universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Comprehensive analysis and prediction of pathogenic missense variants in human filamin C“

verfasst von / submitted by

Michael Nagy, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. Kristina Djinovic-Carugo

Acknowledgements

I would like to particularly thank my supervisor Kristina Djinovic-Carugo, who came up with the idea of our project during the Covid lockdown. It was during this period of time where I realized that I don't see myself in the future working in the wet lab. Instead, she introduced me together with Oliviero Carugo towards bioinformatics and finally even towards data science in the form of machine learning techniques. Although I lacked any prior knowledge about computing or bioinformatics due to my biology major, she set her trust in me and guided our project step by step towards the final end result, AMIVA-F. In hindsight, this whole project appears gigantic and beyond the scope of an usual master thesis, but now we can proudly look back after reaping the fruits of our labour. It was her trust, patience and also the urge to not disappoint her expectations in me that motivated me to learn on the fly new topics like programming in python and introductions to machine learning even without any prior knowledge, which immediately found purpose in the AMIVA-F bioinformatics tool.
There were times where I first doubted if I would manage to solve all computational issues that arose, but fortunately hard work paid off.
It took many meetings, discussing potential strategies and candidate parameters which could correlate with disease outcome until we managed to reach the final state of AMIVA-F. On the one hand, Kristina was contributing more with ideas towards chemical aspects of disease outcome, while on the other hand, Oliviero hinted at potential flaws from a statistical side regarding interdependence of parameters within themselves. After countless hours of ups and downs during weeks of trials with multiple combinations of different parameters, algorithms and iterative refinements which lead nowhere, we finally managed to find the best (to my knowledge) setup which hopefully will be of most use for clinical personal utilizing AMIVA-F.

Additional thanks belongs especially to Oliviero, who introduced me to machine learning methods and data science. It was his idea to try machine learning based techniques to find correlation within our data and he also introduced me to professor Ian Witten, a valuable teacher with an even more valuable online course for beginners and intermediates for free. It was also his book, Data Mining: Practical machine learning tools and techniques which first sent me back to my high school linear algebra books but then allowed me to elevate my work to a higher level and sparked my interest in data science.

From the side of drafting the thesis, and (hopefully) soon following paper of AMIVA-F, I want to emphasize on the help of Julius Kostan, who provided many nicely illustrated figures, proof-read countless pages and even came up with the name of AMIVA-F. Not only did his contribution upgrade our project, it also hinted out weak points during drafting and helped to improve our draft massively. Even though we sat countless hours in the late evenings preparing topology figures and changing views, colours and even whole arrangements of figures, I think the final result clearly reflects the efforts.

Not forgotten, Georg Mlynek, my former supervisor and valuable mentor in the lab, who provided crystal structures of FLNc-Ig domains and also gave some valuable input at the beginning regarding bioinformatic analysis of inter-domain interactions. Even if he no longer belongs to the Djinovic group, I wish him only the best for his work at Boku. In terms of supervision, his educational approach towards students which relies partially towards independent work suited me well and I managed to learn a lot from him.

I would also like to thank Dalibor Milic, who reached out for a collaboration with a team of students from HTL Spengergasse.
They are a team of young ambitious computational engineers consisting of Stefan Burghuber, Antonia Schwarz, Bernd Cala and Benjamin Wittmann, who now work on a web server based version of AMIVA-F, thereby increasing the accessibility of AMIVA-F towards the audience of non technically versatile people.

Table of contents

Abstract:

In human, mutations in Filamin C (FLNc), the heart specific isoform of filamins, are associated with many different genetically related cardiac diseases.

First, we showed that known pathogenically associated mutations of FLNc are correlated with changes in thermodynamic stability changes, as measured by differential scanning calorimetry (DSC).
To further investigate this behaviour, we crystallized FLNc Ig14-15 and two mutations, Ser1624Leu and Gly1676Arg. We found that a highly conserved PXSPF motif correlates with high pathogenic mutational frequency and that interdomain interactions need to be taken into account when judging the pathogenicity of mutations.

Combining literature and our novel data, we designed a machine learning based bioinformatics tool, AMIVA-F (Analysis of Mutations In Variants of Filamin-C), which makes predictions based on biophysical and structural parameters regarding pathogenicity of single point missense mutations in Filamin C. AMIVA-F is freely available and stands out with its high accuracy of nearly 80%, outclassing the most prominently used general protein stability predictors. Its cutting edge against its competitors gains AMIVA-F by incorporating interdomain interactions, as well as taking into account known post translational modifications for its prediction.

For users, AMIVA-F works fully automated and does not require external resources nor deeper knowledge and can be run on any personal computer, working on Linux, MacOS and Windows.
With that in mind, AMIVA-F seeks to aid medical personal in clinical decisions, where preventive diagnosis of a potential pathogenic variant can drastically increase patients life expectancy and prevent surprising sudden cardiac death.

Abstract (german):

Mutationen im menschlichen Filamin C (FLNc), der Herz-spezifischen Isoform der Filamine, sind oftmals mit vielen verschiedenen genetisch vererbten Herzerkrankungen assoziiert.
Mithilfe der thermischen Differentialkalorimetrie zeigten wir die Korrelation zwischen Mutationen des FLNc und der Veränderung in der thermodynamischen Proteinstabilität auf.
Für weitere Evidenz des vorher genannten Phänomens ermittelten wir die Struktur von FLNc Ig1415 wildtyp und Ig1415 Ser1624Leu sowie Gly1676Arg mithilfe der Röntgenkristallographie.
Es ergab sich eine hohe strukturelle Konservierung eines PXSPF Motivs, welches mit hoher pathologischer Mutationsfrequenz korreliert. Ebenso stellte sich heraus, dass Zwischendomain-Interaktionen in die Berücksichtigung der Abschätzung von krankmachenden Mutationen herangezogen werden müssen.

Die experimentell ermittelten Daten wurden mit den bereits in der Literatur vorhandenen Daten zusammengezogen und dienten als Grundlage des Trainings für unseren Vorhersagealgorithmus namens AMIVA-F (Analyse von Mutationen in Varianten von Filamin C). Jene Vorhersagen im Bezug auf krankmachende Punktmutationen sind basierend auf verschiedenen Biophysikalischen sowie Strukturellen Parametern und erzielen eine Vorhersagegenauigkeit von knapp 80%. Somit schneidet AMIVA-F besser als andere gleichwertige allgemeine Konkurrenzprogramme ab, welche allgemeine Proteinstabilitätsveränderungen vorhersagen. Die Besonderheit von AMIVA-F gegenüber den vorhergenannten alternativen Programmen erreicht AMIVA-F durch Berücksichtigung von Post-translationalen chemischen Modifikationen und Interdomain Interaktionen.
Für Benutzer steht ein vollkommen automatisiertes AMIVA-F zur Verfügung welches sowohl ohne externe Ressourcen als auch vertiefendes Fachwissen auskommt. AMIVA-F funktioniert unter Linux, MacOS und Windows Betriebssystemen.

Als Zielgruppe von AMIVA-F sind in erster Linie medizinisches Personal in der Präventivdiagnostik gedacht, in welcher das schnelle Erkennen von potenziellen pathogenen Mutationen die Lebenserwartung der betroffenen Patienten drastisch erhöhen und etwaigen plötzlichen Herztoden zuvorkommen kann.

1) Introduction

1.1) Introduction to striated muscle and muscular organization

In human, around 40% of total body weight is comprised by skeletal muscles, which itself contains up to 70% of all body proteins (Frontera and Ochala 2015). Skeletal muscle fulfils multiple purposes, ranging from mechanical work which is done by utilizing chemical energy to generate force and power, maintenance of body temperature through production of heat, storage of carbohydrates and amino acids and many more metabolic functions.

For most people, the perception of muscles is essentially just the topmost hierarchical level of the muscle, which itself contains bundles of muscle fibers, fat and connective tissue (**Figure 1A top**). The size of the muscle is mostly determined by the number and size of muscle fibers, although there are pathological phenotypes where infiltration of connective tissue and fat leads to increases in size (Javan, Horvath et al. 2013).
In order to accomplish above mentioned metabolic and mechanical tasks, proteins are required.
Muscle fibers are postmitotic and multinucleated, where each nucleus within a muscle fiber governs the expression of specific proteins at specific regions in the cell (Frontera and Ochala 2015). Given the postmitotic state of muscle fibers, satellite cells, the adult stem cells of skeletal muscles are required to form new muscle fibers. When activated by myogenic factors, those stem cells proliferate and form new fibers (Hikida 2011).
The muscle fiber itself consists of many myofibrils (**Figure 1A middle**), which are surrounded by the sarcolemma, a membrane like structure that wraps around the myofibrils.
At the heart of the myofibril, the sarcomere stands as the basic contractile unit which enables muscles to produce force through contraction (**Figure 1A bottom**). In the middle of the sarcomere, the M band connects multiple sarcomeres in space through intermediate filaments (not shown) as well as link the sarcomere to the sarcolemma and T-tubles via obscurin (Gautel and Djinović-Carugo 2016).
Two of the most abundantly found proteins in the sarcomere are myosin and actin (**Figure 1A bottom**). These two proteins form large filamentous structures that are required for force generation.
On the one side, myosin is the major part of the thick filament and acts as the molecular motor which is required for force production. On the other side, actin filaments decorated with troponin and tropomyosin comprise the thin filament where the myosin heads can attach. In order to contract the sarcomere, myosin heads require ATP which gets hydrolysed in a concise manner in order to generate a power stroke, "walking" along the actin filament in a process termed sliding (Frontera and Ochala 2015).
For efficient force production, multiple sarcomeres need to act in concert together. Not only are sarcomeres connected laterally through the M-bands, but also through their boundaries at the Z-discs (**Figure 1A bottom**).
The Z-disc effectively crosslinks actin filaments through α-actinins N-terminal actin binding domain while its C-terminal region binds to Titin-Z repeats (Gautel and Djinović-Carugo 2016).
Furthermore, a plethora of proteins orchestrates additional crosslinking and capping at the Z-disc of the barbed end of the actin filaments, thereby preventing growth of filaments into the adjacent sarcomere (**Figure 1B**).
Fully matured Z-discs consist of hundreds of proteins acting in concert, which often regards them as one of the most complex macromolecular structures in biology (Zou, Pinotsis et al. 2006).

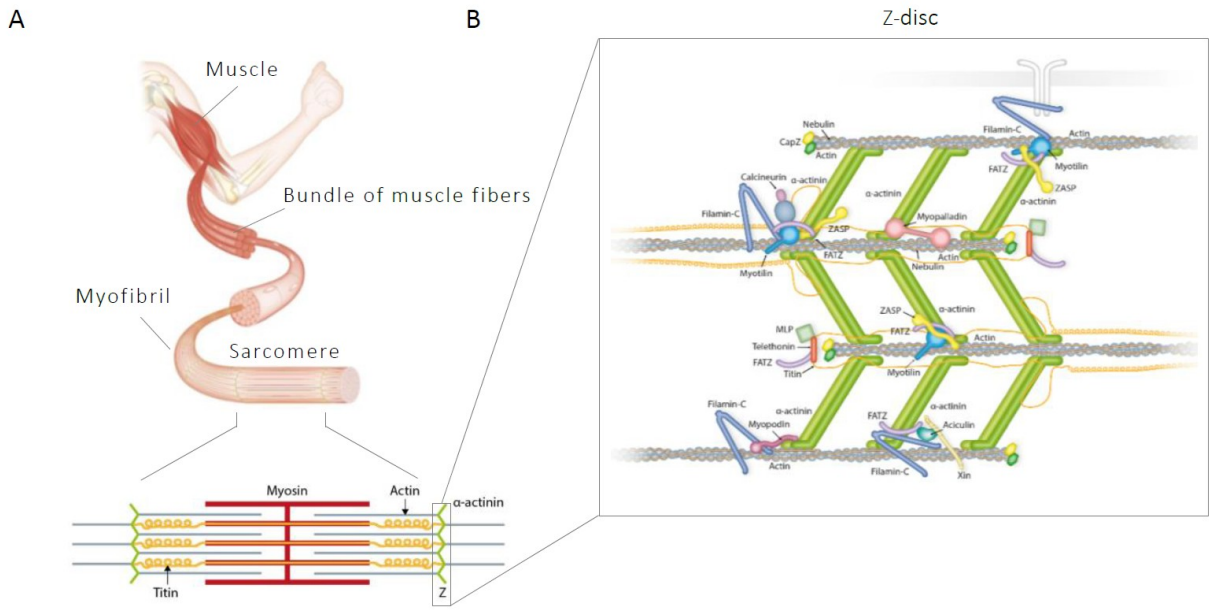


Figure 1. Hierarchical organization of striated muscles.
(A) The muscle consisting of multiple bundles of muscle fibers, fat and connective tissue. Each muscle fiber is composed of multiple myofibrils which by themselves are built out of repeating contractive units, the sarcomeres. This contraction is driven through the interplay between myosin and actin and requires ATP. The boundaries of the sarcomere are called the Z-discs and delineate each sarcomere from the neighbouring one.
(B) The Z-disc consisting of hundreds of proteins which are required to act in concert to fulfil a plethora of tasks, among those the necessity to prevent filaments to stray out into the next neighbouring sarcomere. This is done via capping proteins together with crosslinking between actin filaments in an antiparallel manner.

1.2) Filamins as important key player in the muscle

Filamin A (FlnA), B (FlnB) and Filamin C (FlnC) are important structural cross-linkers of actin filaments (F-actin), with FlnC being specific for both cardiac and skeletal muscle and localised in Z-discs, the sarcolemma, the myotendinous junction and the intercalated discs. Its principal role is to maintain the structural integrity of the sarcomere through cross-linking F-actin and anchoring sarcolemmal proteins to the cytoskeleton (Thompson, Chan et al. 2000).
FlnC interacts with a plethora of other proteins, either Z-disc components (e.g. myotilin, calsarcins, myopodin, LDB3/ZASP), signalling molecules or sarcolemma associated proteins (e.g. integrin ß1, sarcoglycan delta (reviewed in (Mao and Nakamura 2020)). Furthermore, FlnC interacts with the Xin actin-binding repeat-containing proteins (XIRP) and is recruited to sarcomeric lesions together with aciculin to accomplish its role in myofibril stabilization and early myofibril repair processes (Fujita, Mitsuhashi et al. 2012, Molt, Bührdel et al. 2014, Leber, Ruparelia et al. 2016, Reimann, Wiese et al. 2017).

1.3) Human Filamin C at the base of genetically inherited cardiac diseases

In a recent study, both global knockout and cardiac-specific knockout mice were embryonic lethal, while inducible cardiomyocyte-specific knockout in adult mice caused fulminant dilated cardiomyopathy, demonstrating that FlnC is critical in developing cardiomyocytes (Zhou, Chen et al. 2020). In humans, FlnC is dramatically redistributed in a large variety of muscle diseases (Sewry 2000, Thompson, Chan et al. 2000, Bönnemann, Brockmann et al. 2003, Claeys, van der Ven et al. 2009, Nilsson, Nissar et al. 2013). While pathogenic variants in the *FLNC* gene have been firstly reported to cause dominant myofibrillar and distal myopathy (Fürst, Goldfarb et al. 2013), *FLNC* has more recently emerged also as a cardiomyopathy gene with cardiac phenotypes, including hypertrophic cardiomyopathies (HCM), restrictive cardiomyopathies (RCM), arrhythmogenic cardiomyopathy (ACM) and dilated cardiomyopathy (DCM)

(Valdés-Mas, Gutiérrez-Fernández et al. 2014, Ortiz-Genga, Cuenca et al. 2016, Ader, De Groote et al. 2019, Verdonschot, Vanhoutte et al. 2020, Eden and Frey 2021, Wadmore, Azad et al. 2021).
Given the increasing number of *FLNC* variants potentially associated with cardio(myopathies), a bioinformatics tool is needed to assess their pathogenicity.

1.4) Architecture of human Filamin C

Filamins are composed of an N-terminal actin-binding domain (ABD) consisting of two calponin homology domains, followed by 24 immunoglobulin (Ig)-like domains (Ig1-Ig24) with two intervening calpain-sensitive hinges separating Ig15 and Ig16 (hinge 1), and Ig23 and Ig24 (hinge-2) (**Figure 2A**).

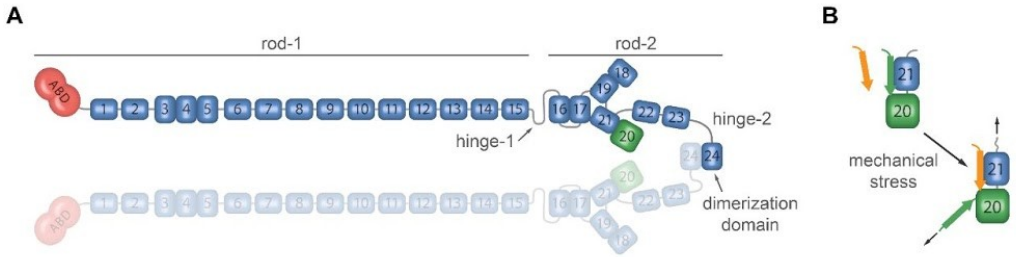


Figure 2. Domain structure and mechanosensing.
(**A**) Domain structure of FlnC dimer. ABD is depicted in red, Ig-domains in blue, domain 20 hosting the insertion region of 82 aa in green. Inter-and intra-molecular domain pairing is schematically presented. Hinge-1 is spliced out during myogenesis (Thompson, Chan et al. 2000, van der Flier, Kuikman et al. 2002). (**B**) Proposed mechanism of mechanosensing and ligand binding regulation (Lad, Kiema et al. 2007, Rognoni, Stigler et al. 2012, Rognoni, Möst et al. 2014). Under unstrained conditions β-strand A of Ig20 covers the ligand-binding site present in domain Ig21. Mechanical force will lead to domain pair opening, exposing therefore the cryptic binding site in Ig21 to interactions with binding partners (Lad, Kiema et al. 2007, Rognoni, Stigler et al. 2012, Rognoni, Möst et al. 2014).

The last Ig domain - Ig24 - is responsible for filamin dimerization, leading to a V-shaped homodimer (Nakamura, Osborn et al. 2007). Contrary to the other isoforms, FlnC contains an insertion of 82 amino acids in Ig20 (insertion region, IR), and its hinge-1 part is spliced out during myogenesis (Thompson, Chan et al. 2000, van der Flier, Kuikman et al. 2002). Filamin domains are not lined as pearls on a string but are organized with intricate inter-domain interactions (Lad, Kiema et al. 2007, Ruskamo, Gilbert et al. 2012, Tossavainen, Koskela et al. 2012, Sethi and Ylänne 2014, Seppälä, Bernardi et al. 2017) (**Figure 2A**), which are the basis of the force-/mechano-sensing and ligand binding of filamin, which was described for FlnA (Lad, Kiema et al. 2007, Ehrlicher, Nakamura et al. 2011, Rognoni, Stigler et al. 2012). Lad et al. (2007) showed that Ig20 and Ig21 form a dynamic domain pair, where the first β-strand of Ig20 binds to the subsequent domain Ig21, thus covering and auto inhibiting its transmembrane receptor-binding site (**Figure 2B**). This domain pair has a mechanosensor function (Ehrlicher, Nakamura et al. 2011, Rognoni, Stigler et al. 2012), whereby mechanical force transmitted through the cytoskeleton leads to the opening of the domain pair and an increase in receptor-binding affinity of up to 17-fold (Ehrlicher, Nakamura et al. 2011). Furthermore, the crystal structure of FlnA displays a *cis* Pro in front of the last β-strand of Ig20, and multiple-sequence alignment (**Supplementary Figure 1. Multiple Ig-domain sequence alignment and the conserved PXSPF motif**) shows that this *cis* proline residue is strictly conserved in all available 3D structures of filamin Ig domains in the context of also highly conserved PXSP motif (Rognoni, Möst et al. 2014).
Single-molecule mechanical studies showed that *cis-trans* isomerization modulates the force-sensing mechanism (Rognoni, Möst et al. 2014). Given sequence similarity of FlnC with FlnA, FlnC may mediate mechanotransduction as well. Moreover, the 82 amino acid insertion in Ig20, typical of FlnC, might ei-

ther modify the fine molecular details of mechanosensing or act as an interaction hub for Z-disc proteins (Mao and Nakamura 2020).

1.5) Aim of the studies

Considering the rise in number and diversity of cardiac diseases related due to FLNc, it seems unreasonable to solve the exact pathogenic mechanisms for all diseases. Therefore, we might ask if there is a common ground between diseases that can be exploited in developing a predictive tool, which in turn can then help to treat patients pre-emptively. In this thesis, we report an integrative approach, based on the findings of calorimetric experiments and structural alignments that shall aid in answering the following questions:

I) Are there any common features observed that are shared between mutations?

In order to address this question, we performed differential scanning calorimetry experiments with constructs of up to three FLNc Ig domains bearing single mutations. This results were then compared against the wild type construct of identical size lacking any mutations to base our conclusions on the effect of given mutations.

II) Can we find some conserved structurally important spots on the structure of Ig domains which correlate with pathogenic mutation frequency?

We structurally aligned all FLNc Ig-domains and observed the literature and experimentally derived known mutations. Ig domain 5, an x-ray crystallography solved experimental structure was used to superimpose mutations occurring in all 24 Ig-domains. The mutational frequency, as well as the conservation of the underlying amino acid were mapped onto Ig5 in order to analyse any specialties.

III) Can we develop a bioinformatics tool that can predict the effect of mutations based on structural, biophysical and chemical properties?

A total of 173 mutations were gathered and comprised of roughly 1/3 neutral and 2/3 pathogenic mutations. For the limited data available, we tried several types of algorithms to try and find some correlation with pathogenicity. To do so, we tried and defined multiple biophysical and biochemical parameters that can be linked to pathogenicity.

