Bogomolovas, J. *et al.* Exploration of pathomechanisms triggered by a single-nucleotide polymorphism in titin's I-band: the cardiomyopathy-linked mutation T2580I. *Open Biology* **6,** 160114 (Sept. 1, 2016).

101. Bertz, M., Wilmanns, M. & Rief, M. The titin-telethonin complex is a directed, superstable molecular bond in the muscle Z-disk. *Proceedings of the National Academy of Sciences* **106,** 13307–133310 (Aug. 11, 2009).

102. Lee, E. H., Gao, M., Pinotsis, N., Wilmanns, M. & Schulten, K. Mechanical Strength of the Titin Z1Z2-Telethonin Complex. *Structure* **14,** 497–509 (Mar. 1, 2006).

103. Zou, P. *et al.* Palindromic assembly of the giant muscle protein titin in the sarcomeric Z-disk. *Nature* **439,** 229–233 (Jan. 2006).

104. Wadmore, K., Azad, A. J. & Gehmlich, K. The Role of Z-disc Proteins in Myopathy and Cardiomyopathy. *International Journal of Molecular Sciences* **22,** 3058 (Jan. 2021).

105. Thirumal Kumar, D. *et al.* Influence of V54M mutation in giant muscle protein titin: a computational screening and molecular dynamics approach. *Journal of Biomolecular Structure and Dynamics* **35,** 917–928 (Apr. 4, 2017).

106. Gerull, B. *et al.* Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nature Genetics* **30,** 201–204 (Feb. 2002).

107. Peddareddygari, L. R., Oberoi, K. & Grewal, R. P. Novel Titin Gene Mutation Causing Autosomal Dominant Limb-Girdle Muscular Dystrophy. *Cureus* **14,** e30550 (2022).

108. Vasli, N. *et al.* Next generation sequencing for molecular diagnosis of neuromuscular diseases. *Acta Neuropathologica* **124,** 273–283 (2012).

109. Linke, W. Stretching molecular springs: Elasticity of titin filaments in vertebrate striated muscle. *Histology and histopathology* **15,** 799–811 (Aug. 1, 2000).

110. Castelmur, E. v. *et al.* A regular pattern of Ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain. *Proceedings of the National Academy of Sciences* **105,** 1186–1191 (Jan. 29, 2008).

111. Labeit, S. & Kolmerer, B. Titins: Giant Proteins in Charge of Muscle Ultrastructure and Elasticity. *Science* **270,** 293–296 (Oct. 13, 1995).

112. Freiburg, A. *et al.* Series of Exon-Skipping Events in the Elastic Spring Region of Titin as the Structural Basis for Myofibrillar Elastic Diversity. *Circulation Research* **86,** 1114–1121 (June 9, 2000).

113. Taylor, M. *et al.* Genetic Variation in Titin in ARVC-Overlap Syndromes. *Circulation* **124,** 876–885 (Aug. 23, 2011).

114. Fleming, J. R., Rigden, D. J. & Mayans, O. The importance of chain context in assessing small nucleotide variants in titin: in silico case study of the I10-I11 tandem and its arrhythmic right ventricular cardiomyopathy linked position T2580. *Journal of Biomolecular Structure and Dynamics* **39,** 3480–3490 (July 3, 2021).

115. Liu, X. *et al.* [Titin gene mutations in Chinese patients with dilated cardiomyopathy]. *Zhonghua Xin Xue Guan Bing Za Zhi* **36,** 1066–1069 (Dec. 2008).

116. Roncarati, R. *et al.* Doubly heterozygous LMNA and TTN mutations revealed by exome sequencing in a severe form of dilated cardiomyopathy. *European Journal of Human Genetics* **21,** 1105–1111 (Oct. 2013).

117. Watanabe, K. *et al.* Molecular Mechanics of Cardiac Titin's PEVK and N2B Spring Elements *. *Journal of Biological Chemistry* **277,** 11549–11558 (Mar. 29, 2002).