2) Materials and Methods

2.1) Structures and structural models of FlnC Ig-domains

The available 3D structures of 25 filamin domains that have been determined experimentally were downloaded from the Protein Data Bank (Bernstein, Koetzle et al. 1977, Berman, Westbrook et al. 2000, Burley, Berman et al. 2019) (**Table 1. Summary of 3D - structures used in algorithm training**). For domains where experimental 3D structures were unavailable, homology models were generated by using the Modeller package (Sali and Blundell 1993), based on available structures of FlnA or FlnB templates. A total of five structures per template were generated based on Modeller Automodel function, energy was minimized according to DOPE and GA341 scores and the lowest energy model was used for further computations (WHICH WAS systematically associated with a GA341 score of 1.00). Due to the presence of intrinsically disordered elements within Ig20 (Insertion region IR) and the hinge region between Ig 15 and Ig16, which separates rod 1 and rod 2 of FLNc (**Figure 2. Domain structure and mechanosensing**), we were unable to accurately model Ig 16 and Ig1921 with Modeller. We therefore resorted due to the absence of homologous structures in those two cases to I-tasser (Roy, Kucukural et al. 2010, Yang and Zhang 2015) models which were generated through threading.

Table 1. Summary of 3D - structures used in algorithm training

Experimental structures of FlnC domains, determined by X-ray crystallography or by NMR, and experimental structures of FlnA and FlnB domains used as templates to generate homology models, with Modeller, of FlnC domains. The resolution of the X-ray crystal structures and the percentage of sequence identity between template and target are provided.

Domain	Position	Template (pdb id/protein name)	Resolution (Å)	Seq. identity (%)
ABD	36-262	3hop/FlnA	2.3	85.8
Ig1	270-368	4b7l/FlnB	2.1	67.7
Ig2	370-468	4b7l/FlnB	2.1	37.4
Ig3	469-565	6ew1/FlnA	2.3	77.3
Ig4-5	568-760	3v8o/FlnC	2.8	na[(b)]
Ig6	759-861	4m9p/FlnA	1.7	38.8
Ig7	862-960	4m9p/FlnA	1.7	40.4
Ig8	961-1056	2dic/FlnB	na[(a)]	41.7
Ig9	1057-1149	2di9/FlnB	na[(a)]	82.8
Ig10	1150-1244	3rgh/FlnA	2.4	61.1

Ig11	1245-1344	2dib/FlnA	na[(a)]	60
Ig12	1345-1437	2dic/FlnB	na[(a)]	69.9
Ig13	1438-1533	2dj4/FlnB	na[(a)]	71.9
Ig14-15	1535-1758	Crystal structure reported here	1.8	na[(b)]
Ig16*	1759-1853	2d7n/FlnC	na[(a)]	na[(b)]
Ig17	1854-1946	2d7o/FlnC	na[(a)]	na[(b)]
Ig18	1947-2033	2k7q/FlnA	na[(a)]	na[(b)]
Ig19-21*	2036-2401	2j3s/FlnA	2.5	81.4*
Ig22	2398-2509	2D7P/FlnC	na[(a)]	na[(b)]
Ig23	2502-2598	2NQC/FlnC	2.1	na[(b)]
Ig24	2633-2725	1V05/FlnC	1.4	na[(b)]

[(a)] If no resolution is given, the structure was determined by NMR.
[(b)] The percentage of sequence identity between target and template is not provided for experimental structures determined either with X-ray crystallography or NMR.
* The model was generated via i-Tasser (Yang and Zhang 2015).

2.2) Mutational variants present in the training set

In order to distinguish between pathogenic and benign variants of FlnC mutations, we gathered a total of 173 mutations and split them into two classes.
The first class consisted of disease-related mutations which were already verified experimentally and were classified as "pathogenic" (**Supplementary Table 1. Pathogenic mutations dataset overview**).
For the second class of mutations, we selected missense mutations from the GnomAD database according to the criteria suggested by Karczewski et al. (Karczewski, Francioli et al. 2020), who found that pathogenic mutations tend to appear with lower frequencies, compared to neutral ones. Furthermore, the presence of homozygotes can give indication that a mutation is potentially benign. Non-pathogenic variants were selected according to the following criteria: (i) high allele count (at least 20, median is 60), (ii) allele frequency >> 10^{-3} (extrapolation of frequency in the population), and/or (iii) presence of homozygotes. Details on the selection of non-pathogenic variants can be found in (**Supplementary Table 2. SGnaomAD-negative Dataset**).

2.3) Machine learning algorithm (AMIVA-F)

We utilized WEKA (Witten, Frank et al. 2017) a workbench for machine learning tools and algorithms in order to set up our machine learning algorithm.
In order to benchmark our algorithm, we used ZeroR, an algorithm that always predicted the most common instance in a dataset and therefore defined a baseline accuracy, which is solely dependent on the underlying datasets and is algorithm independent. For our dataset, ZeroR determined a 62.4% baseline accuracy.

After scouting for potential classification algorithm types, we found the multilayer perceptron, a neural network algorithm to be a promising candidate. The initial default parameters were already able to classify better than the baseline accuracy, which lead us to further optimization.
The nodes in our network are all sigmoidal, except for classes with numeric input. In those cases, the output nodes become unthresholded linear units.
For our learning process we used backpropagation to calculate the gradients of our hyperparameters (weights and bias).
These gradients were then used to optimize our initialized weights and bias by stochastic gradient descent (SGD), a technique commonly used to find the global minimum of the cost function J(w, b) with w and b referring to the weights and bias. The minimization of the cost function J is equivalent to the minimization of the mean squared error (MSE) and thereby increases classification accuracy.
In order to overcome local minima during the gradient descent, we allowed the adjustment of the learning rate during optimization.
The iterative cycle of optimization and parameter update was then repeated until the hyperparameters (weights and bias) converged to their final values.
Ultimately, the final optimized algorithm was found to be a multilayer perceptron neural network with a weighted average true positive (TP) rate of ~80%. The exact metrics are shown in (**Table 2. Summary of the quality of predictions**). Pertinent details are shown in (**Supplementary Table 3. Technical details**).

Table 2. Summary of the quality of predictions.
Summary of the quality of the predictions of AMIVA-F.

Sensitivity (TP[a] / (TP + FN[b]))	**0.858**
Specificity (TN[c] / (TN + FP[d]))	0.689
Accuracy (TP+TN)/(TP+TN+FP+FN)	0.786
Precision (TP / (TP + FP))	0.795
F-measure[e] TP / (TP + ½ *(FP + FN))	0.788
ROC-Area under the curve (AUC)[f]	0.818

MCC[(g)]	0.560
$(TP*TN)(PF*FN)/[(TN+FN)(TN+FP)(TP+FN)(TP+FP)]^{1/2}$	

[(a)] TP = true positives: number of pathogenic variants that are predicted to be pathogenic.
[(b)] FN = false negatives: number of pathogenic mutations that are predicted to be non-pathogenic.
[(d)] FP = false positives: number of non-pathogenic mutations that are predicted to be pathogenic.
[(c)] TN = true negatives: number of non-pathogenic mutations that are predicted to be non-pathogenic.
[(e)] The F-measure is a measure of the accuracy of a model on a dataset. It takes into account precision and recall of the model and is defined as the harmonic mean of precision and recall. The highest possible F-measure is 1 indicating perfect recall and precision and the lowest possible value is 0 if either recall or precision is 0.
[(f)] The area under the curve is defined in Witten IH, Frank E (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann, San Francisco.
[(g)] Matthews correlation coefficient (Matthews 1975)

2.4) Cross-validation and prevention of overfitting

Due to the relatively low amount of data (173 variants), particular precautions needed to be taken. This arose due to the fact that upon normal cross validation, probability would be higher to bin outliers together and therefore artificially increase bias. For bigger datasets, this concern becomes less relevant given that binning large datasets results in lower contribution of potential outliers to the binned averages while in smaller datasets, a few outliers could skew considerably the average. In order to counteract that, we utilized 10-fold stratified cross validation to ensure incorporation of all data, while also preventing disproportionate bias through random sampling which could happen in normal unstratified cross validation.

2.5) Attributes used to distinguish pathogenic and benign mutations

The 11 different parameters used to feed our algorithm were either derived through manual inspection of the structures or computed through external sources.

(1) Absolute solvent accessibility in $Å^2$ of the wild-type residue.

(2) Absolute solvent accessibility in $Å^2$ of the mutant residue.

(3) Relative solvent accessibility of the wild-type residue, discretized into three classes according to (WORTH; PREISSNER; BLUNDELL, 2011): accessible (>43% relative SASA), partially accessible (17-43% relative SASA), and inaccessible (<17% relative SASA).

(4) Relative solvent accessibility of the mutated residue, discretized into three classes according to (WORTH; PREISSNER; BLUNDELL, 2011): accessible (>43% relative SASA), partially accessible (17-43% relative SASA), and inaccessible (<17% relative SASA).

(5) Change in numbers of non-hydrogen atoms upon mutation.

(6) Change in hydrophobicity upon mutation according to the scale of (KYTE; DOOLITTLE, 1982).

(7) Difference between attributes 4 and 3, which are discrete variables: as a consequence this attribute is discretized too and it is classified at “no change” if the attributes 6 and 7 are equal, “better” if the solvent accessibility increases upon mutation, and “worse” in the opposite case.

(8) A composite discrete variable, taking into account known binding partners, introduced side-chain clashes upon mutation as well as potential disruption of a phosphosite. Mutations were deemed to clash if there was no side-chain rotamer in the Pymol library (The PyMOL Molecular Graphics System,

version 2.3, Schrödinger, LLC) that could have been fitted without altering the structure. Mutations were considered to be related to protein phosphorylation, and as a consequence prone to engender pathological consequences, if closer than 8 Å from a phosphorylation site. Cα - Cα - distances were used as a cut off distance here and phosphorylation sites were taken from (REIMANN; SCHWÄBLE; FRICKE; MÜHLHÄUSER et al., 2020). If either a sidechain clashes or the specified amino acid is within 8 Å vicinity of a phosphorylation site this parameter takes the value of "yes" else as "no".

(9) Side-chain orientation of the affected amino acid. In the case of Pro/Gly this is labeled with a "?". This attribute can either assume 2 states, "towards solvent" or "towards protein core".
The decision if an amino acid side chain is directed towards solvent or towards the core was based on the cut off of 20 $Å^2$ solvent exposure. If the side chain of the mutated residue had below 20 $Å^2$ solvent exposure it was deemed to be buried and therefore directed toward the protein core. For glycine and proline residues this parameter was set to "?" by default.

(10) Type of secondary structure of the mutated residues - a special tag was used for mutations located in the intrinsically disordered part of Ig20.

(11) A composite variable, combining solvent accessibility and change in hydrophobicity named SAP-score can be calculated according to the following formula (VOYNOV; CHENNAMSETTY; KAYSER; HELK *et al.*, 2009).

$$SAPatom(i,R) = \sum_{1}^{n} \left(\frac{SAA\ of\ atom\ within\ radius\ R}{SAA\ of\ fully\ exposed\ residual\ atom} * Residue\ Hydrophobicity \right) \quad \text{(eq. 1)}$$

The SAP score of a particular atom (i) is the summation over all atoms within radius R (10 Å) and takes into account the normalized solvent accessibility (with respect to the fully exposed potential atom, ranging from 0 to 1 for fully buried and fully exposed atoms respectively) and its associated residue hydrophobicity (normalized according to VOYNOV et al. with 0 as the assigned valued for glycine). The modified hydrophobicity scale assigns positive values for hydrophobic and negative values for polar residues. Therefore it follows that a large positive value corresponds to a solvent exposed hydrophobic residue atom which would be aggregation prone, while negative values would correspond to solvent exposed polar residue atoms and therefore be acceptable.
For our predictions, we summed up all atoms of the residue of interest and assigned the mean atomic SAP score value to the whole residue.
Due to the summation over all residues of radius 10 Å, the SAP score gives information about the local neighbourhood. In cases where hydrophobic residue clusters are solvent exposed, the SAP scores for the individual residues within the cluster will sum up to larger positive values, while lone hydrophobic solvent exposed residues will still be tolerated at low positive values if they are surrounded by polar exposed residues.

2.6) Plasmids and DNA constructs

Ig14 (residues 1534-1634), Ig15 (residues 1633-1736), Ig14-15 (residues 1534-1736), Ig14-15^{S1624L}(residues 1534-1736), Ig14-15^{G1676R} (residues 1534-1736), Ig19 (residues 2033-2130), Ig21 (residues 2313-2406), Ig2021 (residues 2132-2406), Ig1921 (residues 2036-2406), Ig1921^{I2160F} (residues 2036-2406), Ig1921^{W2164C} (residues 2036-2406), Ig1921$^{P\ 2298S}$ (residues 2036-2406), Ig1921^{Q2058R} (residues 2036-2406) and Ig1921 ^{K2260R} (residues 2036-2406) of the human FLNc gene (UniProt code Q14315) were cloned into p3NH-vector (Mlynek, Lehner et al. 2014) which confers resistance to chloramphenicol and kanamycin, and attaches an N-terminal His_6 tag followed by a human rhinovirus 3C protease cleavage site.

2.7) Protein expression

For protein expression, 6x 500 mL Autoinduction media (no trace metal mix added) supplemented with 100 µg/mL chloramphenicol and 50 µg/mL kanamycin was inoculated with 20 mL overnight culture and grown at 37°C until OD_{600} ~0.8 shaking at 150 rpm. Then the temperature was reduced to 20°C and after 12 hours the culture was centrifuged (4°C, 5000xg) and the resulting cell pellet was either processed immediately or frozen in liquid nitrogen and stored at -80°C.

2.8) Protein purification

Ig21 and Ig19 cell pellets were resuspended in 100 mL lysis buffer (1x PBS, 20 mM imidazole, pH 7.4) while Ig20-21, Ig19-21, Ig19-21^{Q2058R}, Ig19-21^{I2160F}, Ig19-21^{K2260R}, Ig19-21$^{P\ 2298S}$, Ig19-21^{W2164C} cell pellets were resuspended in 100 mL lysis buffer (1 X PBS, 20 mM imidazole, 2 M urea, pH 7.4). Ig14, Ig15, Ig14-15, Ig14-15^{S1624L}, Ig14-15^{G1676R} cell pellets were resuspended in 100 mL lysis buffer (50 mM MES, 150 mM NaCl, 2 M urea, pH 6.5). All lysis buffers were supplemented with 50 uL DNaseI (10 mg/mL) and sonicated twice for 3 min (50% Amplitude; 1s Pulse On; 1s Pulse Off). The lysate was clarified by centrifugation and subsequently, the supernatant was loaded onto a 5 mL HisTrap FF crude (GE Healthcare) column pre-equilibrated with lysis buffer. Constructs eluted with a step gradient using an elution buffer (1 X PBS, 500 mM imidazole, pH 7.4) for Ig19 and Ig21, while Ig20-21, Ig19-21, Ig19-21^{Q2058R}, Ig19-21^{I2160F}, Ig19-21^{K2260R}, Ig19-21^{P2298S} were eluted with 1 X PBS, 500 mM imidazole, 2 M urea, pH 7.4. Ig14, Ig15, Ig14-15, Ig14-15^{S1624L}, Ig14-15^{G1676R} were eluted with 50 mM MES, 500 mM imidazole, 2 M urea, pH 6.5. The N-terminal His-tag of Ig14, Ig15, Ig14-15, Ig14-15^{S1624L}, Ig14-15^{G1676R} Ig19 and Ig21 was cleaved by overnight incubating with human rhinovirus 3C protease with human rhinovirus 3C protease protease and afterwards loaded again onto a 5 mL HisTrap FF crude column pre-equilibrated with corresponding lysis buffer. This constructs were concentrated after pooling the fractions containing pure protein. Individual fractions of purified constructs were frozen in liquid nitrogen without further concentration and stored at -80°C until further use. Prior to the DSC experiments, the samples were overnight inoculated with human rhinovirus 3C protease and afterwards loaded again onto a 5 mL HisTrap FF crude column pre-equilibrated with corresponding lysis buffer. The successfully cleaved samples were subsequently buffer exchanged using a superdex S200 5/150 GL or 10/300 increase column (GE Healthcare) connected to an Agilent HPLC 1260 Affinity equipped with a fraction collector. The column was either equilibrated with 50 mM MES 150 mM NaCl pH 6.5 for Ig14, Ig15, Ig14-15, Ig14-15^{G1676R} , Ig14-15^{S1624L} or 1x PBS 150 mM NaCl pH 7.4 for Ig19, Ig20-21 Ig21, Ig19-21, Ig19-21^{Q2058R}, Ig19-21^{I2160F}, Ig19-21^{K2260R}, Ig19-21$^{P\ 2298S}$ and Ig19-21^{W2164C} constructs.

2.9) Crystallization of Ig14-15, Ig14-15^{S1624L} and Ig14-15^{G1676R}

Crystallization and Data collection including structural refinement was performed by Georg Mlynek, using SWISSCI MRC three-well crystallization plates (Molecular Dimensions) applying the sitting drop vapour diffusion method. Crystallization plates with commercially available screens were set up using the Mosquito crystallization robot (TTP LabTech). The reservoir was filled with 35 µL of precipitant solution and different ratios of protein to precipitant (150:200 nL, 200:200 nL and 250:200 nL) were applied. Crystallization plates were stored in a Formulatrix RI-1000 imaging device at 22°C. Crystals of Ig14-15 were obtained in SG1™ (ShotGun) screen (Molecular dimensions) in well G1 (0.1 M HEPES pH 7.0, 30 % v/v Jeffamine® ED-2003). Crystals of Ig14-15S^{1624L} were obtained in PACT premier™ screens (Molecular dimensions) in well A2 (0.1 M SPG buffer: succinic acid, sodium phosphate monobasic monohydrate, glycine; pH 5.0, 25 % w/v PEG 1500) and crystals of Ig14-15^{G1676R} in 0.1 M Bis-TRIS pH 6, 26.4 % w/v PEG 3350, 5% glycerol (initial conditions SG1 condition F11).

Crystals of Ig14-15, Ig14-15^{S1624L} were soaked with mother liquor supplemented with 20% 2-methyl-2,4-pentanediol (MPD), collected with cryo-loops and flash-vitrified with liquid nitrogen. For Ig14-15^{G1676R} mother liquor supplemented with 20% of glycerol was used as cryoprotectant.

2.10) Data collection and structure refinement

Datasets of Ig14-15 were collected at beamline i24 of the Diamond Light Source (UK) at 100 K using a Pilatus3 6M. Crystals of Ig14-15^{S1624L} on i04 of the Diamond Light Source at 100 K using a Pilatus 6M-F and crystals of Ig14-15^{G1676R} at beamline MASSIF-1 at ESRF using a Pilatus3 2M (http://dx.doi.org/10.1107/S1399004715011918). The dataset was processed with XDS and symmetry equivalent reflections merged with XDSCONV (Kabsch 2010). Intensities were not converted to amplitudes. Initially, we used a conservative high-resolution cutoff 1.7 Å ($CC_{1/2}$ = 67.5; I/σ(I)=1.57) for Ig14-15, 1.92 Å ($CC_{1/2}$ = 86; I/σ(I)=2.1) for Ig14-15^{S1624L} and 1.83 Å ($CC_{1/2}$ = 93.9; I/σ(I)=3.81) for Ig14-15^{G1676R} (Karplus and Diederichs 2012). The phase problem of Ig14-15 was solved by molecular replacement using MORDA from ccp4 online (Winn, Ballard et al. 2011, Vagin and Lebedev 2015). For Ig14-15^{S1624L} and Ig14-15^{G1676R} we used phenix.phaser (McCoy, Grosse-Kunstleve et al. 2007) taking Ig14-15 as the search model. The models were further improved by iterative cycles of a manual model building using COOT (Emsley, Lohkamp et al. 2010) and maximum likelihood refinement using phenix.refine (Adams, Afonine et al. 2010). Phenix.refine converted intensities into amplitudes using the French and Wilson algorithm (French and Wilson 1978). The final high-resolution cutoff was based on performing paired refinement using PAIREF (Malý, Diederichs et al. 2020) and PDB_REDO webserver (Joosten, Salzemann et al. 2009).

Final stages of refinement included translation-libration-screw (TLS) parameters, isotropic B-factor model, automated addition of hydrogens and water molecules, optimization of X-ray/ADP weight, and optimization of X-ray/stereochemistry weight for Ig14-15^{S1624L} and Ig14-15^{G1676R}. For Ig14-15 an anisotropic B-factor model. The models were validated with MolProbity (Adams, Afonine et al. 2010) and the PDB_REDO webserver. The statistics on data-collection and refinement are reported in (**Supplementary Table 4. Crystallographic Table**). Figures were prepared with PyMOL Molecular Graphics System (Version 2.4.0, Schrödinger, LLC). Atomic coordinates have been deposited in the Protein Data Bank under the accession code Ig14-15 (7OUU), Ig14-15^{S1624L} (7OUV) and Ig14-15^{G1676R} (7P0E).

2.11) DSC analysis

Based on experimental buffer screening (not shown here), MES buffer (50 mM MES, 150 mM NaCl pH 6.5) was selected for Ig14-15 constructs and PBS (1x PBS, 150 mM NaCl pH 7.4) buffer for Ig19-21 constructs. The constructs were investigated in the temperature range of 20 to 110°C. Protein concentration was determined by 280 nm absorbance with extinction coefficients being calculated from the primary amino acid sequence using Protparam (https://web.expasy.org/protparam).