118. Raskin, A. *et al.* A Novel Mechanism Involving Four-and-a-half LIM Domain Protein-1 and Extracellular Signal-regulated Kinase-2 Regulates Titin Phosphorylation and Mechanics *. *Journal of Biological Chemistry* **287,** 29273–29284 (Aug. 24, 2012).

119. Hessel, A. L. & Linke, W. A. Unraveling the mysteries of the titin–N2A signalosome. *Journal of General Physiology* **153** (June 22, 2021).

120. Bennett, P. M., Hodkin, T. E. & Hawkins, C. Evidence That the Tandem Ig Domains near the End of the Muscle Thick Filament Form an Inelastic Part of the I-Band Titin. *Journal of Structural Biology* **120,** 93–104 (Oct. 1, 1997).

121. Van der Pijl, R. J. & Ottenheijm, C. A. Titin–N2A: More than a signaling node? *Journal of General Physiology* **153** (June 3, 2021).

122. Kelly, C., Pace, N., Gage, M. & Pfuhl, M. Solution NMR Structure of Titin N2A Region Ig Domain I83 and Its Interaction with Metal Ions. *Journal of Molecular Biology* **433,** 166977 (June 25, 2021).

123. Stronczek, C. *et al.* The N2A region of titin has a unique structural configuration. *Journal of General Physiology* **153** (Apr. 9, 2021).

124. Tsiros, C., Punch, E., Schaffter, E., Apel, S. & Gage, M. J. Identification of the domains within the N2A region of titin that regulate binding to actin. *Biochemical and Biophysical Research Communications* **589,** 147–151 (Jan. 22, 2022).

125. Van der Pijl, R. *et al.* Muscle ankyrin repeat protein 1 (MARP1) locks titin to the sarcomeric thin filament and is a passive force regulator. *Journal of General Physiology* **153** (July 5, 2021).

126. Costa, S., Pons, E., Medeiros-Domingo, A. & Saguner, A. M. Clinical impact of low coverage in whole-exome genetic testing in the assessment of familial arrhythmogenic right ventricular cardiomyopathy: a case report. *European Heart Journal: Case Reports* **5,** ytab111 (June 17, 2021).

127. Sasaki, R. *et al.* TTN missense variants in two siblings with asymmetric facial and limb weakness. *Journal of the Neurological Sciences* **415,** 116885 (Aug. 15, 2020).

128. Fitts, R. H. The cross-bridge cycle and skeletal muscle fatigue. *Journal of Applied Physiology* **104,** 551–558 (Feb. 2008).

129. Park-Holohan, S.-J. *et al.* Stress-dependent activation of myosin in the heart requires thin filament activation and thick filament mechanosensing. *Proceedings of the National Academy of Sciences* **118,** e2023706118 (Apr. 20, 2021).

130. Hedberg, C. *et al.* Hereditary myopathy with early respiratory failure is associated with misfolding of the titin fibronectin III 119 subdomain. *Neuromuscular Disorders* **24,** 373–379 (May 1, 2014).

131. Ohlsson, M. *et al.* Hereditary myopathy with early respiratory failure associated with a mutation in A-band titin. *Brain* **135,** 1682–1694 (June 1, 2012).

132. Pfeffer, G. *et al.* Titin mutation segregates with hereditary myopathy with early respiratory failure. *Brain* **135,** 1695–1713 (June 2012).

133. Palmio, J. *et al.* Hereditary myopathy with early respiratory failure: occurrence in various populations. *Journal of Neurology, Neurosurgery & Psychiatry* **85,** 345–353 (Mar. 1, 2014).

134. Toro, C. *et al.* Exome sequencing identifies titin mutations causing hereditary myopathy with early respiratory failure (HMERF) in families of diverse ethnic origins. *BMC Neurology* **13,** 29 (Mar. 20, 2013).

135. Yue, D. *et al.* New disease allele and de novo mutation indicate mutational vulnerability of titin exon 343 in hereditary myopathy with early respiratory failure. *Neuromuscular Disorders* **25,** 172–176 (Feb. 2015).

136. Bugiardini, E. *et al.* The Diagnostic Value of MRI Pattern Recognition in Distal Myopathies. *Frontiers in Neurology* **9** (2018).