Both buffers are DSC compatible and did not show any extensive contribution to heat capacity which could distort results. We kept concentrations of wild-types and mutants around 1 mg/ml if possible. Due to ongoing aggregation during concentration, Ig19-21^{I2160F} and Ig19-21^{P2298S} were measured at 0.82 mg/ml and 0.67 mg/ml, respectively, which still yielded interpretable thermograms. Normalized values were used for comparison of different mutations in terms of changes in enthalpy upon unfolding. For data analysis, a “non two state” model was employed to fit all thermograms, besides Ig19, Ig14, Ig15 and Ig21, where a “two state” model was used to fit the experimental data based on agreement between calculated Van't Hof and observed calorimetric enthalpy. For each experimental run, the baseline approximation under the transition curve was calculated with a spline approximation. Prior to each experimental construct run, we conducted at least three buffer runs to counteract thermal hysteresis of the device and chose the most stable buffer run for later buffer subtraction. All experiments from this study were conducted on a MicroCal PEAQ-DSC from Malvern Panalytical.

2.12) Data and Code availability

The atomic coordinates of all homology models are available upon request or downloadable from PyPi under the project folder of AMIVA-F (https://pypi.org/project/AMIVA-F/). The published package contains a training set, homology and experimentally derived structures and tutorials. An in depth tutorial guiding the installation process and the usage of AMIVA-F is included. In order to utilize AMIVA-F, a Java virtual machine (JVM) and PyMol are required (instructions on installation for a step by step guidance is specified at https://pypi.org/project/AMIVA-F/).
Java is required in order to access WEKA, the machine learning platform, while Pymol is required to compute some of the input parameters required for WEKA.
Inside the AMIVA-F GUI, an extensive tutorial and additional information for advanced users is available. AMIVA-F is tested on Windows10, as well as in a virtual environment (Anaconda 3, python=3.8), MacOSF and Linux (Ubuntu 20.04.2 LTS) and is designed to be operating system independent.

The crystal structures have been deposited in the Protein Data Bank and are available with these links and will be released upon publication:
https://www.rcsb.org/structure/unreleased/7OUU
https://www.rcsb.org/structure/unreleased/7OUV
https://www.rcsb.org/structure/unreleased/7P0E

3) Results

3.1) Impact of pathogenic variants on structural integrity - thermodynamic and structural analysis

In order to assess whether the pathogenicity is rooted in the altered stability of the domains, we performed a series of thermal stability and structural analyses. We selected eight variants, four of which are reported in the literature (p.Ile2160Phe (Ig20), p.Trp2164Cys(Ig20), p.Pro2298Ser (Ig20), p.Ser1624Leu (Ig14)), and four of which have been deemed interesting and therefore investigated by other laboratories with which we collaborate (p.Met82Lys (ABD), p.Gly1676Arg, (Ig15), p.Gln2058Arg (Ig19) and p.Lys2260Arg (Ig20)).

We failed to recombinantly express ABD p.Met82Lys mutant in *E. coli*, in line with the observation that variants that destabilise the protein fold often display notably reduced expression levels (Chowdhury, Vasmatzis et al. 1998). The Met82 residue maps to the first CH domain of ABD, to a loop connecting α-helices C and E, with its side-chain embedded in a hydrophobic environment provided by α-helices C and E of CH1.

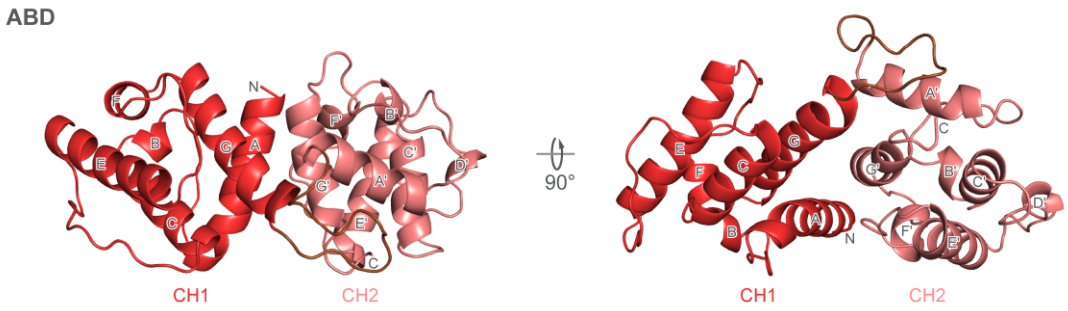


Figure 3. Topology of the actin-binding domain (ABD)
The actin-binding domain and its 2 calponin homology (CH) domains, CH1 and CH2.
Both CH-domains consist of 7 alpha-helices (A-G) and are connected through a flexible linker between CH1 and CH2. In contrary to all other Ig domains in FLNc, there are no beta sheets present.

A substitution of Met by Lys, with a charged and longer side-chain, is expected to cause severe steric clashes and repulsion, leading to destabilisation and potential unfolding of the domain (**Figure 4A ABD_M82K +DSC+ FLNc_Ig14-15_structure**). Impaired proper folding of the ABD is expected to negatively impact the solubility and the interaction with one of the main FlnC binding partners - actin filaments - and thus lead to miss-localisation of the protein. Indeed, this is corroborated by in-cell results from other laboratories, where FlnC bearing p.Met82Lys mutation was observed to form aggregates and cause disruption of the Z-discs (data not published).

We performed differential scanning calorimetry (DSC) on seven Ig14-15 and Ig19-21 constructs bearing single point mutations. A summary of the DSC results and the pathogenicity prediction of AMIVA-F can be found in (**Table 3. DSC results of known mutations together with AMIVA-F predictions (P7)**).

Table 3. DSC results of known mutations together with AMIVA-F predictions (P7)
Impact of selected variants on thermal stability analyzed by DSC, AMIVA-F prediction and associated phenotype.

Domain	c-Notation	p-Notation	Phenotype	ΔT_m (°C)	Pathogenicity prediction (AMIVA-F)	Reference
Ig14	c.4871C>T	p.Ser1624Leu	**RCM**	Ig14 and Ig 15: -12.8	**Yes**	(Brodehl, Ferrier et al. 2016)
Ig15	c.5026G>A	p.Gly1676Arg	**RCM**	Ig14: -6.5 Ig15: -23.9 [(a)]	**Yes**	Collaborators (unpublished)
Ig19	c.6173A>G	p.Gln2058Arg	*No*	Ig19: -2.8 Ig20: +0.3	*No*	Collaborators (unpublished)

				Ig21: +0.4		
Ig20	c.6478A>T	p.Ile2160Phe	**RCM**	Ig19: -0.2 Ig20: +7.1 Ig21: -0.9	**Yes**	(Brodehl, Ferrier et al. 2016)
Ig20	c.6490T>C	p.Trp2164Cys	**HCM**	Ig19: -0.8 Ig20: +7.2 Ig21: +0.9	**Yes**	H. Watkins, personal communicati-on
Ig20	c.6779A>G	p.Lys2260Arg	*No*	Ig19: -1.9 Ig20: +1.5 Ig21: +0.1	*No*	Collaborators (unpublished)
Ig20	c.6892C>T	p.Pro2298Ser	**RCM**	Ig19: -3.6 Ig20: +6.9 Ig21: -0.7	**Yes**	(Gómez, Lorca et al. 2017)

[(a)] Unfolding for Ig14-15 wildtype shows a single transition, this behaviour is lost upon introduction of p.Gly1676Arg.

The thermal unfolding of Ig14-15 in the range 20 - 110°C showed a single transition with melting temperature (T_m) of 77°C (**Figure 4B ABD_M82K +DSC+ FLNc_Ig14-15_structure**). The single unfolding transition showed a ratio between Van't Hoff and calorimetric enthalpy of 1.9, with 2.0 being expected for two equivalent domains unfolding in a single transition, thereby supporting the single entity unfolding behaviour. The crystal structure of Ig14-15 corroborates the single unfolding transition due to extensive interaction between the domains and formation of a hydrogen-bond stabilized interface, resulting in coupled unfolding (**Figure 4B and 3C ABD_M82K +DSC+ FLNc_Ig14-15_structure**). Notably, an interdomain interaction between Ig14 and Ig15 was suggested by Sethi et al (2014) based on a small-angle X-ray scattering based study (Sethi and Ylänne 2014), but the fine details of the interaction could not be resolved.

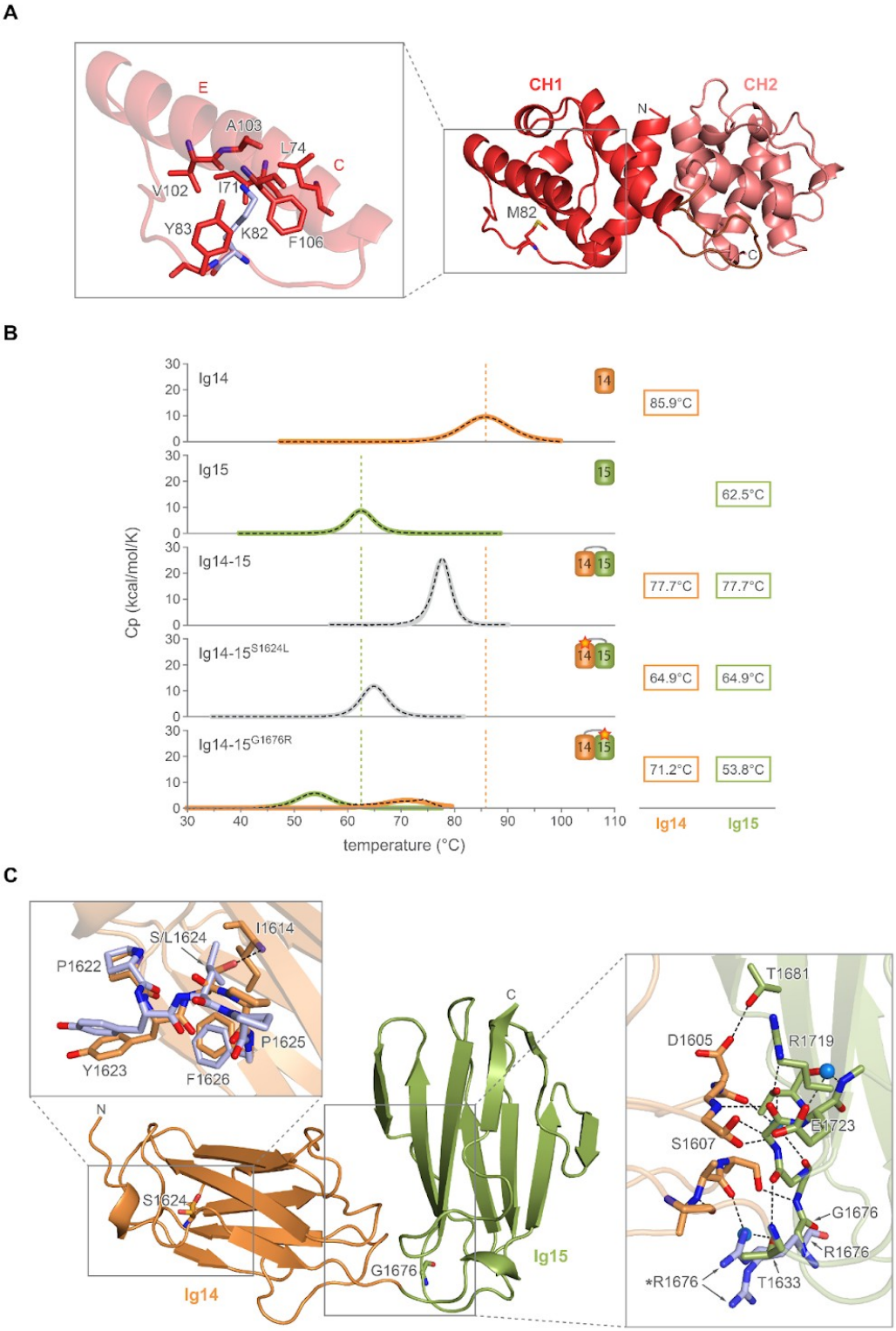


Figure 4. ABD_M82K +DSC+ FLNc_Ig14-15_structure

Structural and thermal stability analysis of selected mutations in ABD, and Ig14-15 domain. (A) ABD of human FlnC is shown in ribbon diagram, with CH1 and CH2 colored in dark and light red. Inlet shows the model of p.Met82Lys mutant generated by Duet (Pires, Ascher et al. 2014), with positively charged Lys side-chain clashing with the hydrophobic cluster formed by residues of h3 and h4. (B) Thermograms of Ig14 wild-type, Ig15wt, Ig14-15 wt, Ig14-15^{G1676R}, Ig14-15^{S1624L}. (C) Crystal structure of Ig14-15. Left inlet: superposition of PXSP loops of Ig14-15^{S1624L} (blue) and Ig14-15 wild-type, showing the crucial hydrogen bond between Ser1624 to Ile1614, which is lost upon mutation to Ile, with concomitant distortion of the loop to accommodate for a bigger, hydrophobic residue at this position. Right inlet: the details of hydrogen-bonding network between Ig14 and Ig15 do-

mains, with hydrogen bonds highlighted with dashed lines, solvent molecules are shown as light blue spheres, the position of the Gly1676, mutated into Arg is labelled.

The thermal stability of constructs Ig14-15^{S1624L} and Ig14-15^{G1676R} bearing mutations p.Ser1624Leu and p.Gly1676Arg found in patients with RCM (Brodehl, Ferrier et al. 2016), were examined under the same experimental conditions. The p.Ser1624Leu mutation is located in Ig14 (**Figure 4C, left inset ABD_M82K +DSC+ FLNc_Ig14-15_structure**), and maps to a highly conserved and functionally important PXSP motif, which is present in nearly all Ig domains of human FlnC and was shown to be highly phosphorylated *in vivo* in several FlnC Ig domains (Reimann, Wiese et al. 2017, Reimann, Schwäble et al. 2020). Ig14-15^{S1624L} displayed similarly to the wild-type a single transition upon unfolding (**Figure 4B ABD_M82K +DSC+ FLNc_Ig14-15_structure**), and a ratio between Van't Hoff and calorimetric enthalpy of 1.85, in agreement with both domains unfolding in a coupled manner. In terms of fold stability, Ig14-15^{S1624L} displayed T_m of 64.9°C, suggesting a destabilising effect of p.Ser1624Leu mutation on the overall domain pair stability. While the overall structure and interaction interface between Ig14 and Ig15 are essentially unaffected, the hydrogen bond between Ser1624 OG and Ile1614 backbone N residing in a ß-strand of Ig14 is lost (**Figure 4C, left inset ABD_M82K +DSC+ FLNc_Ig14-15_structure**). Replacement of Ser to Leu prevents the formation of this stabilising interaction with destabilisation reflected in reduced T_m of Ig14-15 domain couple (**Figure 4C, left inlet ABD_M82K +DSC+ FLNc_Ig14-15_structure**).

On the contrary, the construct Ig14-15^{G1676R} hosting the p.Gly1676Arg displays a different unfolding trajectory compared to the wild-type. Instead of the single unfolding transition, we observed two, with T_ms for Ig14 and Ig15 of 71.2°C and 53.8°C, respectively. The mutation is located in Ig15 (**Figure 4C, right insert ABD_M82K +DSC+ FLNc_Ig14-15_structure**), close to the hydrogen bonded interface between the two domains and consequently, affects inter-domain coupling. This leads to decreased folding stability with Ig15 more affected than Ig14 and two independent unfolding events (**Figure 4B, ABD_M82K +DSC+ FLNc_Ig14-15_structure**). Some clues about the unfolding decoupling are provided by the X-ray crystal structure of the Ig14-15^{G1676R} construct: the introduction of a larger and positively charged residue perturbs the interface structure between Ig14 and Ig15 considerably. The number of interdomain interactions mediated by water molecules increases, and this might limber the interface between the two domains and favour their independent unfolding (**Figure 5. G1676R and WT analysis, Ligplot).**

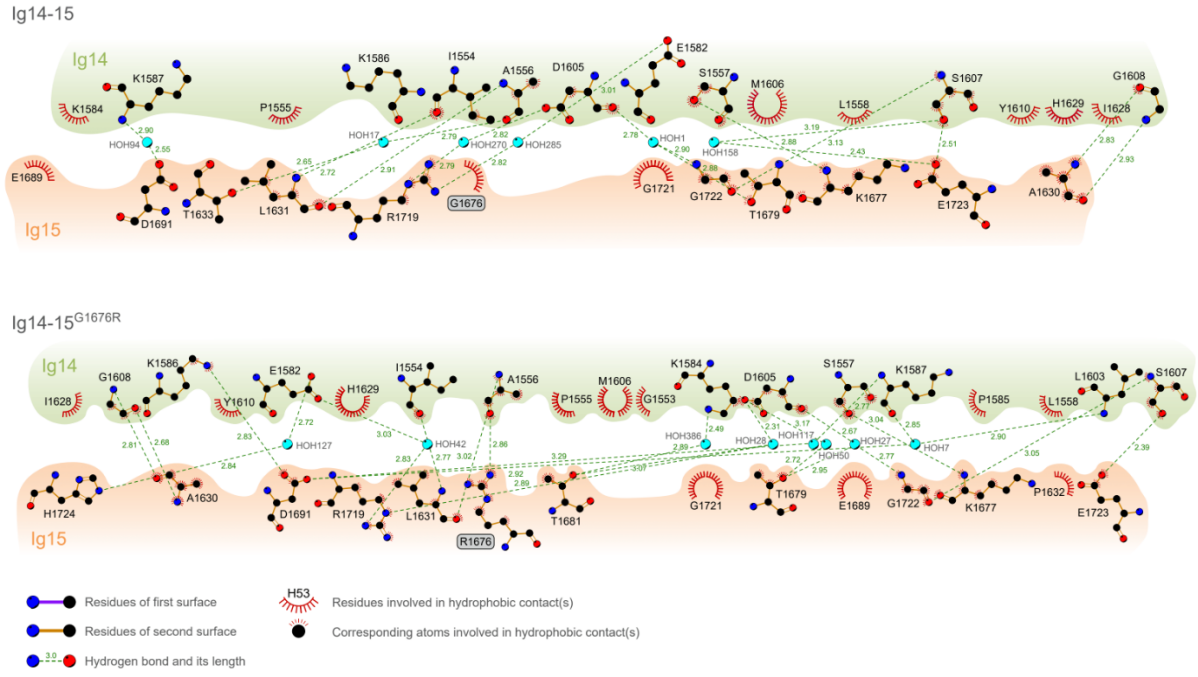


Figure 5. G1676R and WT analysis, Ligplot
Top: Hydrogen bond network of the wildtype construct and its associated residues depicted with dotted lines and dots respectively. Blue dots resemble conserved water molecules, found in the experimental crystal structure.
Bottom: Hydrogen bond network of the mutant variant.

Based on these results, the molecular basis of the pathogenicity of both mutations p.Ser1624Leu and p.Gly1676Arg resides in decreasing the fold stability and/or association strength and communication between the two domains.

Analysis of thermal stability on a second set of mutations (p.Ile2160Phe (Ig20, RCM), p.Trp2164Cys (Ig20, HCM), p.Lys2260Arg (Ig20, non-pathogenic) p.Gln2058Arg (Ig19, non-pathogenic) and p.Pro2298Ser (Ig20, RCM)) was performed on the Ig19-21 construct. The thermogram of the wild type displayed three overlapping transitions, at 60.9°C, 75.2°C and 84.8°C (**Figure 6A DSC + 3D Model Ig19-21**). These were assigned to Ig21, Ig20, and Ig19, respectively, by comparison with the thermograms of individual domains (**Supplementary figure 2. DSC 19-21 assignment**).

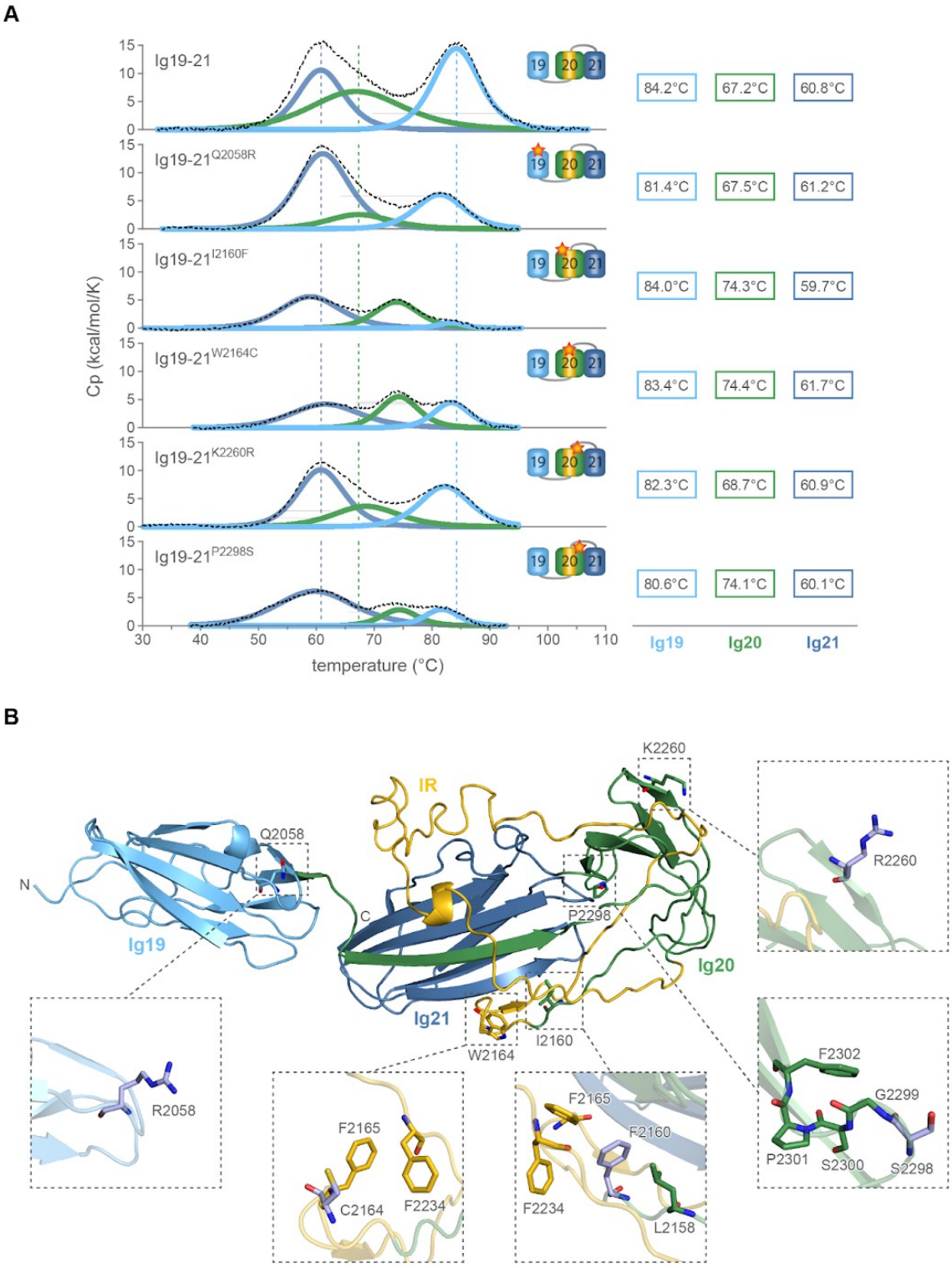


Figure 6. DSC Ig19-21 + 3D Model Ig19-21.
Structural and thermal stability analysis of selected mutations in Ig19-21 domains. (A) Thermograms of wild-type Ig19-21 and its mutants; (B) Cartoon diagram of Ig19-21 structural model generated by threading approach as encoded in I-tasser {Roy, 2010; Yang, 2015; (10.1093/nar/gkv342), with domain color code as in A. Positions of mutated residues p.Gln2058Arg, p.Ile2160Phe, p.Pro2298Ser, p.Lys2260Arg, p.Trp2164Cys are depicted as stick models. A detailed view of the mutations generated with Duet {Pires, 2014} are shown in the inlets.

In summary, two mutations (p.Gln2058Arg and p.Lys2260Arg) displayed no impact on fold integrity ofIg19-21 constructs. On the other hand, three pathogenic mutations (p.Ile2160Phe, p.Trp2164Cys, and p.Pro2298Ser) resulted in increased stability of Ig20 (**Table 3. DSC results of known mutations together with AMIVA-F predictions (P7)**).

Mutation p.Lys2260Arg, located in Ig20, does not show any significant change of T_m. Overall, mutation from Lys to Arg brings only a small physicochemical change, resulting in a similar unfolding trajectory as the wild type. Similarly, a modest decrease (from 84.2°C to 81.4°C) of Ig19 T_m is observed in p.Gln2058Arg, which maps to a solvent exposed loop in Ig19 (**Figure 6B (DSC + 3D Model Ig19-21)**)

Mutation p.Ile2160Phe, associated with RCM (Brodehl, Ferrier et al. 2016), closely precedes the IR in Ig20 (aa 2162-2243) (**Figure 6B. DSC + 3D Model Ig19-21, Figure 7. (IDP+HCA) Prediction of intrinsic disorder and hydrophobicity cluster analysis for Ig20 , Supplementary Figure 3. Topologies of human FLNc Ig-domains Part I-III**), and leads to a stabilization of Ig20 compared to the wild-type by 7°C, while the other domains are not affected. The p.Ile2160Phe side-chain is buried and involved in a structure-stabilizing hydrophobic cluster enriched in aromatic residues (**Supplementary Figure 3. Topologies of human FLNc Ig-domains Part III), Figure 6B. DSC + 3D Model Ig19-21**), where replacement of Ile with an aromatic Phe can increase the stability of the domain by aromatic stacking interactions. A nearby mutation p.Trp2164Cys, associated with HCM (H. Watkins, personal communication), is particularly interesting since it maps to the beginning of the IR that is considered to lack a unique and well-defined 3D structure. The mutation leads to an increase in the melting temperature of Ig20 by 7°C, suggesting a local structuring rearrangement, as inferred from the relative increase in enthalpy change upon unfolding of Ig20 compared to Ig19 and Ig21 (**Figure 6A. DSC + 3D Model Ig19-21**). In the case of p.Ile2160Phe and p.Trp2164Cys, we hypothesise that increased relative enthalpy of unfolding, suggestive of an increased structuring of the surrounding of the mutation, can negatively impact interactions with binding partners, where structural plasticity of a disordered region is required, leading to RCM (Brodehl, Ferrier et al. 2016) or HCM (H. Watkins, personal communication), respectively.
p.Pro2298Ser mutation, associated with RCM ((Gómez, Lorca et al. 2017)) maps to Ig21 (**Figure 6B. DSC + 3D Model Ig19-21**), and shares an important similarity with p.Ser1624Leu in Ig14: both mutations are part of the highly conserved PXSP motif (Rognoni, Möst et al. 2014). Similar to the p.Ser1624Leu, the p.Pro 2298Ser mutation also influences the stability of neighbouring domains (**Figure 6B. DSC + 3D Model Ig19-21**), possibly through inter-domain contacts observed by NMR studies and small-angle X-ray scattering of Ig16-21 (Ruskamo, Gilbert et al. 2012, Tossavainen, Koskela et al. 2012, Seppälä, Bernardi et al. 2017) and cryo-EM study of Ig16-24 of FlnA (Ruskamo, Gilbert et al. 2012). The observed increased T_m (7°C) could be attributed to the replacement of Pro with Ser, which can form additional hydrogen bonds, both with the side-chain and the main-chain amide group. The mutation p.Pro2298Ser affects the first proline residue in the conserved PXSP motif, which was shown to be involved in the regulation of mechanosensing in FlnA Ig20 through *cis-trans* isomerisation of the second proline of the motif (Rognoni, Möst et al. 2014). Substitution of proline with a multivalent hydrogen-bond acceptor and donor residue, serine, could disturb the finely tuned *cis-trans* isomerisation balance involved in mechanosensing regulation.

3.2) Development of predictive bioinformatics tool AMIVA-F

Given that a growing number of *FLNC* variants are associated with (cardio)myopathies, we were motivated to develop a prediction tool that uses machine learning (Analysis of MIssense VAriants in human Filamin C, AMIVA-F) to predict variants pathogenicity.

3.3) Training sets

To develop a predictive bioinformatics tool, we first selected a training set of variants associated with HCM, DCM, RCM, myofibrillar and distal myopathy, and some others. A total of over 250 unique variants were retrieved from the international peer-reviewed literature. Among those, we omitted variants from the collection that either lacked a clear description of the phenotype associated with the variants or where no clinical information was available. Furthermore, we only collected single point missense variants, excluding frameshift or truncating variants, leading to a total of 108 disease related variants (**Figure 7. All variants overview)**.

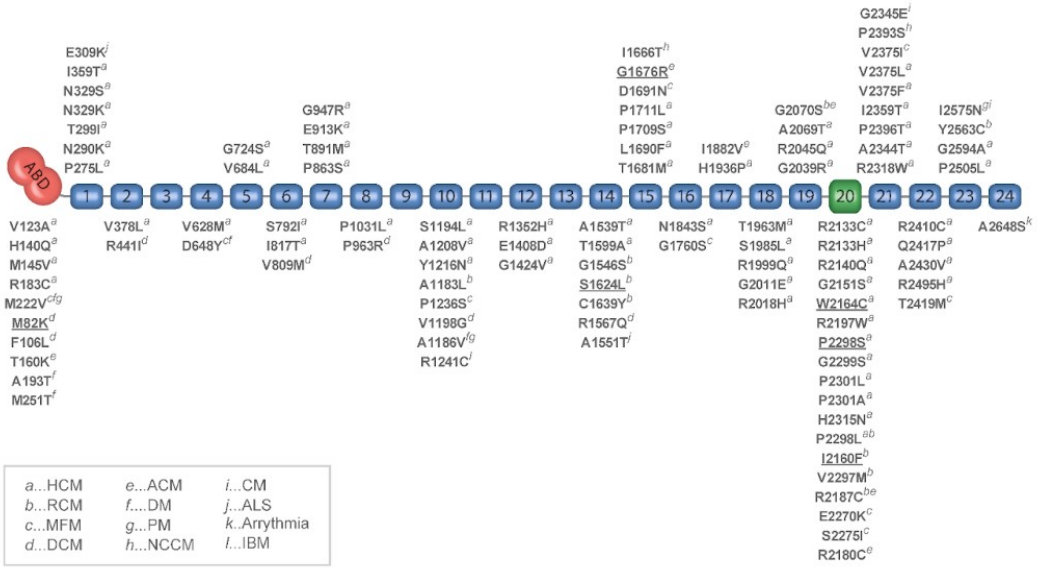


Figure 7. All variants overview. Variants in the *FLNc* gene that have been previously or in this study reported in individuals with myopathy or cardiomyopathy, mapped to corresponding FlnC domain. Variants are annotated at protein level. Abbreviations: HCM (hypertrophic cardiomyopathy), RCM (restrictive cardiomyopathy), MFM (myofibrillar myopathy), DCM (dilated cardiomyopathy), ACM (arrhythmogenic cardiomyopathy), DM (distal myopathy), PM (peripheral myopathy), NCCM (non compaction cardiomyopathy), CM (cardiomyopathy), ALS (amyotrophic lateral sclerosis), IBM (inclusion body myositis). Mutants experimentally characterized in this work are highlighted in bold.

There is at least one literature report for each variant found in the disease-related dataset, which documents its pathogenicity (**Supplementary Table 1. Pathogenic mutations dataset overview)).** 65 variants are associated with HCM, eight with RCM, five with myofibrillar myopathies, five with distal myopathies, four with ACM, three with DCM, two with IBM (inclusion body myositis), two with congenital heart disease, two with left ventricular non-compaction, and 12 with other diverse phenotypes.
For the selection of a negative or neutral mutation set, we gathered single point missense variants found in the Genome Aggregation Database (gnomAD; https://gnomad.broadinstitute.org). This yielded a collection of 1500 missense variants, which were analyzed and selected according to the following criteria: (i) the selection was based on the study by Karczewski et al. (2020) (Karczewski, Francioli et al. 2020), who from analysis of 141,456 exome variants derived that loss of function and disease related mutations tend to be found at extremely low frequencies compared to neutral ones; (ii) the presence of individuals having homozygous alleles for variants was taken as further evidence of a variant being neutral. We selected only variants with high allele frequency (> 10^{-3}), and high allele counts based on these criteria. By comparison, variants found in the gnomAD associated with pathogenicity were found at frequencies around 10^{-5}, a difference of up to three orders of magnitude.

We discarded variants that showed less than 10 allele counts as these variants might be rare but harmless. In total, these stringent criteria selections led to a reduction from 1500 to 65 variants and constituted the neutral dataset. The average allele count in the neutral variants dataset was 504 counts, and some selected variants also showed homozygous alleles.

In summary, our two datasets, consisting of 108 disease-related variants (**Figure 7. All variants overview, Supplementary Table 1. Pathogenic mutations dataset overview**) and 65 presumably neutral variants (**Supplementary Table 2. SGnaomAD-negative Dataset**), were merged to 173 instances and used as training sets for our algorithm. An overview of the distribution of pathogenic variants through FlnC domains is reported in (**Figure 7. All variants overview)**, showing that they tend to cluster in the rod-2 region, which is a mutational and protein-protein interaction hot-spot of FlnC (van der Ven, Ehler et al. 2006, Mao and Nakamura 2020, Verdonschot, Vanhoutte et al. 2020, Eden and Frey 2021).
We observed that mutations of arginines and methionines result in pathogenicity more frequently than expected based on aminoacid composition, in particular when arginine is mutated into cysteine. Pertinent details on the propensity of each amino acid type to result in pathogenic variants upon mutation are reported in the **Pathogenicity propensity upon mutation section**.

3.4) Description of the attributes

AMIVA-F is based on the hypothesis that pathogenic variants may modify protein function in a complex way by impacting the fold stability and integrity as well as by influencing interaction interfaces or phosphorylation sites.

In order to characterize the impact of variants at a biophysical and structural level, we first prepared a collection of 3D structures for all of wild-type FlnC domains (**Supplementary Figure 3. Topology and architecture of human FLNc Part I-IV, Table 1. Summary of 3D - structures used in algorithm training**), composed of 6 experimental structures covering 8 Ig domains, and 15 3D computational models covering the ABD and 16 Ig domains. The structural models of the mutants were prepared with the DUET server (Pires, Ascher et al. 2014). 13 computational models were generated by homology modeling with Modeller (Sali and Blundell 1993), while models of Ig16 and Ig19-21 were generated by threading with I-tasser (Roy, Kucukural et al. 2010, Yang and Zhang 2015). The structure of Ig16 was built with threading due to the lack of 25 amino acids N-terminally in the deposited NMR structure (2d7n). This lead to a missing beta strand and overall lacking structure. In the case of Ig19-21, threading was done due to the presence of the 82 amino acid IR, for which no 3D structure exists and thus cannot be handled by homology modeling. Conformational disorder predictions generated with GeneSilico (Kozlowski and Bujnicki 2012) suggest that this insertion is partially disordered (residues 2171-2210) and partially ordered (residues 2210-2243) (**Figure 8. (IDP+HCA). Prediction of intrinsic disorder and hydrophobicity cluster analysis for Ig20. A and B**). Similarly, hydrophobic cluster analysis (HCA) (Eudes, Le Tuan et al. 2007) predicts for this region islands of hydrophobic clusters after residue 2210 (**Figure 8A. (IDP+HCA). Prediction of intrinsic disorder and hydrophobicity cluster analysis for Ig20.).**

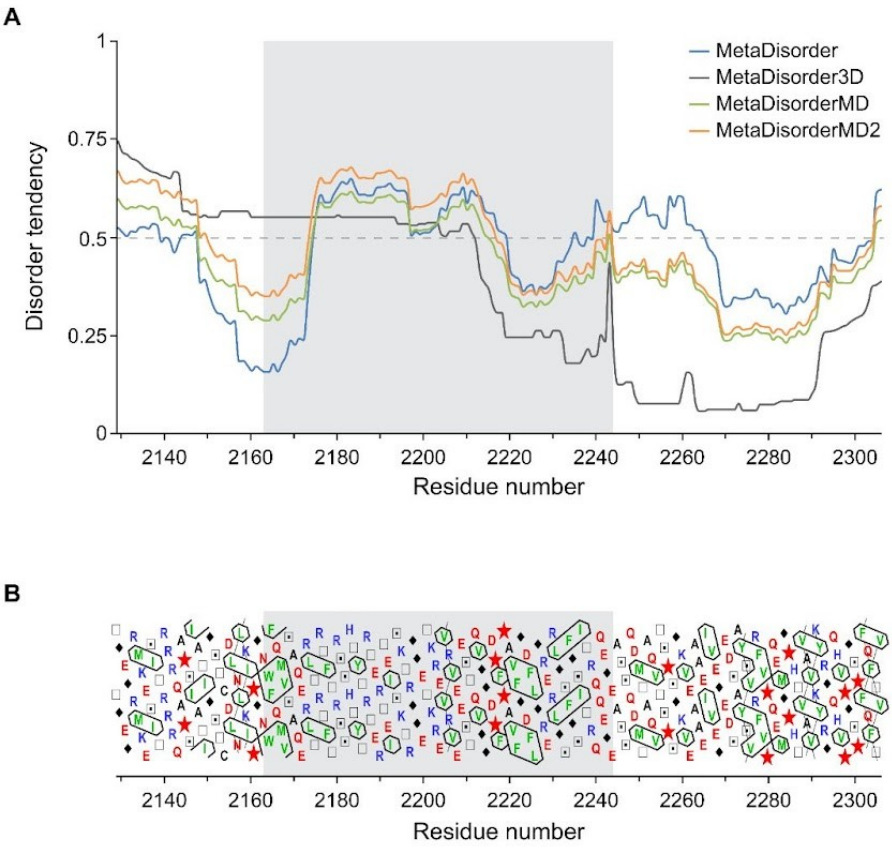


Figure 8. (IDP+HCA). Prediction of intrinsic disorder and hydrophobicity cluster analysis for Ig20.
(A) Four conformational disorder predictions generated with the GeneSilico metaserver (Kozlowski and Bujnicki 2012) showing partial disorder in the insertion region, highlighted in grey; (B) Hydrophobic cluster analysis indicates the presence of secondary structure in insertion region of Ig20, highlighted in grey. Pro residues are represented by red stars, Gly by black diamonds, square and dotted square indicate Thr and Ser, respectively. Vertical clusters are associated with ß-strands whereas horizontal ones correspond to α-helices. The 82 amino acid insertion region (IR) encompasses residues 2162-2243.

The N-terminal region is rich in Arg, Gly, and Glu residues, where high frequency and alternation of positive and negative charges is typically observed in intrinsically disordered regions prone to liquid-liquid phase separation and formation of macromolecular condensates (Banani, Lee et al. 2017).

The partial compact nature of the insertion was corroborated by differential scanning calorimetry (DSC) analysis (see above, section **Impact of pathogenic variants on structural integrity - thermodynamic and structural analysis**), where a change in heat capacity upon unfolding with respect to temperature is observed, suggesting that some secondary structure is present in this region. Otherwise, the DSC thermogram would not show any transition, but a flat line, due to the absence of the intermolecular hydrogen bonding that characterizes secondary structures.

These are the reasons why a threading model of the Ig19-21 construct was generated. However, since the structural details on the insertion region are uncertain, the secondary structure of the residues within the insertion region was not assigned, and it was labelled differently from all other residues.

Several features were selected to analyze the potential effects of variants, ranging from hydrophobicity changes caused by mutations to variables strictly related to 3D structures, like changes in solvent acces-

sible surface areas caused by mutations. In order to fully characterize a variant, we defined 11 different attributes (see **Materials and Methods** for details) that were derived through analysis of 3D structures by using scripts included in the AMIVA-F package.

In the course of the development of the prediction tool we initially included an attribute related to the predicted change in folding free energy upon mutation: ΔΔG ($\Delta G(mutant)^{folding}$ - $\Delta G(wild\text{-}type)^{folding}$) as predicted with the DUET server (Pires, Ascher et al. 2014), where the negative ΔΔG value corresponds to a destabilizing mutation. This attribute was subsequently removed because we managed to reach the same accuracy without it, while its exclusion substantially simplified the development of a user-friendly automated prediction tool.

Several attributes are related to the solvent accessibility of the residues: the absolute solvent accessible surface areas of both wild-type and mutant residues, and the discretized solvent accessibilities, grouped into the three categories (inaccessible, partially accessible, and accessible, as defined by Worth et al., (2011) (Worth, Preissner et al. 2011), and detailed in **Materials and Methods**).

Other attributes are related to residue properties, like the change in hydrophobicity upon mutation, according to the scale defined by Kyte et al. (1982) (Kyte and Doolittle 1982), the spatial aggregation propensity score (SAP-score) (Voynov, Chennamsetty et al. 2009), the change in the number of non-hydrogen atoms upon mutation, and the secondary structure of the residue that is mutated.

Furthermore, we included attributes that do not depend just on residue properties: the side-chain orientation of the mutated residue - this is ignored for Gly and Pro residues - and a composite variable, taking into account known binding partners, side-chain clashes introduced upon mutation as well as potential disruption of phosphorylation sites.

3.5) Prediction algorithm AMIVA-F

Predictions were made with a multilayer perceptron neural network, by using WEKA (Witten, Frank et al. 2017), an open-source workbench that allows the use of several machine learning techniques.

In order to benchmark the performance of AMIVA-F, we used ZeroR algorithm (Witten, Frank et al. 2017) that predicts the most common instance in a dataset and therefore defines a baseline accuracy. In general, baseline accuracy is solely dependent on the underlying dataset and it is algorithm independent. For our dataset, ZeroR determined 62.4% as baseline accuracy.

Given the relatively low number of instances (173 variants), particular precautions had to be taken to train the network because, upon regular cross-validation, there is a probability to bin outliers together and artificially increase bias. For larger datasets, this concern becomes less relevant given that binning large datasets results in a lower contribution of potential outliers to the binned averages. In comparison, in smaller datasets, a few outliers could considerably skew the average. To counteract that, we sampled by utilizing 10-fold stratified cross-validation to ensure equal incorporation of all data while also preventing disproportionate bias through random sampling, which could happen in regular unstratified cross-validation.

A tutorial on using the algorithm AMIVA-F and preparing a test set containing variants of interest is provided at https://pypi.org/project/AMIVA-F/**.**

3.6) Prediction validation

Prediction performance, evaluated by means of a 10-fold cross validation is satisfactory: sensitivity and accuracy are averaged close to or slightly around 0.8 as well as the area under the ROC curve (AUC) and the F-statistics; specificity, though slightly lower, is still in a comfortable range close to 0.7 (see **Table 2. Summary of the quality of predictions** in **Materials and Methods** for details).

Besides the canonical cross validation, we used 8 variants, classified to be likely pathogenic (Fokkema, Taschner et al. 2011) as external test cases. Their exclusion from the training sets was justified by the absence of concrete information about the specific disease underlying those variants, even though the probability of the variant to be disease causing is sufficient to warrant clinical actions (Richards, Aziz et al. 2015). As shown in (**Table 4. Validation of AMIVA-F against clinically classified data**), six predictions were correct (75%), in agreement with the cross validation of AMIVA-F and this reinforces the cross validation estimations, despite the small dimension of the external test set (8 variants).

Table 4. Validation of AMIVA-F against clinically classified data.
Pathogenicity prediction of eight "likely pathogenic" classified variants which were absent from the training set.