137. Morais, J. *et al.* Titinopathy, an atypical respiratory failure. *BMJ Case Reports* **13,** e235378 (Sept. 9, 2020).

138. Izumi, R. *et al.* Exome sequencing identifies a novel TTN mutation in a family with hereditary myopathy with early respiratory failure. *Journal of Human Genetics* **58,** 259–266 (May 2013).

139. Sano, Y. *et al.* A Japanese Patient with Hereditary Myopathy with Early Respiratory Failure Due to the p.P31732L Mutation of Titin. *Internal Medicine,* 7733–21 (2022).

140. Liu, G. *et al.* Novel missense variant in TTN cosegregating with familial atrioventricular block. *European Journal of Medical Genetics* **63,** 103752 (Mar. 1, 2020).

141. Yu, M. *et al.* Novel TTN mutations and muscle imaging characteristics in congenital titinopathy. *Annals of Clinical and Translational Neurology* **6,** 1311–1318 (2019).

142. Evilä, A. *et al.* Targeted Next-Generation Sequencing Reveals Novel TTN Mutations Causing Recessive Distal Titinopathy. *Molecular Neurobiology* **54,** 7212–7223 (Nov. 1, 2017).

143. Mihailov, E. *et al.* Whole-exome sequencing identifies a potential TTN mutation in a multiplex family with inguinal hernia. *Hernia* **21,** 95–100 (2017).

144. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* **0 7,** Unit7.20 (Jan. 2013).

145. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31,** 3812–3814 (July 1, 2003).

146. Lopez-Bravo, A., Roche-Bueno, J. C., Romera-López, A. & Larrode-Pellicer, P. A novel TTN variant in a patient with distal myopathy of lower limbs and dilated cardiomyopathy. *Neurología (English Edition)* **36,** 721–723 (Nov. 1, 2021).

147. Peled, Y. *et al.* Titin Mutation in Familial Restrictive Cardiomyopathy. *International Journal of Cardiology* **171,** 24–30 (Jan. 2014).

148. Fukuzawa, A. *et al.* Interactions with titin and myomesin target obscurin and obscurin-like 1 to the M-band – implications for hereditary myopathies. *Journal of Cell Science* **121,** 1841–1851 (June 1, 2008).

149. Van den Bergh, P. Y. K. *et al.* Tibial muscular dystrophy in a Belgian family. *Annals of Neurology* **54,** 248–251 (2003).

150. Pollazzon, M. *et al.* The first Italian family with tibial muscular dystrophy caused by a novel titin mutation. *Journal of Neurology* **257,** 575–579 (Apr. 1, 2010).

151. Udd, B., Vihola, A., Sarparanta, J., Richard, I. & Hackman, P. Titinopathies and extension of the M-line mutation phenotype beyond distal myopathy and LGMD2J. *Neurology* **64,** 636–642 (Feb. 22, 2005).

152. Zheng, W. *et al.* Identification of a Novel Mutation in the Titin Gene in a Chinese Family with Limb-Girdle Muscular Dystrophy 2J. *Molecular Neurobiology* **53,** 5097–5102 (Oct. 1, 2016).

153. Lange, S., Edström, L., Udd, B. & Gautel, M. Reply: Hereditary myopathy with early respiratory failure is caused by mutations in the titin FN3 119 domain. *Brain* **137,** e279 (June 1, 2014).

154. Luo, X.-h., Zhu, R., Chen, Q., Shi, P.-h. & Na, L.-s. Early Diagnosis of Abnormal Left Ventricular Systolic Functions of Rare Pathogenic Titin Mutation Gene Carriers in FHCM by Three-Dimensional Speckle Tracking Echocardiography Combined with Gene Detection. *International Journal of Clinical Practice* **2022,** 3415545 (Oct. 11, 2022).

155. Tharp, C. A., Haywood, M. E., Sbaizero, O., Taylor, M. R. G. & Mestroni, L. The Giant Protein Titin's Role in Cardiomyopathy: Genetic, Transcriptional, and Post-translational Modifications of TTN and Their Contribution to Cardiac Disease. *Frontiers in Physiology* **10,** 1436 (2019).