Domain	c-Notation	p-Notation	Classification	AMIVA-F prediction
ABD	c.368T>C	Val123Met	Likely pathogenic	pathogenic
ABD	c.643G>A	Val215Met	Likely pathogenic	pathogenic
ABD	c.t54A>C	Thr252Pro	Likely pathogenic	pathogenic
Ig3	c.1546C>A	Pro516Thr	Likely pathogenic	non-pathogenic
Ig5	c.2084G>C	Arg695Pro	Likely pathogenic	pathogenic
Ig6	c.2311G>A	Gly771Ser	Likely pathogenic	pathogenic
Ig14	c.4706C>T	Ala1569Val	Likely pathogenic	non-pathogenic
Ig22	c.7409C>A	Pro2470His	Likely pathogenic	pathogenic

Additionally, we verified if predictions are biased by the frequency of occurrence of the variants in the GnomAD database. This is demanded by the fact that the non-pathogenic variant learning set was assembled based on occurrence frequency – variants observed frequently are likely to be harmless (Lek, Karczewski et al. 2016, Karczewski, Francioli et al. 2020) – while the pathogenic variant learning set is constituted of infrequent variants, even if it was assembled based on published experimental information on their pathogenicity and not on the basis of their frequency.

For this purpose, 16 rare variants (<10^{-6} allele frequency), with no reported pathogenicity, were selected and used as test cases. Nine of them were predicted to be pathogenic and seven to be non-pathogenic (**Supplementary Table 5. (P16)**).
While there is no experimental proof that these predictions are correct, they indicate that the algorithm is not simply classifying based on rarity. In other words, this indicates that predictions made with AMIVA-F are independent of the variant frequency within the population.

3.7) AMIVA-F prediction capacity at a glance

Six pathogenic variants, with clinical data reported in this work, and six non-pathogenic variants, were selected for testing the prediction capacity of AMIVA-F (**Table 5 (P12)**).

The AMIVA-F algorithm correctly predicted 10 out of 12 variants, one pathogenic variant was predicted to be non-pathogenic and one non-pathogenic variant was predicted to be pathogenic (**Table 5. (P12)**), highlighting a high quality predictive power of AMIVA-F for pathogenic and non-pathogenic variants.

Table 5. (P12).
Pathogenicity prediction of 12 mutations, six of which are pathogenic, according to the clinical data reported in this work (bold) and six of which are deemed to be non-pathogenic according to our selection criteria or to clinical data reported in this publication (*italic*).

Domain	c-Notation	p-Notation	Phenotype	Pathogenicity prediction (AMIVA-F)	Reference
ABD	c.245T>A	p.Met82Lys	**DCM**	**Yes**	This study
Ig2	c.1142G>A	p.Arg381His	*No known disease association*	*No*	
Ig6	c.2392G>A	p.Asp798Asn	*No known disease association*	*No*	
Ig11	c.3812C>G	p.Thr1271Ser	*No known disease association*	**Yes**	
Ig13	c.4459G>A	p.Val1487Met	*No known disease association*	*No*	
Ig14	c.4871C>T	p.Ser1624Leu	**HCM/RCM**	**Yes**	This study and (Brodehl, Ferrier et al. 2016)
Ig15	c.5026G>A	p.Gly1676Arg	**RCM**	**Yes**	This study
Ig18	c.6053G>C	p.Arg2018Pro	**HCM/RCM**	*No*	Collaborators
Ig19	c.6173A>G	p.Gln2058Arg	*No disease association*	*No*	This study

Ig20 (IR)	c.6490T>C	p.Trp2164Cys	**HCM**	**Yes**	H. Watkins, personal communication
Ig20	c.6779A>G	p.Lys2260Arg	*No disease association*	*No*	This study
Ig20	c.6892C>T	p.Pro2298Leu	**RCM**	**Yes**	(Ader, De Groote et al. 2019)

3.8) Pathogenic mutations in FlnC Ig domains cluster to highly conserved Ig domain regions

A further important issue to illustrate and comprehend variants' pathogenicity is their location in sequence and 3D structure within the 24 FlnC Ig domains. To address this, 3D experimental structures or homology models of all 24 Ig domains were superposed on Ig5, which was selected as the reference structure since it shows the lowest average RMSD with other Ig domains. All superpositions were performed with PyMol (The PyMOL Molecular Graphics System, version 2.3, Schrödinger, LLC). The RMSD values ranged from 0.64 Å (Ig19 superposed to Ig5) to 2.467 Å (Ig22 superposed to Ig5), with an average value of 1.15 Å; and the percentage of Cα atom pairs included in the RMSD minimization ranged from 0.48 (Ig22 superposed to Ig5) to 0.89 (Ig13 superposed to Ig5), with an average value of 0.75.

The pathogenic mutation sites were mapped on the structure-based sequence alignment derived from the above superpositions. Mutations are clustered in a few 3D regions as shown in (**Figure 8A. MHSmap + MHSPconserv + PXSPF loop surrounding**), where the thickness and color of the ribbon are related to the pathogenic mutation frequency in each sequence position, ranging from narrow-blue where there are no pathogenic mutations to thick-red where mutations are frequent. Moreover, the frequently mutated sites are conserved in FlnC Ig domains, as shown in (**Figure 8B (MHSmap + MHSPconserv + PXSPF loop surrounding)**, where low and high conservation degrees are shown with blue-narrow and red-large ribbons, respectively.

3.9) Pathogenicity propensity upon mutation

The 108 pathogenic variants of human filamin C imply that, on average, there are 4.3 deleterious mutations per domain (one ABD domain and 24 Ig domains). However, the mutations are not equally distributed (**Figure 6. All variants overview**). Some domains are preserved: no pathogenic variants imply mutations in Ig3, Ig9, Ig11 and Ig13 – interestingly, these are odd numbers and might indicate a feeble evolutionary signal. Other domains, on the contrary, are heavily affected, especially in the N-terminal moiety (nine variants imply mutations in the ABD and seven in the Ig1 domain) and in the rod 2, where 6.8 deleterious mutations per domain are observed. The rod 2 mutational hotspot is the Ig20 domain, which hosts the mutations implied by 18 pathogenic variants.

Given that the 108 pathogenic variants of human FlnC are associated with a mutation from amino acid R1 to amino acid R2 it is possible to compute the propensity of each residue type to be R1. For example, the propensity of alanine to be R1, $p(Ala_{R1})$ is defined as:

$$p(Ala_{R1}) = \left(\frac{n_{Ala_{R1}}}{108}\right) * \left(\frac{n_{Ala}}{n_{all}}\right) \quad \text{(eq. 2)}$$

where n_{AlaR1} is the number of times alanine is observed in the 108 R1 positions, n_{Ala} is the number of alanine residues in human FlnC, and n_{all} = 2725 is the total number of residues in human FlnC (Uniprot Q14315-1). The numerator of equation 1 is the probability to observe an alanine in position R1 amongst the pathogenic variants and the denominator of equation 1 is the probability to observe an alanine in human FlnC. The propensity values are equal to one when the numerator and the denominator are equal, which means that the two probabilities are identical. On the contrary, the propensity is larger than one (or lower than one) if the amino acid tends to produce pathogenic variants upon its mutation (or it does not tend to produce pathogenic variants upon its mutation).

For obvious reasons it is not possible to define the propensity of each residue type to be R2, since the denominator of equation 1 is undetermined in that case.

The 20 propensity values reported in (**Table 6. Pathogenicity propensities upon mutation of the 20 amino acids**) are quite interesting.

Table 6. Pathogenicity propensities upon mutation of the 20 amino acids.
Different amino acids show different propensities to correlate upon mutation with pathogenic outcome in patients

Amino acid type	Propensity (acc. to equation 2)
R	3.666
M	3.058
P	1.584
W	1.402
I	1.318
A	1.294
N	1.216
H	1.183
G	1.086
T	1.062
V	0.924

S	0.751
Q	0.655
E	0.573
C	0.561
D	0.321
Y	0.315
F	0.274
L	0.180
K	0.000

Two amino acids, methionine and arginine, have high propensity to be R1, which means that their mutations are often associated with disease onset.
The high propensity of methionine is likely related to two facts. This residue is usually buried into the protein core and it is unique, with its long aliphatic side-chain, amongst apolar residues. Interestingly, the sulfur atom of the methionine side-chain may be involved in stabilizing chalcogen bonds (Iwaoka, Komatsu et al. 2002). Methionine mutations are therefore very likely deleterious for protein stability and, as a consequence, for human FlnC function.
On the contrary, the high propensity to be R1 of arginine is surprising, since this amino acid is usually exposed to the solvent and surface residues of globular proteins are known to be often mutable without major protein destabilization. Moreover, the high propensity of arginine does not compare with that of lysine, which is positively charged like arginine and which has the lowest propensity to be R1 (**Table 6. Pathogenicity propensities upon mutation of the 20 amino acids**).

Obviously, the sample of pathogenic variants is quite small and it would be temerarious to pretend definitive conclusions from these observations. It is nevertheless possible to hypothesize that arginine methylation, which is involved in numerous and diversified biological processes (Wu, Schapira et al. 2021) plays a role in human filamin C function.
Interestingly, in about one third of the mutations from Arg to R2, R2 is a cysteine. This means that replacements of arginines by cysteines have a considerable probability to be pathogenic. This might depend, at least in part, from the fact that solvent accessible cysteines can be easily oxidized and degraded, especially under stress.

4) Discussion

While some variants are pathogenic, others are perfectly tolerated by the organisms and constitute the natural genetic variability within populations that drives evolution. Even if pathogenic mutations are sometimes predictable, the molecular basis of pathogenicity grounds are often unclear. Multiple sequence alignments allow identifying amino acids conserved during evolution, enabling ascertaining the physicochemical reasons of their conservation based on their positions in the 3D structure. However, even when evolutionary and 3D information is available, mutation pathogenicity may remain elusive since it may affect post-translational modifications of adjacent residues, or/and modify gene expression rate, *etc.* The spectrum of possible hypotheses is large. Therefore, numerous biophysical and structural variables, which may be variably related to pathogenicity, must be considered, and machine learning techniques can be used to find satisfactory classifications of the data.

Many patients suffering from cardiomyopathies bear mutations in the FLNC gene, and therefore early identification of disease-associated mutations in this gene is becoming particularly important from a clinical perspective, given that next-generation sequencing is getting more accessible. While in vitro and in silico analysis will never replace necessary clinical studies to identify disease related mutations correctly, they certainly can suggest the need for further clinical examinations and prevention.

In the present study, a predictive, computational tool for automatic identification of pathogenic FlnC variants, AMIVA-F, has been developed, and comprehensive and systematic analysis of all available information on pathogenic variants has been performed.

4.1) AMIVA-F predictions

AMIVA-F is based on a multilayer perceptron neural network, a machine learning method. Its performance, estimated by 10-fold stratified cross validation and using an external test set, is remarkable: several figures of merit reach considerable high values (**Table 2. Summary of the quality of predictions**).

A user-friendly interface is freely distributed for allowing easy fully automated AMIVA-F computations.

Interestingly, AMIVA-F seems to perform better than other predictive techniques, like Polyphen2 (Adzhubei, Schmidt et al. 2010), SIFT (Ng and Henikoff 2003) and Provean (Choi, Sims et al. 2012), which were designed to predict changes in protein stability and are used in clinical diagnostics (Gómez, Lorca et al. 2017). By assuming that pathogenicity is associated with stability change and by using the same mutations used in the learning set of AMIVA-F, we observed that AMIVA-F outperforms all three predictors, with 78.6% accuracy against 73.4%, 64.2%, and 65.9% for Polyphen2, SIFT, and Provean, respectively (**Table 7. P15 comparisons of predictors**).

However, Polyphen2, SIFT, and Provean are protein stability predictors applicable to all types of proteins, while AMIVA-F was specifically designed to analyze variants in FlnC.

In principle, this machine learning approach, incorporating structural, biochemical, and biophysical attributes, can be adapted to analyze miss sense variants of any protein of medical interest, providing enough structural information - even generated computationally - and sufficiently large learning sets can be assembled.

Table 7. P15 comparisons of predictors

Comparison of predictors using 173 variants used in the learning set.

Predictor	AMIVA-F	Polyphen2	SIFT	Provean
TP[(a)] + TN[(b)]	136	127	111	114
Accuracy((TP+TN)/all mutations)	0.78%	0.73%	0.64%	0.66%

[(a)] TP = true positives: number of pathogenic mutations that are predicted to be pathogenic.
[(b)] TN = true negatives: number of non-pathogenic mutations that are predicted to be non-pathogenic.

4.2) Mechanistic impact of pathogenic variants

For the seven variants shown in (**Table 3. DSC results of known mutations together with AMIVA-F predictions (P7)**). - two non-pathogenic, five pathogenic, and all correctly predicted by AMIVA-F , we assessed the impact of these mutations on fold stability, by DSC analyses, and we evaluated their impact on the 3D structure, by solving the crystal structures of Ig14-15 wild-type, and of Ig14-15^{S1624L}, and Ig14-15^{G1676R} mutants (**Supplementary Table 4. Crystallographic Table**). The Ig14-15 construct was used to examine two of these mutations and the Ig19-21 construct was used to examine the other five.

As expected, the thermal stability of the two non-pathogenic mutants (p.Gln2058Arg in Ig19 and p.Lys2260Arg in Ig20) is substantially unaffected upon mutation (**Table 3. DSC results of known mutations together with AMIVA-F predictions (P7)**).

On the contrary, thermal stability is considerably reduced in two pathogenic mutants (p.Ser1624Leu and p.Gly1676Arg in Ig14-15), suggesting that their pathogenicity is due, at least in part, to decreased conformational stability (**Figure 4B (ABD_M82K +DSC+ FLNc_Ig14-15_structure)**). Furthermore, the unfolding pattern is clearly perturbed in p.Gly1676Arg, where the two domains (Ig 14 and Ig15) fold as a unique entity in the wild-type construct and as two separate entities upon mutation. This well agrees with the 3D structural data showing a perturbation of the hydrogen bonding network and the protein hydration upon mutation, especially at the interface between Ig14 and Ig15. Moreover, the p.Ser1624Leu mutation is located in the highly conserved and functionally important PXSP motif, which is present in nearly all Ig domains of human FlnC and was shown to be highly phosphorylated *in vivo* in several FlnC Ig domains (Reimann, Wiese et al. 2017, Reimann, Schwäble et al. 2020). It is thus reasonable to suppose that the disruption of phosphorylation capacity in p.Ser1624Leu is related to its pathogenicity.

A different trend is observed for the remaining three pathogenic mutants (p.Ile2160Phe, p.Trp2164Cys, and p.Pro2298Ser in Ig19-21), which show increased thermal stability of Ig20 upon mutation (**Figure 6A DSC Ig19-21 + 3D Model Ig19-21**). Two of these mutations (p.Ile2160Phe and p.Trp2164Cys) are close to or in the IR of Ig20, which is partially disordered (**Figure 8. (IDP+HCA). Prediction of intrinsic disorder and hydrophobicity cluster analysis for Ig20**.). We hypothesize that the folding stabilization, suggestive of an increased structuring of the surrounding of the mutation, can negatively impact interactions with binding partners, where structural plasticity is required. The third mutation (p.Pro2298Ser in Ig20) affects the first proline residue in the conserved PXSP motif, where the serine is not phosphorylated *in vivo* (Reimann, Schwäble et al. 2020), and which was shown to be involved in the regulation of mechanosensing in Ig20 of FlnA through *cis-trans* isomerisation of the second proline of the motif (Rognoni,

Möst et al. 2014). Interestingly, the p.Pro2298Ser mutation causes a destabilization of the adjacent domain Ig19. Furthermore, the altered stability of the three Ig20 mutants might perturb recognition of mechanically unfolded FlnC domain with co-chaperone, BAG3 responsible for chaperone-assisted selective autophagy (CASA) (Ulbricht, Felix et al. 2013), causing accumulation of aggregates and leading to HCM or RCM.

Therefore, it is pretty clear that the relationship between variant pathogenicity and thermal stability of the human FlnC domains is rather complex and somehow unexpected. Two issues are nevertheless quite clear: (i) pathogenicity is associated with perturbations of folding stability and (ii) can be associated with inter-domain interactions, suggesting that pathogenicity-stability studies need to be conducted not with single but multi domain constructs.

4.3) Systematic analysis of mutations in the Ig domains

We asked if there is a potential common feature shared between the pathogenic variants in our dataset. All FlnC Ig domains were superposed on one of them and all pathogenic variants were mapped to this reference structure. This revealed that pathogenic mutations cluster in regions intercalated, in the 3D structure, between the loop hosting the PXSP motif and the N-terminal side of the Ig domain (**Figure 9A (MHSmap + MHSPconserv + PXSP loop surrounding)**). The multiple sequence alignment of all human FlnC Ig domains (**Supplementary Figure S3 (Ig_domain_sequence_alignment)**) and projection of the amino acid conservation on the reference structure further showed that the mutational hotspots tend to be well conserved throughout the Ig domains of human FlnC (**Figure 9B (MHSmap + MHSPconserv + PXSP loop surrounding)**).

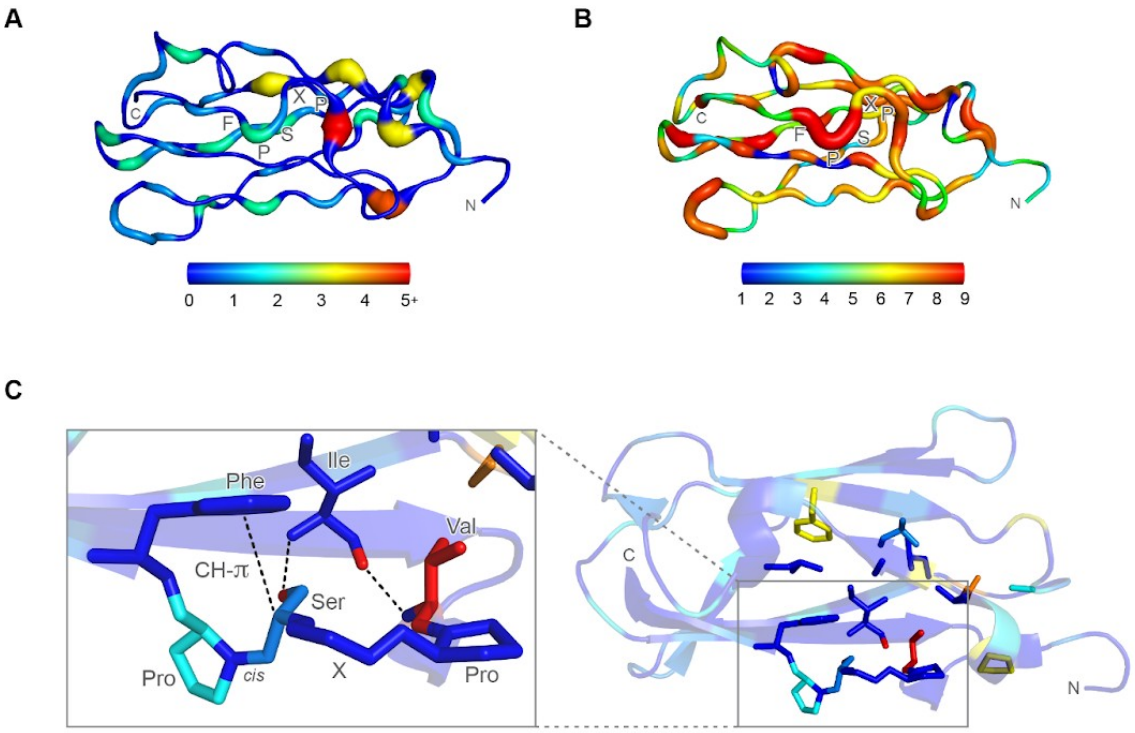


Figure 9. (MHSmap + MHSPconserv + PXSP loop surrounding).
Frequency of pathogenic mutations and amino acid conservation in human FlnC Ig domains mapped on the representative structure of Ig5. (A) Frequency of pathogenic mutations. Positions that do not host pathogenic mutations are shown in blue and with a narrow ribbon. Positions that host five (or plus) pathogenic mutations are shown in a red and thick ribbon. Residues with intermediate numbers of pathogenic mutations are represented according to the color code shown in the bar shown below; (B)

Degree of conservation of residues in human FlnC Ig domains. The conservation score, determined by ConSurf (Ashkenazy, Abadi et al. 2016) is mapped according to the bar shown below, with blue and narrow ribbon corresponding to low, and red and large ribbon corresponding to high conservation. PXSP motif is labeled in both panels; (C) PXSP loop and surrounding hydrophobic cluster, colored according to the mutational frequency as shown in A. Inlet: The crucial hydrogen-bond between the serine side-chain oxygen and Leu main-chain nitrogen residing on the adjacent β-strand is shown. A highly conserved hydrophobic residue colored in red (either valine or isoleucine in all Ig domains) forms another hydrogen main-chain main-chain hydrogen bond with Leu, further stabilising the structure. Carbonyl oxygens omitted with the exception of the relevant hydrogen bond engaging one for clarity.

A closer inspection of the residues, which are both well conserved amongst the Ig domains and the mutation of which is frequently pathogenic, informs on the biophysical and structural basis. These residues are either involved in extensive hydrogen bond networks or efficient hydrophobic packing. Any mutation that involves a change of side-chain volume will most likely not fit into the dense core without significant remodeling to avoid steric clashes or the formation of empty and apolar buried cavities.