156. Cazorla, O. *et al.* Differential expression of cardiac titin isoforms and modulation of cellular stiffness. *Circulation Research* **86,** 59–67 (Jan. 7, 2000).

157. Trombitás, K., Wu, Y., Labeit, D., Labeit, S. & Granzier, H. Cardiac titin isoforms are coexpressed in the half-sarcomere and extend independently. *American Journal of Physiology. Heart and Circulatory Physiology* **281,** H1793–1799 (Oct. 2001).

158. Savarese, M. *et al.* The complexity of titin splicing pattern in human adult skeletal muscles. *Skeletal Muscle* **8,** 11 (Mar. 29, 2018).

159. Opitz, C. A., Leake, M. C., Makarenko, I., Benes, V. & Linke, W. A. Developmentally Regulated Switching of Titin Size Alters Myofibrillar Stiffness in the Perinatal Heart. *Circulation Research* **94,** 967–975 (Apr. 16, 2004).

160. Neagoe, C. *et al.* Titin Isoform Switch in Ischemic Human Heart Disease. *Circulation* **106,** 1333–1341 (Sept. 10, 2002).

161. Makarenko, I. *et al.* Passive Stiffness Changes Caused by Upregulation of Compliant Titin Isoforms in Human Dilated Cardiomyopathy Hearts. *Circulation Research* **95,** 708–716 (Oct. 1, 2004).

162. Nagueh, S. F. *et al.* Altered Titin Expression, Myocardial Stiffness, and Left Ventricular Function in Patients With Dilated Cardiomyopathy. *Circulation* **110,** 155–162 (July 13, 2004).

163. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nature Communications* **13,** 3895 (July 6, 2022).

164. Savarese, M., Sarparanta, J., Vihola, A., Udd, B. & Hackman, P. Increasing Role of Titin Mutations in Neuromuscular Disorders. *Journal of Neuromuscular Diseases* **3,** 293–308 (Aug. 30, 2016).

165. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **44,** D862–D868 (Database issue Jan. 4, 2016).

166. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46,** D1062–D1067 (D1 Jan. 4, 2018).

167. Lopes, L. R. *et al.* Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *Journal of Medical Genetics* **50,** 228–239 (Apr. 1, 2013).

168. Yeo, Y., Park, J. E. & Kwon, H. S. A Novel TTN Gene Variant c.95136T>G (p.Cys31712Trp) and Associated Clinical Characteristics in a Family With Suspected Hereditary Myopathy With Early Respiratory Failure. *Annals of Laboratory Medicine* **41,** 604–607 (Nov. 1, 2021).

169. Su, K., Mayans, O., Diederichs, K. & Fleming, J. R. Pairwise sequence similarity mapping with PaSiMap: Reclassification of immunoglobulin domains from titin as case study. *Computational and Structural Biotechnology Journal* **20,** 5409–5419 (Sept. 26, 2022).

170. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (Aug. 2021).

171. *AlphaFold Protein Structure Database* `https://alphafold.ebi.ac.uk/` (2022).

172. Cavallo, L., Kleinjung, J. & Fraternali, F. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Research* **31,** 3364–3366 (July 1, 2003).

173. González-Pérez, A. & López-Bigas, N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics* **88,** 440–449 (Apr. 8, 2011).

174. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30,** 335–342 (Feb. 1, 2014).

175. Pires, D. E., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research* **42,** W314–W319 (Web Server issue July 1, 2014).

176. Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N. & Bahar, I. Rhapsody: predicting the pathogenicity of human missense variants. *Bioinformatics* **36,** 3084–3092 (May 1, 2020).

177. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* **12,** 103 (Dec. 2, 2020).

178. *Jmol: an open-source browser-based HTML5 viewer and stand-alone Java viewer for chemical structures in 3D* `http://jmol.sourceforge.net/` (2019).

179. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Research* **46,** W200–W204 (W1 July 2, 2018).

180. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19,** 679–682 (June 2022).

181. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35,** 1026–1028 (Nov. 2017).

182. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science : A Publication of the Protein Society* **27,** 293–315 (Jan. 2018).

183. Sobolev, O. V. *et al.* A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. *Structure* **28,** 1249–1258.e2 (Nov. 3, 2020).

184. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics* **99,** 877–885 (Oct. 6, 2016).

185. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nature Communications* **12,** 510 (Jan. 21, 2021).

186. Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66,** 630–645 (Feb. 15, 2007).

187. Kleinjung, J. & Fraternali, F. POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Research* **33,** W342–W346 (Web Server issue July 1, 2005).

188. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods* **11,** 801–807 (Aug. 2014).

189. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

190. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302,** 205–217 (Sept. 8, 2000).

191. Pei, J., Kim, B.-H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research* **36,** 2295–2300 (Apr. 2008).

192. Yagawa, K. *et al.* Structural basis for unfolding pathway-dependent stability of proteins: Vectorial unfolding versus global unfolding. *Protein Science : A Publication of the Protein Society* **19,** 693–702 (Apr. 2010).

193. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47,** D886–D894 (D1 Jan. 8, 2019).

194. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* **34,** 57–65 (Jan. 2013).

195. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics* **48,** 1581–1586 (Dec. 2016).

196. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24,** 2125–2137 (Apr. 15, 2015).

197. Samocha, K. E. *et al. Regional missense constraint improves variant deleteriousness prediction* June 12, 2017.

198. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39,** e118 (Sept. 1, 2011).

199. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods* **11,** 361–362 (Apr. 2014).

200. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications* **11,** 5918 (Nov. 20, 2020).

201. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics* **50,** 1161–1170 (Aug. 2018).

202. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE* **7,** e46688 (Oct. 8, 2012).

203. Cao, H., Wang, J., He, L., Qi, Y. & Zhang, J. Z. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *Journal of Chemical Information and Modeling* **59,** 1508–1514 (Apr. 22, 2019).

204. Lai, J., Yang, J., Gamsiz Uzun, E. D., Rubenstein, B. M. & Sarkar, I. N. LYRUS: a machine learning model for predicting the pathogenicity of missense variants. *Bioinformatics Advances* **2,** vbab045 (Jan. 1, 2022).

205. Van Gunsteren, W. F. *et al.* Biomolecular Modeling: Goals, Problems, Perspectives. *Angewandte Chemie International Edition* **45,** 4064–4092 (2006).

206. Ponzoni, L. & Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings of the National Academy of Sciences of the United States of America* **115,** 4164–4169 (Apr. 17, 2018).

207. Van Gunsteren, W. F. & Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition in English* **29,** 992–1023 (1990).

208. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (Oct. 5, 1990).

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

210. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46,** W296–W303 (W1 July 2, 2018).

211. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27,** 343–350 (Feb. 1, 2011).

212. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2,** 19–25 (Sept. 1, 2015).

213. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98,** 10089–10092 (June 15, 1993).

214. Essmann, U. *et al.* A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103,** 8577 (Aug. 31, 1998).

215. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **11,** 3696–3713 (Aug. 11, 2015).

216. Mark, P. & Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *The Journal of Physical Chemistry A* **105,** 9954–9960 (Nov. 1, 2001).

217. *User contributions — GROMACS webpage https://www.gromacs.org documentation* `https://www.gromacs.org/user_contributions.html` (2023).

218. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **4,** 116–122 (Jan. 1, 2008).

219. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126,** 014101 (Jan. 7, 2007).

220. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81,** 3684–3690 (Oct. 15, 1984).

221. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52,** 7182–7190 (Dec. 1981).

222. Alexandrov, N. N., Takahashi, K. & Go, N. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *Journal of Molecular Biology* **225,** 5–9 (May 5, 1992).

223. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. & Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry* **16,** 273–284 (1995).

224. Bondi, A. van der Waals Volumes and Radii. *The Journal of Physical Chemistry* **68,** 441–451 (Mar. 1, 1964).

225. *R: The R Project for Statistical Computing* `https://www.r-project.org/` (2019).

226. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *The R Journal* **7,** 4 (2015).

227. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24,** 417–441 (1933).