The PXSP motif and its hydrophobic surrounding are highly conserved (**Figure 9B (MHSmap + MHSPconserv + PXSP loop surrounding)**). The peptide bond between the serine and the following proline residue raises attention being in *cis* conformation (**Figure 9C (MHSmap + MHSPconserv + PXSPF loop surrounding)**), which is infrequent due to steric clashes between side-chains of consecutive amino acids. It is estimated, thought, that roughly 5% of peptide bonds involving proline are in *cis* conformation, with peptide bonds between non- proline showing less than 0.5% *cis* conformational share (Stewart, Sarkar et al. 1990). Single-molecule mechanical experiments in FlnA revealed that this peptide bond in Ig20 is involved in mechanosensing modulation, regulated through *cis-trans* isomerization (Rognoni, Möst et al. 2014). It is known that CH-π electron interactions can further stabilize the *cis* conformation (Ganguly, Majumder et al. 2012). This interaction is observed between the phenylalanine following the PXSP motif and the CαH of the serine in the motif, preceding the *cis* proline involved in the *cis* petide bond (**Figure 9C, inlet (MHSmap + MHSPconserv + PXSPF loop surrounding)**). This feature can be observed in all experimentally determined Ig domain structures of FlnC. The importance of this stabilizing interaction is highlighted by high conservation of this position in human FlnC Ig domains (**Figure 9B, MHSmap + MHSPconserv + PXSP loop surrounding, Supplementary Figure 1. Multiple Ig-domain sequence alignment and the conserved PXSPF motif).** Moreover, this phenylalanine side-chain is in contact with an apolar aliphatic residue, the backbone of which forms two hydrogen bonds, one with the serine of the PXSP motif and one with a highly conserved Ile/Val residue that precedes the motif. The latter position is a pathogenic hotspot that cumulates seven pathogenic mutations (**Figure 9C, inlet (MHSmap + MHSPconserv + PXSP loop surrounding)**).

Given the importance of the phenylalanine residue in orchestrating both the PXSP motif and the stability of the Ig domain, we propose extending the motif to PXSPF.

Further evidence for the critical role of the PXSPF motif and specifically the serine residue comes from a phosphoproteomics study (Reimann, Schwäble et al. 2020), which revealed this site is phosphorylated in several Ig domains of FlnC, particularly in rod 2. This serine, mutated into leucine at position 1624 in Ig14, is one among several pathogenic mutations in the PXSPF motif. Based on its high conservation, we hypothesize that mutations of the PXSPF motif in other domains will likely be pathogenic as well. This is corroborated by the fact that both mutations p.Ser1624Leu in Ig14 and p.Pro2298Ser in Ig20, in the PXSPF motif, are pathogenic.

As Ig20 hosts a unique 82 amino acid insertion and since Ig20 presents the highest number of reported pathogenic mutations (**Figure 7. All variants overview.**), we looked into their spatial distribution in Ig20. Notably, only four mutations map to the insertion (p.Trp2164Cys, p.Arg2180Cys, p.Arg2187Cys and p.Arg2197Trp), while six mutations map to the PXSPF motif, highlighting it as a mutational hotspot. We further noticed that three mutations map to the beginning or end of the β-strand of Ig20, which binds to the Ig21 domain (β-strand C). In turn, mutations in the β-strand C of Ig21 also map to its beginning or end and could disturb this dynamic interdomain interaction responsible for mechanosensing and mechano-regulation of FlnC. Furthermore, amongst the most frequently occurring mutations is the replace-

ment of positively charged arginine (6 in Ig20, and 17 out of 108 pathogenic mutations reported in (**Figure 7. All variants overview.**), often with a cysteine, leading not only to a removal of a positive charge but potentially also increased aggregation propensity via disulfide bond formation under oxidative stress (Brennan, Wait et al. 2004). Pertinent details on the amino acid mutations involved in pathogenic variants, especially those involving arginine, can be found in the **Pathogenicity propensity upon mutation section.**

4.4) Conclusions

We developed a bioinformatics tool, AMIVA-F (https://pypi.org/project/AMIVA-F/), which was trained to predict the pathogenicity of human FlnC missense variants. It is distributed as a single user-friendly package that does not require advanced computational knowledge and comes with a step-by-step installation and execution guide. Its accuracy, estimated by internal cross-validation and validation with external and independent data, is close to 80% and outperforms alternative general pathogenicity predictors. Similar machine learning approaches can be customized to analyze missense variants of any given system, provided that structural information and sufficiently large learning sets are available.

Constructs bearing pathogenic variants showed systematically altered thermal stabilities *in vitro*. We further observed that mutations can affect the stability of adjacent Ig domains, which might be ascribed to interdomain interactions.

We found that many of the mutations cluster around the highly conserved and functionally relevant PXSP motif, which we propose to extend to the next residue (PXSPF), a well-conserved phenylalanine, which might stabilize the cis proline conformation.

Although pathogenicity predictions, based on biochemical and biophysical knowledge or computational and statistical techniques, cannot substitute clinical evaluations and analyses, they can assist triage in clinical settings. Moreover, they can allow a better understanding of the “sequence-structure-function-disease” relationships.

This is, nevertheless, just the first step towards the profound characterization of the molecular features of human FlnC. Further studies will require the systematic assessment of the relationship between pathogenicity and fold stability. This must be determined in overlapping multi-domain constructs, since mutations can affect the stability of adjacent domains. However, given that this deterministic experimental approach might be challenging, further refinement of machine learning techniques with inclusions on novel and accurate input attributes is needed.

5) Supplementary

Supplementary Figure 1. Multiple Ig-domain sequence alignment and the conserved PXSPF motif.
Sequence alignment of FLNc Ig-domains showing the high conservation of the PXSPF motif (cyan highlighted) in each domain. The first Pro in the motif happens to be in trans configuration while the second Pro forming a cis peptide bond with the preceeding Ser residue.

Supplementary Table 1. Pathogenic mutations dataset overview
Table of human FlnC mutations reported in this work. Mutants experimentally characterized in this work are highlighted in bold.

Domain	c-Notation	p-Notation	Phenotype	Reference
ABD	c.245T>G	p.Met82Lys	DCM	This study
ABD	c.318C>G	Phe106Leu	DCM	(IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018)
ABD	c.368T>C	Val123Ala	HCM	(IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018), (HALL; AKHTAR; SABATER-MOLINA; FUTEMA et al., 2020), (GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017), (VALDÉS-MAS; GUTIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014)
ABD	c.420C>A	His140Gln	HCM	(IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018), (EDEN; FREY, 2021)

ABD	?	Met145Val	HCM	(IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018), (EDEN; FREY, 2021)
ABD	c.547C>T	Thr160Lys	ACM	(IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018), (HALL; AKHTAR; SABATER-MOLINA; FUTEMA et al., 2020)
ABD	c.577G>A	Arg183Cys	HCM	(CUI; WANG; ZHANG; WU et al., 2018), (EDEN; FREY, 2021)
ABD	c.664A>G	Ala193Thr	distal myopathy	(DUFF; TAY; HACKMAN; RAVENSCROFT et al., 2011; IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018)
ABD	c.664A>G	Met222Val	DM PM MFM	(GEMELLI; PRADA; FIORILLO; FABBRI et al., 2019)
ABD	c.752T>C	Met251Thr	distal myopathy	(DUFF; TAY; HACKMAN; RAVENSCROFT et al., 2011; IWAMOTO; HUEHN; SIMON; HUET-CALDERWOOD et al., 2018)
Ig1	c.824C>T	Pro275Leu	HCM	(EDEN; FREY, 2021)
Ig1	c.870C>A	Asn290Lys	HCM	(VALDÉS-MAS; GU-TIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014), (GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig1	c.896C>T	Thr299Ile	HCM	(EDEN; FREY, 2021)
Ig1	c.925G>A	Glu309Lys	IBM , ALS, Dementia	(WEIHL; BALOH; LEE; CHOU et al., 2015)
Ig1	c.987C>A	Asn329Lys	HCM	(ALEJANDRA RE-STREPO-CORDOBA; CAMPUZANO; RIPOLL-VERA; COBO-MARCOS et

				al., 2017)
Ig1	c.986A>G	Asn329Ser	HCM	(EDEN; FREY, 2021), (ALEJANDRA RE-STREPO-CORDOBA; CAMPUZANO; RIPOLL-VERA; COBO-MARCOS et al., 2017)
Ig1	c.1076T>C	Ile359Thr	HCM	(EDEN; FREY, 2021)
Ig2	c.1132G>T	Val378Leu	HCM	(EDEN; FREY, 2021)
Ig2	c.1322G>T	Arg441Ile	DCM	(XIAO; WEI; WU; LIU *et al.*, 2020)
Ig4	c.1882G>A	Val628Met	HCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig4	c.1942G>T	Asp648Tyr	MFM, DM	(ZHANG; PU; BAN; LIU et al., 2018)
Ig5	c.2050G>C	Val684Leu	HCM	(CUI; WANG; ZHANG; WU et al., 2018),
Ig5	c.2170G>A	Gly724Ser	HCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig6	c.2375G>T	Ser792Ile	HCM	(JAAFAR; GÓMEZ; KAMMOUN; ZAIRI et al., 2016)
Ig6	c.2425G>A	Val809Met	DCM	(JANIN; N'GUYEN; HABIB; DAUPHIN et al., 2017)
Ig6	c.2450T>C	Ile817Thr	HCM	(CIRINO ALLISON; LAK-DAWALA NEAL; MCDONOUGH; CONNER et al., 2017)
Ig7	c.2587C>T	Pro863Ser	HCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig7	c.2672C>T	Thr891Met	HCM	(EDEN; FREY, 2021)
Ig7	c.2737G>A	Glu913Lys	HCM	(CUI; WANG; ZHANG; WU et al., 2018)

Ig7	c.2839G>C	Gly947Ar	HCM	(EDEN; FREY, 2021)
Ig8	c.2888G>A	Pro963Arg	DCM	(XIAO; WEI; WU; LIU *et al.*, 2020)
Ig8	c.3092>X	Pro1031Leu	HCM	(EDEN; FREY, 2021)
Ig10	c.3547_3548delinsCT	Ala1183Leu	RCM	(KISELEV; VAZ; KNYAZEVA; KHUDIAKOV et al., 2018)
Ig10	c.3557C>T	Ala1186Val	PM, DM	(GHAOUI; COOPER; LEK; JONES et al., 2015)
Ig10	c.3581C>T	Ser1194Leu	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig10	c.3592>X	Val1198Gly	DCM	(XIAO; WEI; WU; LIU *et al.*, 2020)
Ig10	c.3623C>T	Ala1208Val	HCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig10	c.3646T>A	Tyr1216Asn	HCM/MFM	(AVILA-SMIRNOW; BÉHIN; GUENEAU; CLAEYS et al., 2010), (VALDÉS-MAS; GU-TIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014)
Ig10	c.3706C>T	Pro1236Ser	muscular phenotype	(YU; ZHENG; JIN; GANG et al., 2017)
Ig10	c.3721C>T	Arg1241Cys	Inclusion body formation	(WEIHL; BALOH; LEE; CHOU et al., 2015)
Ig12	c.4055G>A	Arg1352His	HCM	(EDEN; FREY, 2021)
Ig12	c.4222G>A	Glu1408Asp	HCM	(EDEN; FREY, 2021)
Ig12	c.4270G>A	Gly1424Val	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig14	c.4615G>A	Ala1539Thr	HCM	(VALDÉS-MAS; GU-TIÉRREZ-FERNÁNDEZ;

				GÓMEZ; COTO et al., 2014)
Ig14	c.4636G>A	Gly1546Ser	RCM	(SANOJA; LI; FRICKER; KINGSMORE et al., 2018)
Ig14	c.4651G>A	Ala1551Thr	ALS Dementia	(JANSSENS; PHILTJENS; KLEINBERGER; VAN MOSSEVELDE et al., 2015)
Ig14	c.4700G>A	Arg1567Gln	DCM	(ESSLINGER; GARNIER; KORNIAT; PROUST *et al.*, 2017; VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN *et al.*, 2020)
Ig14	c.4795A>G	Thr1599Ala	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig14	**c.4871C>T**	**Ser1624Leu**	**RCM**	**(BRODEHL; FERRIER; HAMILTON; GREENWAY et al., 2016), This study, DSC data**
Ig14	c.4916G>A	Cys1639Tyr	RCM	(XIAO; WEI; WU; LIU *et al.*, 2020)
Ig15	c.4997T>C	Ile1666Thr	NCCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig15	**c.5026G>A**	**Gly1676Arg**	**RCM**	**This study, DSC data**
Ig15	c.5042C>T	Thr1681Met	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig15	c.5068C>T	Leu1690Phe	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig15	c.5071G>A	Asp1691Asn	MFM	(ZHANG; PU; BAN; LIU et al., 2018)
Ig15	c.5125C>T	Pro1709Ser	HCM	(CUI; WANG; ZHANG; WU et al., 2018)

Ig15	c.5132C>T	Pro1711Leu	HCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig16	c.5278G>A	Gly1760Ser	MFM	(YU; ZHENG; JIN; GANG et al., 2017)
Ig16	c.5524G>A	Asn1843Ser	HCM	(EDEN; FREY, 2021)
Ig17	c.5644A>G	Ile1882Val	ACM	(HALL; AKHTAR; SABATER-MOLINA; FUTEMA et al., 2020)
Ig17	c.5807A>C	His1936Pro	HCM	(EDEN; FREY, 2021), (CUI; WANG; ZHANG; WU et al., 2018)
Ig18	c.5888C>T	Thr1963Met	HCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig18	c.5954C>T	Ser1985Leu	HCM	(VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020)
Ig18	c.5996G>A	Arg1999Gln	HCM	(JAAFAR; GÓMEZ; KAMMOUN; ZAIRI et al., 2016)
Ig18	c.6032G>A	Gly2011Glu	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019), (GARCIA HERNANDEZ; ORTIZ-GENGA; ANALIA RAMOS; OCHOA et al., 2020)
Ig18	c.6053G>A	Arg2018His	HCM	(CHANAVAT; JANIN; MILLAT, 2016)
Ig19	c.6115G>A	Gly2039Arg	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig19	c.6134G>A	Arg2045Gln	HCM	(CHANAVAT; JANIN; MILLAT, 2016)
Ig19	c.6205G>A	Ala2069Thr	HCM	((VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020); VANHOUTTE; CLAES;

				HELDERMAN-VAN DEN ENDEN et al., 2020)
Ig19	c.6208G>A	Gly2070Ser	RCM, ACM	(ORTIZ-GENGA; CUENCA; DAL FERRO; ZORIO et al., 2016)
Ig20	c.6397C>T	Arg2133Cys	HCM	(CUI; WANG; ZHANG; WU et al., 2018), (VAN WANING; HOEDEMAEKERS; TE RIJDT; JPMA et al., 2019)
Ig20	c.6398G>A	Arg2133His	HCM	(VALDÉS-MAS; GUTIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014), (VAN WANING; HOEDEMAEKERS; TE RIJDT; JPMA et al., 2019)
Ig20	c.6419G>A	Arg2140Gln	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig20	c.6451G>A	Gly2151Ser	HCM	(VALDÉS-MAS; GUTIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014)
Ig20	**c.6478A>T**	**Ile2160Phe**	**RCM**	**(BRODEHL; FERRIER; HAMILTON; GREENWAY et al., 2016), DSC data**
IR	**?**	**Trp2164Cys**	**HCM**	**DSC / Perry Elliot personal communication**
IR	c.6538C>T	Arg2180Cys	ACM	(VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020)
IR	c.6559C>T	Arg2187Cys	ACM	(VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020)
IR	c.6589C>T	Arg2197Trp	HCM	(ALEJANDRA RESTREPO-CORDOBA; CAMPUZANO; RIPOLL-VERA; COBO-MARCOS et al., 2017)

Ig20	c.6808G>A	Glu2270Lys	muscular phenotype	(VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020)
Ig20	c.6824G>T	Ser2275Ile	muscular phenotype	(VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020)
Ig20	c.6889G>A	Val2297Met	RCM	(TUCKER NATHAN; MCLELLAN MICHEAL; HU; YE *et al.*, 2017)
Ig20	**c.6892C>T**	**Pro2298Ser**	**RCM**	**(SCHUBERT; TARIQ; GEDDES; KINDEL et al., 2018), This study**
Ig20	c.6892C>T	Pro2298Leu	HCM	(SCHUBERT; TARIQ; GEDDES; KINDEL et al., 2018)
Ig20	c.6895G>A	Gly2299Ser	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig20	c.6902C>T	Pro2301Leu	HCM	(ROLDÁN-SEVILLA; PALOMINO-DOZA; DE JUAN; SÁNCHEZ et al., 2019)
Ig20	c.6901C>G	Pro2301Ala	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig20	c.6943C>A	His2315Asn	HCM	(VALDÉS-MAS; GU-TIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014)
Ig21	c.6952C>T	Arg2318Trp	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig21	c.7030G>A	Ala2344Thr	HCM	(CUI; WANG; ZHANG; WU et al., 2018), (EDEN; FREY, 2021)
Ig21	c.7034G>A	Gly2345Glu	congenital heart disease	(KOSMICKI; SAMOCHA; HOWRIGAN; SANDERS et al., 2017)

Ig21	c.7076T>C	Ile2359Thr	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig21	c.7123G>T	Val2375Phe	HCM	(GÓMEZ; LORCA; REGUERO JULIAN; MORÍS et al., 2017)
Ig21	c.7123G>C	Val2375Leu	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig21	c.7123G>A	Val2375Ile	MCM	(CUI; WANG; ZHANG; WU et al., 2018)
Ig21	c.7177C>T	Pro2393Ser	noncompaction cardiomyopathy	(KOSMICKI; SAMOCHA; HOWRIGAN; SANDERS et al., 2017), (VAN WANING; HOEDEMAEKERS; TE RIJDT; JPMA et al., 2019)
Ig21	?	Pro2396Thr	HCM	(EDEN; FREY, 2021)
Ig22	c.7228C>T	Arg2410Cys	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig22	c.7250A>C	Gln2417Pro	HCM	(ADER; DE GROOTE; RÉANT; ROORYCK-THAMBO et al., 2019)
Ig22	c.7256C>T	Thr2419Met	MFM	(TASCA; ODGEREL; MONFORTE; AURINO et al., 2012)
Ig22	c.7289C>T	Ala2430Val	HCM	(VALDÉS-MAS; GUTIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014), (SCHÄNZER; SCHUMANN; ZENGELER; GULATZ et al., 2021)
Ig22	c.7484G>A	Arg2495His	HCM	(VALDÉS-MAS; GUTIÉRREZ-FERNÁNDEZ; GÓMEZ; COTO et al., 2014)
Ig23	c.7514C>T	Pro2505Leu	HCM	(CUI; WANG; ZHANG; WU et al., 2018)

Ig23	c.7688A>G	Tyr2563Cys	RCM	(SCHUBERT; TARIQ; GEDDES; KINDEL et al., 2018)
Ig23	c.7724T>A	Ile2575Asn	PM, CON	((VERDONSCHOT; VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020); VANHOUTTE; CLAES; HELDERMAN-VAN DEN ENDEN et al., 2020)
Ig23	c.7781G>C	Gly2594Ala	HCM	(CHANAVAT; JANIN; MILLAT, 2016)
Ig24	c.7942G>T	Ala2648Ser	Arrhytmia	(XIAO; WEI; WU; LIU *et al.*, 2020)

Supplementary Table 2. SGnaomAD-negative Dataset.

Mutations selected to represent the non pathogenic dataset in AMIVA-F. All mutations were selected from GnomAD database and fulfil the criteria outlined in **Mutational variants present in the training set**.

Domain	Chromosome	c-Notation	p-Notation	Allele Count	Allele Number	Allele Frequency	Homo-zygote Count
ABD	7	c.22T>C	p.Ser8Pro	31	275298	0.00011261	0
Ig1	7	c.1081C>T	p.Arg361Cys	49	280936	0.00017442	0
Ig2	7	c.1108A>G	p.Met370Val	31	280940	0.00011034	0
	7	c.1142G>A	p.Arg381His	31	280900	0.00011036	0
	7	c.1147C>G	p.Pro383Ala	31	249528	0.00012424	0
	7	c.1166G>A	p.Gly389Asp	35	280866	0.00012	0

						462	
	7	c.1354G>A	p.Val452Met	38	279848	0.00013579	0
Ig3	7	c.1474A>G	p.Lys492Glu	73	280852	0.00025992	0
	7	c.1519G>A	p.Gly507Arg	221	280698	0.00078732	0
	7	c.1576C>T	p.Arg526Trp	46	280892	0.00016376	**1**
	7	c.1577G>A	p.Arg526Gln	614	280918	0.00218569	**4**
	7	c.1600G>A	p.Glu534Lys	235	280962	0.00083641	0
	7	c.1657G>A	p.Gly553Ser	76	280938	0.00027052	0
Ig4	7	c.1976T>G	p.Leu659Arg	102	280798	0.00036325	**1**
Ig5	7	c.2068T>C	p.Phe690Leu	48	280970	0.00017084	0
	7	c.2078A>C	p.Asp693Ala	967	280970	0.00344165	**2**
	7	c.2125G>A	p.Ala709Thr	295	279720	0.00105463	**1**
	7	c.2128G>A	p.Asp710Asn	86	279874	0.00030728	0
	7	c.2180G>A	p.Arg727His	108	280538	0.00038	0

						498	
Ig6	7	c.2296C>T	p.Arg766Trp	21	183654	0.00011435	0
	7	c.2392G>A	p.Asp798Asn	63	279436	0.00022545	**1**
	7	c.2501C>T	p.Thr834Met	2083	280520	0.00742 55	**62**
	7	c.2507C>A	p.Pro836Gln	193	280496	0.00068807	**0**
Ig7	7	c.2635C>T	p.Arg879Cys	39	277130	0.00014073	0
	7	c.2686G>A	p.Gly896Arg	110	280648	0.00039195	0
	7	c.2839G>C	p.Gly947Arg	45	280820	0.00016025	0
Ig8	7	c.3145G>T	p.Gly1049Cys	43	278222	0.00015455	0
	7	c.3146G>T	p.Gly1049Val	43	278172	0.00015458	0
Ig9	7	c.3242C>T	p.Ala1081Val	36	280412	0.00012838	0
	7	c.3304C>T	p.Pro1102Ser	30	280776	0.00010685	0

Ig10	7	c.3506A>G	p.Lys1169Arg	283	280580	0.00100863	**2**
	7	c.3721C>T	p.Arg1241Cys	1729	279050	0.00619602	**17**
Ig11	7	c.3757G>A	p.Val1253Ile	365	278306	0.00131151	**5**
	7	c.3812C>G	p.Thr1271Ser	46	248868	0.00018484	0
	7	c.3847A>G	p.Thr1283Ala	32	280250	0.00011418	0
	7	c.3938G>A	p.Arg1313Gln	57	280412	0.00020327	0
	7	c.4022G>A	p.Arg1341Gln	264	280120	0.00094245	**1**
Ig12	7	c.4097A>G	p.Asn1366Ser	44	273358	0.00016096	0
Ig13	7	c.4459G>A	p.Val1487Met	60	278014	0.00021582	**1**
	7	c.4553A>G	p.Lys1518Arg	52	264562	0.00019655	0
Ig14	7	c.4651G>A	p.Ala1551Thr	33	280162	0.00011779	0
	7	c.4700G>A	p.Arg1567Gln	20774	279986	0.07419657	**872**

	7	c.4763C>G	p.Ala1588Gly	45	280652	0.00016034	0
Ig15	7	c.4970G>A	p.Arg1657Gln	29	280592	0.00010335	0
	7	c.5020G>A	p.Gly1674Ser	38	280830	0.00013531	0
	7	c.5042C>G	p.Thr1681Arg	199	280908	0.00070842	0
	7	c.5143G>A	p.Val1715Ile	62	280826	0.00022078	0
Ig16	7	c.5278G>A	p.Gly1760Ser	29	176390	0.00016441	0
	7	c.5284C>T	p.Arg1762Cys	44	175150	0.00025121	0
	7	c.5311C>G	p.Pro1771Ala	71	280832	0.00025282	0
	7	c.5374G>A	p.Ala1792Thr	68	280634	0.00024231	0
	7	c.5375C>T	p.Ala1792Val	27	249226	0.00010834	0
Ig17	7	c.5578C>T	p.Arg1860Cys	1463	280876	0.00520 87	**9**
	7	c.5644A>G	p.Ile1882Val	312	280886	0.00111077	**1**
	7	c.5764G>A	p.Ala1922Thr	415	282370	0.00146 97	**4**

Ig18	7	c.5944C>T	p.Arg1982Cys	45	280056	0.00016068	0
	7	c.5954C>T	p.Ser1985Leu	56	280118	0.00019992	0
Ig19	7	c.6175G>A	p.Val2059Met	238	279776	0.00085068	0
Ig20	7	c.6808G>A	p.Glu2270Lys	193	260954	0.00073959	0
	7	c.6865G>A	p.Ala2289Thr	35	265342	0.00013191	0
Ig21	7	c.6988G>A	p.Gly2330Ser	166	277392	0.00059843	**2**
	7	c.6991G>A	p.Val2331Met	129	277812	0.00046434	0
	7	c.7091G>A	p.Arg2364His	484	280776	0.00172379	**1**
Ig22	7	c.7256C>T	p.Thr2419Met	42	245700	0.00017094	0
	7	c.7289C>T	p.Ala2430Val	28	278082	0.00010069	0
	7	c.7291G>A	p.Val2431Met	47	246748	0.00019048	0

Ig23	7	c.7614G>T	p.Leu2538Phe	435	279574	0.00155594	3
Ig24	7	c.8003T>C	p.Met2668Thr	219	262788	0.00083337	0

Supplementary Table 3. Technical details.
Technical details on the algorithm used to predict human FlnC variants pathogenicity.

Algorithm	Multilayer perceptron
Batch size	100
Hidden layers	a,2,3
Learning rate	0.78
Momentum	0.2
Epochs	517
Threshold function	sigmoid
Validation threshold	20
Seed	0
Filters	Nominal to binary, normalized attributes, normalized numeric classes

Supplementary Table 4. Crystallographic Table

Data collection	Ig_d14-15 (7OUU)	Ig_14-15^{S1624L} (7OUV)	Ig_14-15^{G1676R} (7P0E)

Wavelength (Å)	0.9686	0.9795	0.9655
Resolution range (Å)	19.83 - 1.47 (1.523 - 1.47)[a]	37.97 - 1.8 (1.864 - 1.8)[a]	38.26 - 1.6 (1.657 - 1.6)[a]
Space group	P 21 21 21	P 21 21 21	P 21 21 21
Unit cell (Å)	45.04 58.82 178.31 90 90 90	39.133 70.498 156.937 90 90 90	39.15 59.46 180.07 90 90 90
Total reflections	792556 (20899)	520298 (54865)	212835 (18674)
Unique reflections	71856 (3680)	41232 (4057)	53624 (4752)
Multiplicity	11.0 (5.7)	12.6 (13.5)	4.0 (3.9)
Completeness (%)	87.97 (45.84)	99.64 (99.80)	94.02 (79.64)
Mean *I*/sigma (*I*)	11.10 (0.11)	9.99 (0.76)	9.97 (0.50)
Wilson B-factor (Å^2)	24.15	33.68	29.81
Refinement			
R-merge (%)	0.1136 (9.86)	0.1469 (2.913)	0.06401 (2.205)
R-meas (%)	0.1187 (10.84)	0.1535 (3.027)	0.0733 (2.508)
R-pim (%)	0.03397 (4.376)	0.04388 (0.8159)	0.03494 (1.166)
CC1/2	0.999 (0.019)	0.998 (0.478)	0.998 (0.134)
CC*	1 (0.193)	1 (0.804)	0.999 (0.486)
Reflections used in refinement	71848 (3680)	41103 (4050)	53270 (4458)
Reflections used for *R*-free	1903 (97)	1417 (140)	1446 (121)
R-work	0.1881 (0.3991)	0.2134 (0.3352)	0.1859 (0.3857)

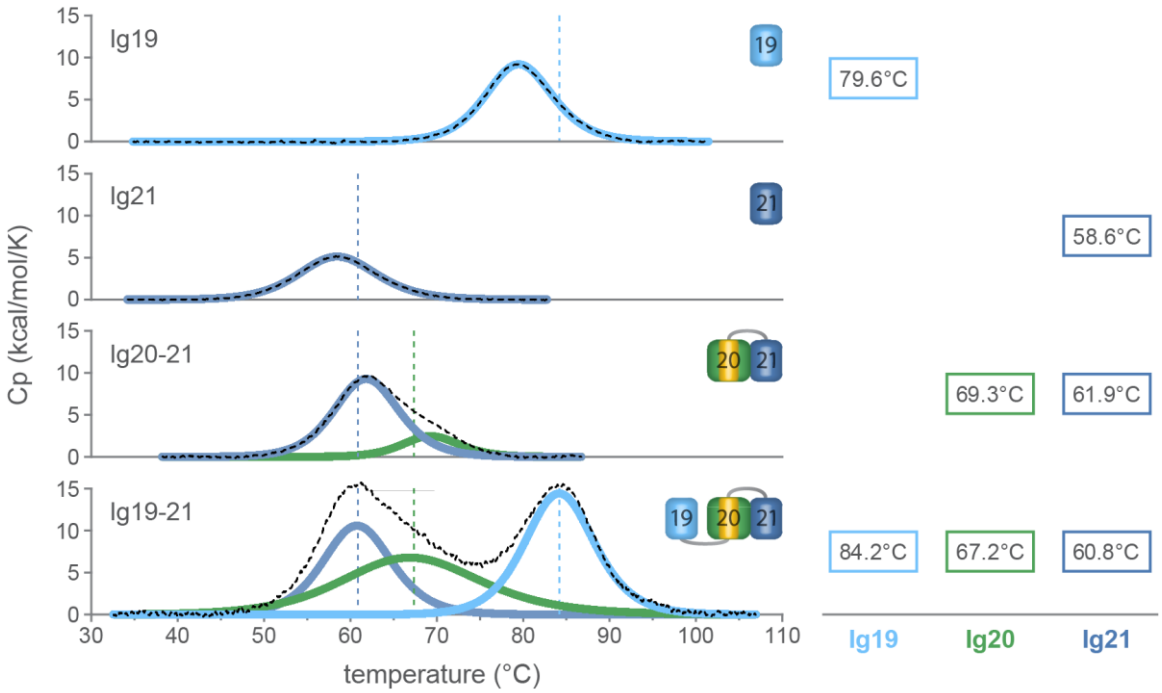
R-free	0.2277 (0.3750)	0.2472 (0.3460)	0.2211 (0.4270)
CC (work)	0.955 (0.103)	0.938 (0.722)	0.964 (0.399)
CC (free)	0.954 (0.096)	0.940 (0.599)	0.938 (0.285)
No. of non-hydrogen atoms	3379	3376	3376
No. of atoms macromolecules	3015	3032	3028
No. of atoms in ligands	0	0	46
No. of solvent atoms	364	344	320
Protein residues	406	406	404
RMS (bonds) (Å)	0.002	0.004	0.015
RMS (angles) (°)	0.54	0.68	1.24
Ramachandran favored (%)	98.51	98.51	97.5
Ramachandran allowed (%)	1.49	1.24	2.5
Ramachandran outliers (%)	0	0.25	0
Rotamer outliers (%)	0	0.61	0
Clashscore	0.34	2.5	3.79
Average B-factor ($Å^2$)	35.02	42.36	38.93
macromolecules ($Å^2$)	34.25	42.19	38.47
ligands ($Å^2$)	-	-	50.69
solvent ($Å^2$)	41.39	43.86	42.31

[a] Values in parentheses represent the highest resolution shell.

Supplementary Table 5. (P16).

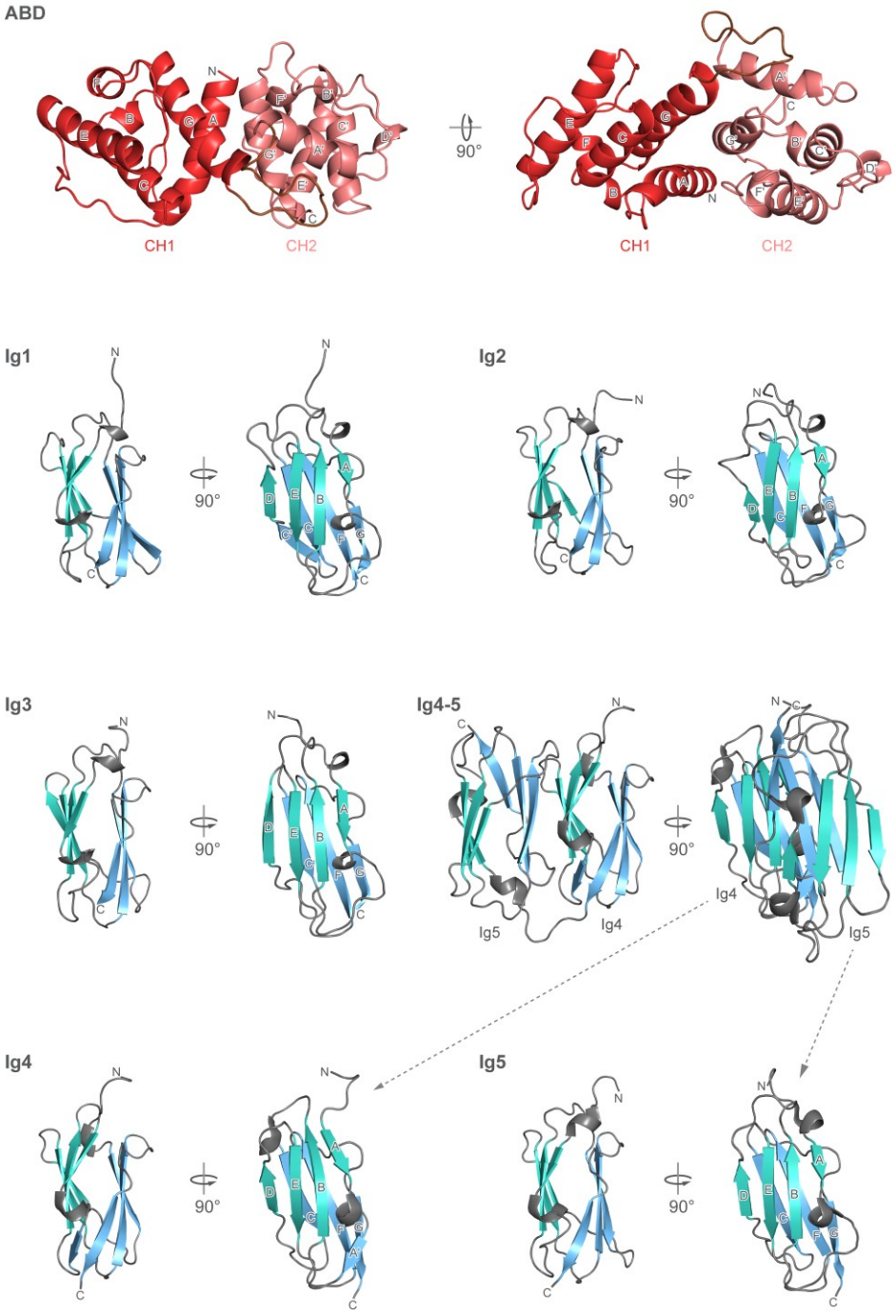
Pathogenicity prediction of 16 mutations absent from the learning sets used for neural network training, which are rarely observed in the gnomeAD database, and for which the pathogenicity is unknown. Pathogenic predictions are indicated in **bold**, neutral in *italics*.

Domain	c-Notation	p-Notation	Pathogenicity prediction (AMIVA-F)
ABD	c.431C>T	**Ser144Phe**	**Yes**
Ig1	c.1019A>G	**Tyr340Cys**	**Yes**
Ig1	c.1027A>G	*Lys343Glu*	*No*
Ig3	c.1649C>A/G/T	*Thr550Met*	*No*
Ig5	c.2272G>A	**Val758Met**	**Yes**
Ig8	2911G>A/T	**Val971Ile**	**Yes**
Ig11	c.3742G>A/T	*Val1248Phe*	*No*
Ig11	c.3905C>T	*Thr1302Ile*	*No*
Ig14	c.5036C>A	**Thr1679Lys**	**Yes**
Ig15	c.5205T>G	**Thr1735Trp**	**Yes**
Ig16	c.5333G>A	**Arg1811Gln**	**Yes**
Ig17	c.5764G>A/T	*Pro1920Ser*	*No*
Ig19	c.6133C>A/T	**Arg2045Trp**	**Yes**
Ig20	c.6808G>A/C	*Glu2270Gln*	*No*
Ig20	c.6809A>C/T	*Glu2270Ala*	*No*
Ig22	c.7302C>G/T	**Asn2434Lys**	**Yes**

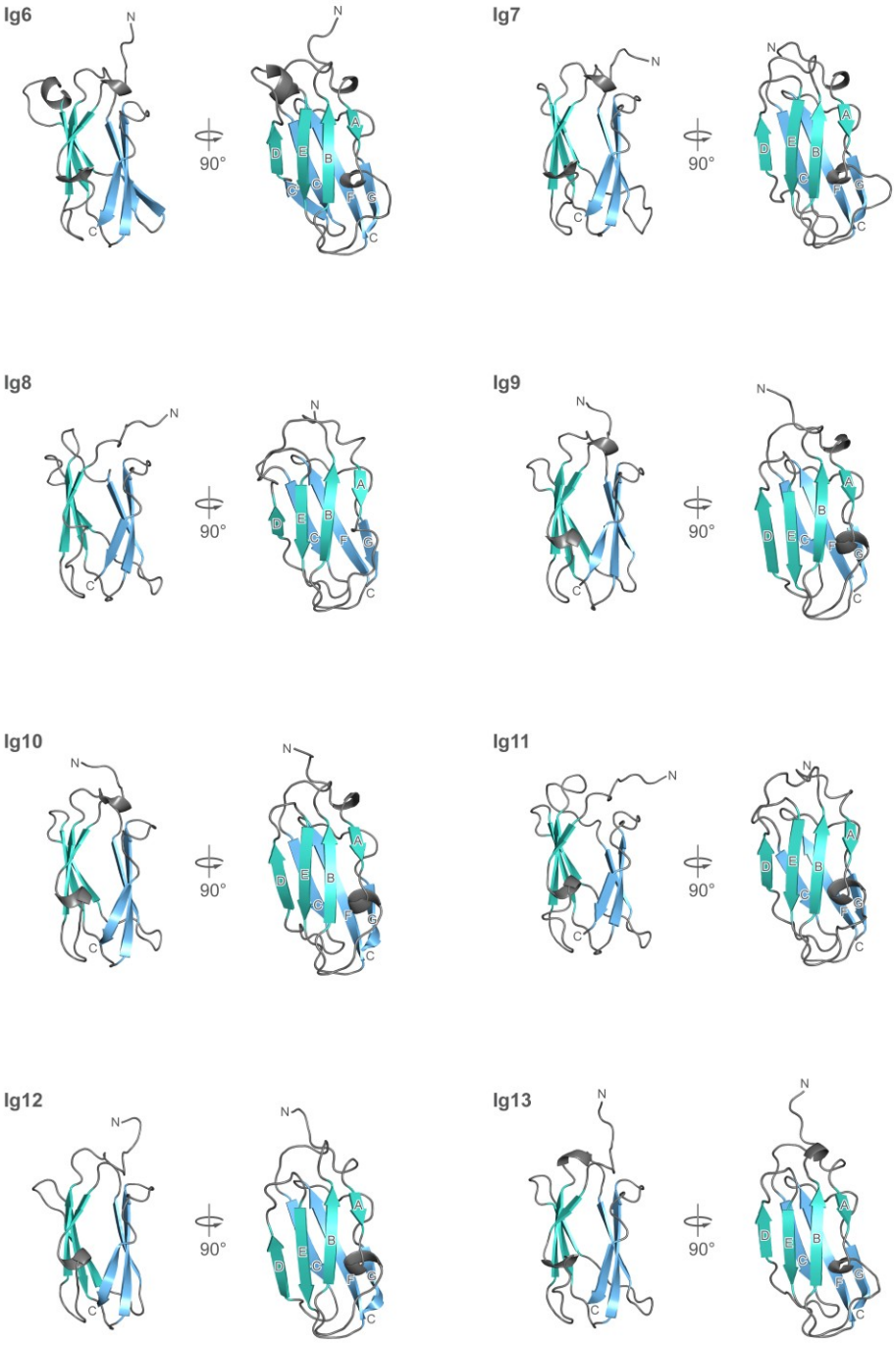


Supplementary Figure 2. DSC 19-21 assignment.

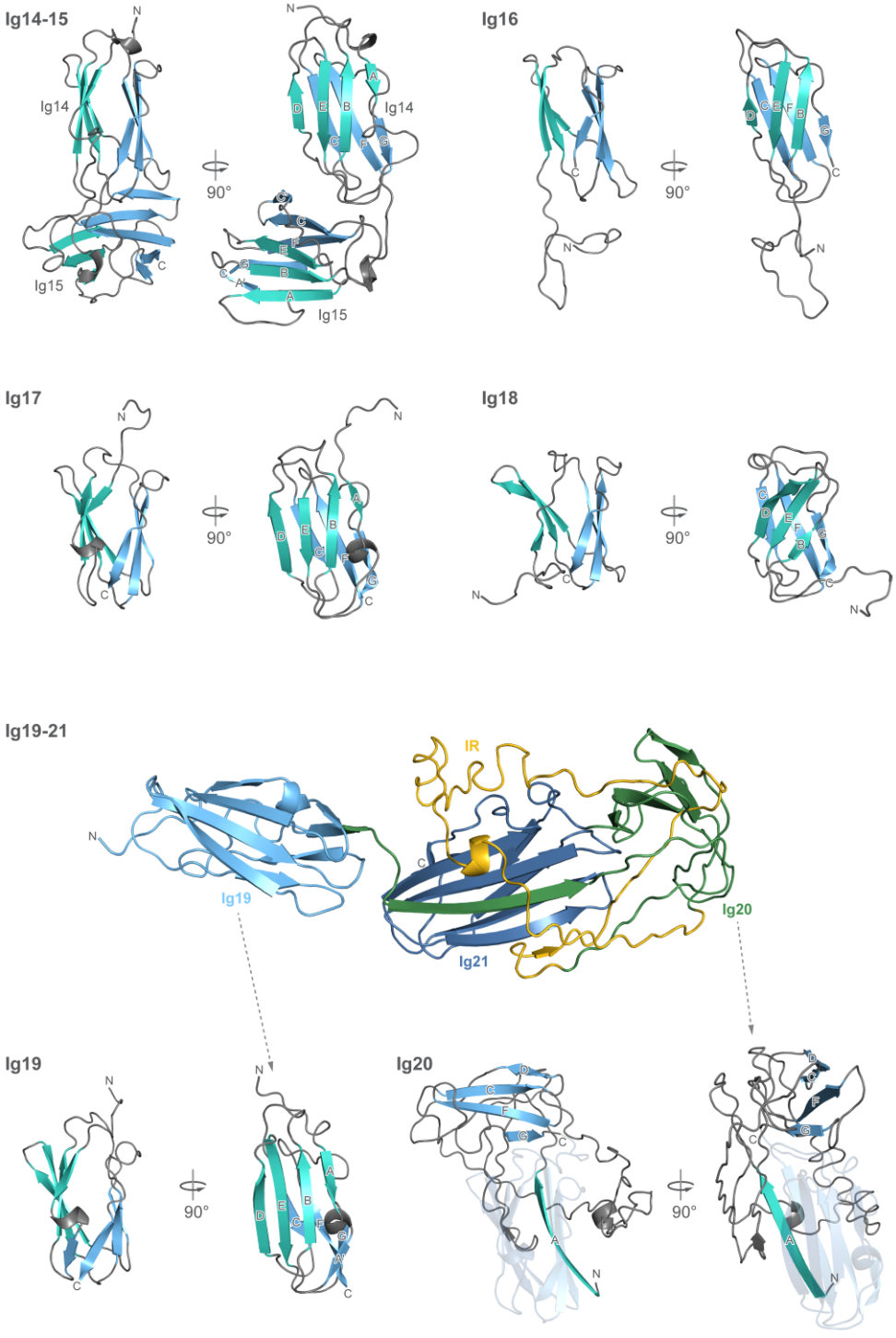
The single domain constructs Ig19 and Ig21 together with Ig20-21 are used to identify the corresponding Ig domains in the Ig19-21 construct below. Due to aggregation issues of Ig20 alone, we used Ig20-21 where we identified Ig20 around 69.3°C. Dotted lines indicate the melting temperatures for the proposed Ig domains in Ig19-21 in each single domain construct, validating our assignment based on the single domain construct T_ms.