228. Grant, B. J., Skjærven, L. & Yao, X.-Q. The Bio3D packages for structural bioinformatics. *Protein Science* **30,** 20–30 (2021).

229. Pandini, A., Fornili, A. & Kleinjung, J. Structural alphabets derived from attractors in conformational space. *BMC Bioinformatics* **11,** 97 (Feb. 20, 2010).

230. Macpherson, J. A. *et al.* Functional cross-talk between allosteric effects of activating and inhibiting ligands underlies PKM2 regulation. *eLife* **8,** e45068 (July 2, 2019).

231. Gracia Carmona, O. *AlloHubPy* July 8, 2022.

232. Pandini, A., Fornili, A., Fraternali, F. & Kleinjung, J. GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinformatics* **29,** 2053–2055 (Aug. 15, 2013).

233. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18,** 50–60 (Mar. 1947).

234. Sunyaev, S. R. *et al.* PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering, Design and Selection* **12,** 387–394 (May 1, 1999).

235. Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H. A. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry* **4,** 23–55 (Feb. 1, 1985).

236. Ittisoponpisan, S. *et al.* Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *Journal of Molecular Biology* **431,** 2197–2212 (May 17, 2019).

237. Kleinjung, J., Fraternali, F., Martin, S. R. & Bayley, P. M. Thermal unfolding simulations of apo-calmodulin using leap-dynamics. *Proteins: Structure, Function, and Bioinformatics* **50,** 648–656 (2003).

238. Zeiske, T., Stafford, K. A. & Palmer, A. G. Thermostability of Enzymes from Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **12,** 2489–2492 (June 14, 2016).

239. Bekker, G.-J., Ma, B. & Kamiya, N. Thermal stability of single-domain antibodies estimated by molecular dynamics simulations. *Protein Science : A Publication of the Protein Society* **28,** 429–438 (Feb. 2019).

240. Allison, J. R. Computational methods for exploring protein conformations. *Biochemical Society Transactions* **48,** 1707–1724 (Aug. 5, 2020).

241. Qi, R., Wei, G., Ma, B. & Nussinov, R. Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide

Aggregation and Self-Assembly Example. *Methods in molecular biology (Clifton, N.J.)* **1777,** 101–119 (2018).

242. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation* **11,** 3584–3595 (Aug. 11, 2015).

243. Zhou, H., Wang, F. & Tao, P. t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. *Journal of Chemical Theory and Computation* **14,** 5499–5510 (Nov. 13, 2018).

244. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* Sept. 17, 2020. arXiv: `1802.03426[cs,stat]`.

245. Trozzi, F., Wang, X. & Tao, P. UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study. *The Journal of Physical Chemistry B* **125,** 5022–5034 (May 20, 2021).

246. Marabotti, A., Scafuri, B. & Facchiano, A. Predicting the stability of mutant proteins by computational approaches: an overview. *Briefings in Bioinformatics* **22,** bbaa074 (May 1, 2021).

247. Huang, P. *et al.* Evaluating Protein Engineering Thermostability Prediction Tools Using an Independently Generated Dataset. *ACS Omega* **5,** 6487–6493 (Mar. 31, 2020).

248. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds Hastie, T., Tibshirani, R. & Friedman, J.) 219–259 (Springer, New York, NY, 2009).

249. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **18,** 1–5 (2017).

250. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds Hastie, T., Tibshirani, R. & Friedman, J.) 587–604 (Springer, New York, NY, 2009).

251. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds Hastie, T., Tibshirani, R. & Friedman, J.) 295–336 (Springer, New York, NY, 2009).

252. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds Hastie, T., Tibshirani, R. & Friedman, J.) 101–137 (Springer, New York, NY, 2009).

253. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds Hastie, T., Tibshirani, R. & Friedman, J.) 417–458 (Springer, New York, NY, 2009).

254. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

255. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21,** 6 (Jan. 2, 2020).

256. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89,** 10915–10919 (Nov. 15, 1992).

257. Munro, D. & Singh, M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* **36,** 5322–5329 (Dec. 1, 2020).

258. Laddach, A. *Anatomy of Missense Variants in Health and Disease: Towards Better Impact Prediction with a Focus on Titinopathies* Doctor of Philosophy (King's College London, Randall Centre for Cell and Molecular Biophysics, Oct. 2018).

259. Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Reports* **38,** 110207 (Jan. 11, 2022).

260. Montanucci, L., Martelli, P. L., Ben-Tal, N. & Fariselli, P. A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics* **35,** 1513–1517 (May 1, 2019).

261. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, New York, NY, 2009).

262. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (Aug. 20, 2021).

263. Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D. & Gromiha, M. M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Research* **49,** D420–D424 (D1 Jan. 8, 2021).

264. Xavier, J. S. *et al.* ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Research* **49,** D475–D479 (D1 Jan. 8, 2021).

265. Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human Genomics* **11,** 10 (May 16, 2017).

266. Zaucha, J., Heinzinger, M., Tarnovskaya, S., Rost, B. & Frishman, D. Family-specific analysis of variant pathogenicity prediction tools. *NAR Genomics and Bioinformatics* **2,** lqaa014 (Feb. 28, 2020).

267. Rivas-Pardo, J. A. *et al.* Work Done by Titin Protein Folding Assists Muscle Contraction. *Cell Reports* **14,** 1339–1347 (Feb. 16, 2016).

268. Bianco, P., Reconditi, M., Piazzesi, G. & Lombardi, V. Is muscle powered by springs or motors? *Journal of Muscle Research and Cell Motility* **37,** 165–167 (Oct. 1, 2016).

269. Eckels, E. C., Tapia-Rojo, R., Rivas-Pardo, J. A. & Fernández, J. M. The Work of Titin Protein Folding as a Major Driver in Muscle Contraction. *Annual Review of Physiology* **80,** 327–351 (2018).

270. Eckels, E. C., Haldar, S., Tapia-Rojo, R., Rivas-Pardo, J. A. & Fernández, J. M. The Mechanical Power of Titin Folding. *Cell reports* **27,** 1836–1847.e4 (May 7, 2019).

271. Brockwell, D. J. *et al.* The Effect of Core Destabilization on the Mechanical Resistance of I27. *Biophysical Journal* **83,** 458–472 (July 1, 2002).

272. Oroz, J. *et al.* The Y9P Variant of the Titin I27 Module: Structural Determinants of Its Revisited Nanomechanics. *Structure (London, England: 1993)* **24,** 606–616 (Apr. 5, 2016).

273. Joshi, T. *et al.* Interdomain linkers tailor the stability of immunoglobulin repeats in polyproteins. *Biochemical and Biophysical Research Communications* **550,** 43–48 (Apr. 23, 2021).

274. Yu, M., Lu, J.-H., Le, S. & Yan, J. Unexpected Low Mechanical Stability of Titin I27 Domain at Physiologically Relevant Temperature. *The Journal of Physical Chemistry Letters,* 7914–7920 (Aug. 12, 2021).

275. Alegre-Cebollada, J. *et al.* S-Glutathionylation of Cryptic Cysteines Enhances Titin Elasticity by Blocking Protein Folding. *Cell* **156,** 1235–1246 (Mar. 13, 2014).

276. Bianco, P., Mártonfalvi, Z., Naftz, K., Kőszegi, D. & Kellermayer, M. Titin Domains Progressively Unfolded by Force Are Homogenously Distributed along the Molecule. *Biophysical Journal* **109,** 340–345 (July 2015).

277. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599,** 91–95 (Nov. 2021).

278. Lin, W., Wells, J., Wang, Z., Orengo, C. & Martin, A. C. R. *VariPred: Enhancing Pathogenicity Prediction of Missense Variants Using Protein Language Models* Mar. 20, 2023.

279. Degiacomi, M. T. Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure* **27,** 1034–1040.e3 (June 4, 2019).

280. Ramaswamy, V. K., Musson, S. C., Willcocks, C. G. & Degiacomi, M. T. Deep Learning Protein Conformational Space with Convolutions and Latent Interpolations. *Physical Review X* **11,** 011052 (Mar. 15, 2021).