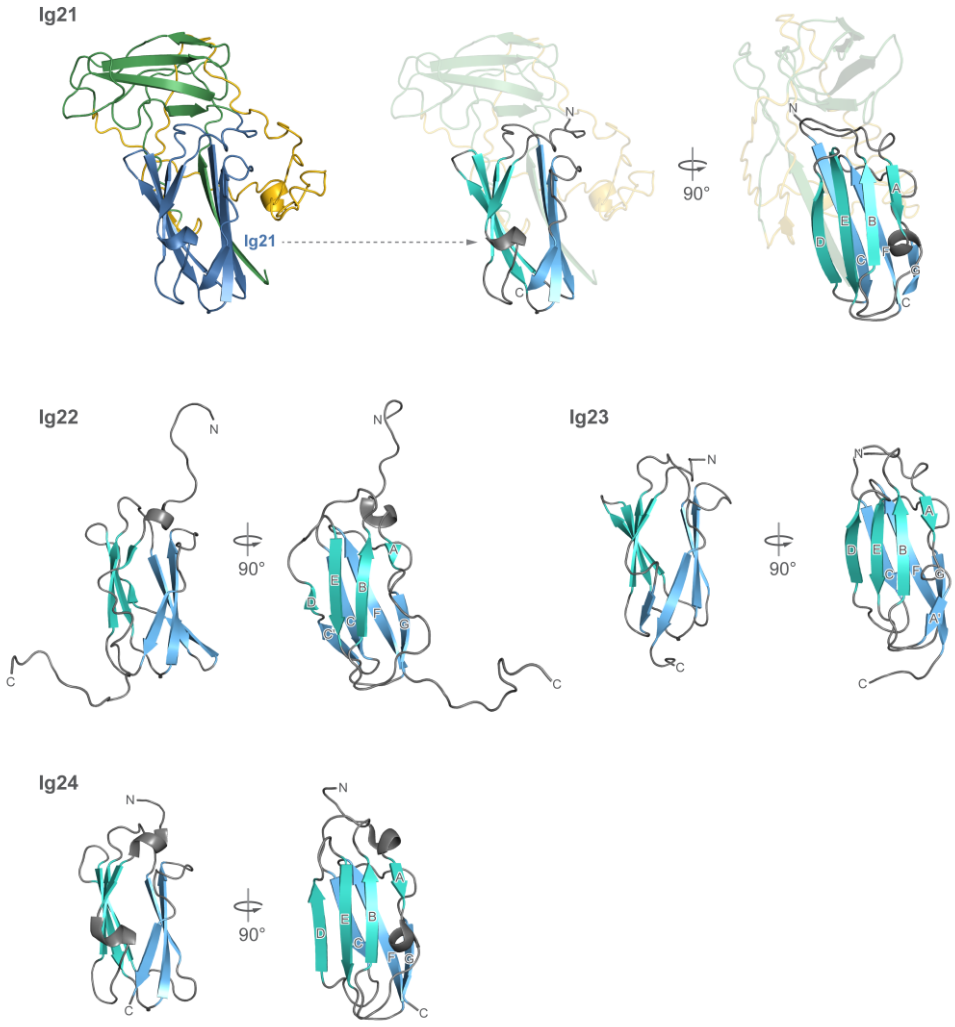
Supplementary Figure 3. Topology and architecture of human FLNc Part I.
FLNc ABD and Ig1 to Ig5



Supplementary Figure 3. Topology and architecture of human FLNc Part II.
FLNc Ig6 to Ig13



Supplementary Figure 3. Topology and architecture of human FLNc Part III.
FLNc Ig14 to Ig21
Ig14-15 and Ig19-21 depicted as multi-domain constructs due to interdomain interactions.



Supplementary Figure 4. Topology and architecture of human FLNc Part IV.
FLNc Ig21 to Ig24
Ig21 shown with fainted Ig20 in the background to emphasize on interdomain interactions.

References:

ADAMS, P. D.; AFONINE, P. V.; BUNKÓCZI, G.; CHEN, V. B. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. **Acta Crystallographica Section D Biological Crystallography**, 66, n. 2, p. 213-221, 2010.

ADER, F.; DE GROOTE, P.; RÉANT, P.; ROORYCK-THAMBO, C. *et al.* FLNC pathogenic variants in patients with cardiomyopathies: Prevalence and genotype-phenotype correlations. **Clinical Genetics**, 96, n. 4, p. 317-329, 2019/10/01 2019. https://doi.org/10.1111/cge.13594.

ADER, F.; DE GROOTE, P.; RÉANT, P.; ROORYCK-THAMBO, C. *et al.* FLNC pathogenic variants in patients with cardiomyopathies: Prevalence and genotype-phenotype correlations. **Clinical Genetics**, 96, n. 4, p. 317-329, 2019.

ADZHUBEI, I. A.; SCHMIDT, S.; PESHKIN, L.; RAMENSKY, V. E. *et al.* A method and server for predicting damaging missense mutations. *In*: **Nat Methods**, 2010. v. 7, p. 248-249.

ASHKENAZY, H.; ABADI, S.; MARTZ, E.; CHAY, O. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. **Nucleic Acids Research**, 44, n. W1, p. W344-W350, 2016.

BANANI, S. F.; LEE, H. O.; HYMAN, A. A.; ROSEN, M. K. Biomolecular condensates: organizers of cellular biochemistry. **Nature Reviews Molecular Cell Biology**, 18, n. 5, p. 285-298, 2017.

BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G. *et al.* The Protein Data Bank. **Nucleic Acids Res**, 28, n. 1, p. 235-242, Jan 1 2000.

BERNSTEIN, F. C.; KOETZLE, T. F.; WILLIAMS, G. J.; MEYER, E. F., JR. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. **J Mol Biol**, 112, n. 3, p. 535-542, May 25 1977.

BRENNAN, J. P.; WAIT, R.; BEGUM, S.; BELL, J. R. *et al.* Detection and mapping of widespread intermolecular protein disulfide formation during cardiac oxidative stress using proteomics with diagonal electrophoresis. **J Biol Chem**, 279, n. 40, p. 41352-41360, Oct 1 2004.

BRODEHL, A.; FERRIER, R. A.; HAMILTON, S. J.; GREENWAY, S. C. *et al.* Mutations in FLNC are Associated with Familial Restrictive Cardiomyopathy. **Human Mutation**, 37, n. 3, p. 269-279, 2016/03/01 2016. https://doi.org/10.1002/humu.22942.

BRODEHL, A.; FERRIER, R. A.; HAMILTON, S. J.; GREENWAY, S. C. *et al.* Mutations inFLNCare Associated with Familial Restrictive Cardiomyopathy. **Human Mutation**, 37, n. 3, p. 269-279, 2016.

BURLEY, S. K.; BERMAN, H. M.; BHIKADIYA, C.; BI, C. *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. **Nucleic Acids Res**, 47, n. D1, p. D464-D474, Jan 8 2019.

BÖNNEMANN, C. G.; BROCKMANN, K.; HANEFELD, F. Muscle ultrasound in Bethlem myopathy. **Neuropediatrics**, 34, n. 6, p. 335-336, Dec 2003.

CHOI, Y.; SIMS, G. E.; MURPHY, S.; MILLER, J. R. *et al.* Predicting the Functional Effect of Amino Acid Substitutions and Indels. **PLoS ONE**, 7, n. 10, p. e46688, 2012.

CHOWDHURY, P. S.; VASMATZIS, G.; BEERS, R.; LEE, B. *et al.* Improved stability and yield of a Fv-toxin fusion protein by computer design and protein engineering of the Fv 1 1Edited by I. Wilson. **Journal of Molecular Biology**, 281, n. 5, p. 917-928, 1998.

CLAEYS, K. G.; VAN DER VEN, P. F.; BEHIN, A.; STOJKOVIC, T. *et al.* Differential involvement of sarcomeric proteins in myofibrillar myopathies: a morphological and immunohistochemical study. **Acta Neuropathol**, 117, n. 3, p. 293-307, Mar 2009.

EDEN, M.; FREY, N. Cardiac Filaminopathies: Illuminating the Divergent Role of Filamin C Mutations in Human Cardiomyopathy. **Journal of Clinical Medicine**, 10, n. 4, 2021.

EHRLICHER, A. J.; NAKAMURA, F.; HARTWIG, J. H.; WEITZ, D. A. *et al.* Mechanical strain in actin networks regulates FilGAP and integrin binding to filamin A. **Nature**, 478, n. 7368, p. 260-263, Sep 18 2011.

EMSLEY, P.; LOHKAMP, B.; SCOTT, W. G.; COWTAN, K. Features and development ofCoot. **Acta Crystallographica Section D Biological Crystallography**, 66, n. 4, p. 486-501, 2010.

EUDES, R.; LE TUAN, K.; DELETTRÉ, J.; MORNON, J.-P. *et al.* A generalized analysis of hydrophobic and loop clusters within globular protein sequences. **BMC structural biology**, 7, p. 2-2, 2007.

FOKKEMA, I. F.; TASCHNER, P. E.; SCHAAFSMA, G. C.; CELLI, J. *et al.* LOVD v.2.0: the next generation in gene variant databases. **Hum Mutat**, 32, n. 5, p. 557-563, May 2011.

FRENCH, S.; WILSON, K. Treatment of Negative Intensity Observations. **Acta Crystallographica Section A**, 34, n. Jul, p. 517-525, 1978.

FRONTERA, W. R.; OCHALA, J. Skeletal muscle: a brief review of structure and function. **Calcif Tissue Int**, 96, n. 3, p. 183-195, Mar 2015.

FUJITA, M.; MITSUHASHI, H.; ISOGAI, S.; NAKATA, T. *et al.* Filamin C plays an essential role in the maintenance of the structural integrity of cardiac and skeletal muscles, revealed by the medaka mutant zacro. **Dev Biol**, 361, n. 1, p. 79-89, Jan 1 2012.

FÜRST, D. O.; GOLDFARB, L. G.; KLEY, R. A.; VORGERD, M. *et al.* Filamin C-related myopathies: pathology and mechanisms. **Acta Neuropathol**, 125, n. 1, p. 33-46, Jan 2013.

GANGULY, H.; MAJUMDER, B.; CHATTOPADHYAY, S.; CHAKRABARTI, P. *et al.* Direct Evidence for CH···π Interaction Mediated Stabilization of Pro- cis Pro Bond in Peptides with Pro-Pro-Aromatic motifs. **Journal of the American Chemical Society**, 134, p. 4661-4669, 03/01 2012.

GAUTEL, M.; DJINOVIĆ-CARUGO, K. The sarcomeric cytoskeleton: from molecules to motion. **J Exp Biol**, 219, n. Pt 2, p. 135-145, Jan 2016.

GÓMEZ, J.; LORCA, R.; REGUERO JULIAN, R.; MORÍS, C. *et al.* Screening of the Filamin C Gene in a Large Cohort of Hypertrophic Cardiomyopathy Patients. **Circulation: Cardiovascular Genetics**, 10, n. 2, p. e001584, 2017/04/01 2017.

HIKIDA, R. S. Aging changes in satellite cells and their functions. **Curr Aging Sci**, 4, n. 3, p. 279-297, Dec 2011.

IWAOKA, M.; KOMATSU, H.; KATSUDA, T.; TOMODA, S. Quantitative evaluation of weak nonbonded Se...F interactions and their remarkable nature as orbital interactions. **J Am Chem Soc**, 124, n. 9, p. 1902-1909, Mar 6 2002.

JAVAN, R.; HORVATH, J. J.; CASE, L. E.; AUSTIN, S. *et al.* Generating color-coded anatomic muscle maps for correlation of quantitative magnetic resonance imaging analysis with clinical examination in neuromuscular disorders. **Muscle Nerve**, 48, n. 2, p. 293-295, Aug 2013.

JOOSTEN, R. P.; SALZEMANN, J.; BLOCH, V.; STOCKINGER, H. *et al.* PDB_REDO: automated re-refinement of X-ray structure models in the PDB. **Journal of Applied Crystallography**, 42, n. 3, p. 376-384, 2009.

KABSCH, W. XDS. **Acta Crystallographica Section D Biological Crystallography**, 66, n. 2, p. 125-132, 2010.

KARCZEWSKI, K. J.; FRANCIOLI, L. C.; TIAO, G.; CUMMINGS, B. B. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. **Nature**, 581, n. 7809, p. 434-443, 2020/05/01 2020.

KARPLUS, P. A.; DIEDERICHS, K. Linking Crystallographic Model and Data Quality. **Science**, 336, n. 6084, p. 1030-1033, 2012.

KOZLOWSKI, L. P.; BUJNICKI, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. **BMC Bioinformatics**, 13, n. 1, p. 111, 2012.

KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. **J Mol Biol**, 157, n. 1, p. 105-132, May 5 1982.

LAD, Y.; KIEMA, T.; JIANG, P.; PENTIKAINEN, O. T. *et al.* Structure of three tandem filamin domains reveals auto-inhibition of ligand binding. **EMBO J**, 26, n. 17, p. 3993-4004, Sep 5 2007.

LEBER, Y.; RUPARELIA, A. A.; KIRFEL, G.; VAN DER VEN, P. F. *et al.* Filamin C is a highly dynamic protein associated with fast repair of myofibrillar microdamage. **Hum Mol Genet**, 25, n. 13, p. 2776-2788, Jul 1 2016.

LEK, M.; KARCZEWSKI, K. J.; MINIKEL, E. V.; SAMOCHA, K. E. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. **Nature**, 536, n. 7616, p. 285-291, 2016/08/01 2016.

MALÝ, M.; DIEDERICHS, K.; DOHNÁLEK, J.; KOLENKO, P. Paired refinement under the control of PAIREF. **IUCrJ**, 7, n. 4, p. 681-692, 2020.

MAO, Z.; NAKAMURA, F. Structure and Function of Filamin C in the Muscle Z-Disc. **Int J Mol Sci**, 21, n. 8, Apr 13 2020.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochim Biophys Acta**, 405, n. 2, p. 442-451, Oct 20 1975.

MCCOY, A. J.; GROSSE-KUNSTLEVE, R. W.; ADAMS, P. D.; WINN, M. D. *et al.* Phasercrystallographic software. **Journal of Applied Crystallography**, 40, n. 4, p. 658-674, 2007.

MLYNEK, G.; LEHNER, A.; NEUHOLD, J.; LEEB, S. *et al.* The Center for Optimized Structural Studies (COSS) platform for automation in cloning, expression, and purification of single proteins and protein-protein complexes. **Amino Acids**, 46, n. 6, p. 1565-1582, Jun 2014.

MOLT, S.; BÜHRDEL, J. B.; YAKOVLEV, S.; SCHEIN, P. *et al.* Aciculin interacts with filamin C and Xin and is essential for myofibril assembly, remodeling and maintenance. **J Cell Sci**, 127, n. Pt 16, p. 3578-3592, Aug 15 2014.

NAKAMURA, F.; OSBORN, T. M.; HARTEMINK, C. A.; HARTWIG, J. H. *et al.* Structural basis of filamin A functions. **J Cell Biol**, 179, n. 5, p. 1011-1025, Dec 3 2007.

NG, P. C.; HENIKOFF, S. SIFT: Predicting amino acid changes that affect protein function. **Nucleic Acids Res**, 31, n. 13, p. 3812-3814, Jul 1 2003.

NILSSON, M. I.; NISSAR, A. A.; AL-SAJEE, D.; TARNOPOLSKY, M. A. *et al.* Xin is a marker of skeletal muscle damage severity in myopathies. **Am J Pathol**, 183, n. 6, p. 1703-1709, Dec 2013.

ORTIZ-GENGA, M. F.; CUENCA, S.; DAL FERRO, M.; ZORIO, E. *et al.* Truncating FLNC Mutations Are Associated With High-Risk Dilated and Arrhythmogenic Cardiomyopathies. **J Am Coll Cardiol**, 68, n. 22, p. 2440-2451, Dec 6 2016.

PIRES, D. E. V.; 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, n. W1, p. W314-W319, 2014.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics**, 30, n. 3, p. 335-342, 2014.

REIMANN, L.; SCHWÄBLE, A. N.; FRICKE, A. L.; MÜHLHÄUSER, W. W. D. *et al.* Phosphoproteomics identifies dual-site phosphorylation in an extended basophilic motif regulating FILIP1-mediated degradation of filamin-C. **Commun Biol**, 3, n. 1, p. 253, May 22 2020.

REIMANN, L.; WIESE, H.; LEBER, Y.; SCHWABLE, A. N. *et al.* Myofibrillar Z-discs Are a Protein Phosphorylation Hot Spot with Protein Kinase C (PKCalpha) Modulating Protein Dynamics. **Mol Cell Proteomics**, 16, n. 3, p. 346-367, Mar 2017.

RICHARDS, S.; AZIZ, N.; BALE, S.; BICK, D. *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. **Genet Med**, 17, n. 5, p. 405-424, May 2015.

ROGNONI, L.; MÖST, T.; ŽOLDÁK, G.; RIEF, M. Force-dependent isomerization kinetics of a highly conserved proline switch modulates the mechanosensing region of filamin. **Proceedings of the National Academy of Sciences**, 111, n. 15, p. 5568, 2014.

ROGNONI, L.; STIGLER, J.; PELZ, B.; YLANNE, J. *et al.* Dynamic force sensing of filamin revealed in single-molecule experiments. **Proc Natl Acad Sci U S A**, 109, n. 48, p. 19679-19684, Nov 27 2012.

ROY, A.; KUCUKURAL, A.; ZHANG, Y. I-TASSER: a unified platform for automated protein structure and function prediction. **Nature Protocols**, 5, n. 4, p. 725-738, 2010.

RUSKAMO, S.; GILBERT, R.; HOFMANN, G.; JIANG, P. *et al.* The C-terminal rod 2 fragment of filamin A forms a compact structure that can be extended. **Biochem J**, 446, n. 2, p. 261-269, Sep 1 2012.

SALI, A.; BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. **J Mol Biol**, 234, n. 3, p. 779-815, Dec 5 1993.

SEPPÄLÄ, J.; BERNARDI, R. C.; HAATAJA, T. J. K.; HELLMAN, M. *et al.* Skeletal Dysplasia Mutations Effect on Human Filamins' Structure and Mechanosensing. **Scientific Reports**, 7, n. 1, 2017.

SETHI, R.; YLÄNNE, J. Small-Angle X-Ray Scattering Reveals Compact Domain-Domain Interactions in the N-Terminal Region of Filamin C. **PLoS ONE**, 9, n. 9, p. e107457, 2014.

SEWRY, C. A. Immunocytochemical analysis of human muscular dystrophy. **Microsc Res Tech**, 48, n. 3-4, p. 142-154, Feb 1-15 2000.

STEWART, D. E.; SARKAR, A.; WAMPLER, J. E. Occurrence and role of cis peptide bonds in protein structures. **J Mol Biol**, 214, n. 1, p. 253-260, Jul 5 1990.

THOMPSON, T. G.; CHAN, Y. M.; HACK, A. A.; BROSIUS, M. *et al.* Filamin 2 (FLN2): A muscle-specific sarcoglycan interacting protein. **J Cell Biol**, 148, n. 1, p. 115-126, Jan 10 2000.

TOSSAVAINEN, H.; KOSKELA, O.; JIANG, P.; YLÄNNE, J. *et al.* Model of a Six Immunoglobulin-Like Domain Fragment of Filamin A (16–21) Built Using Residual Dipolar Couplings. **Journal of the American Chemical Society**, 134, n. 15, p. 6660-6672, 2012.

ULBRICHT, A.; FELIX; VICTOR; PETER *et al.* Cellular Mechanotransduction Relies on Tension-Induced and Chaperone-Assisted Autophagy. **Current Biology**, 23, n. 5, p. 430-435, 2013.

VAGIN, A.; LEBEDEV, A. MoRDa, an automatic molecular replacement pipeline. **Acta Crystallographica Section A Foundations and Advances**, 71, n. a1, p. s19-s19, 2015.

VALDÉS-MAS, R.; GUTIÉRREZ-FERNÁNDEZ, A.; GÓMEZ, J.; COTO, E. *et al.* Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. **Nat Commun**, 5, p. 5326, Oct 29 2014.

VAN DER FLIER, A.; KUIKMAN, I.; KRAMER, D.; GEERTS, D. *et al.* Different splice variants of filamin-B affect myogenesis, subcellular distribution, and determine binding to integrin [beta] subunits. **J Cell Biol**, 156, n. 2, p. 361-376, Jan 21 2002.

VAN DER VEN, P. F.; EHLER, E.; VAKEEL, P.; EULITZ, S. *et al.* Unusual splicing events result in distinct Xin isoforms that associate differentially with filamin c and Mena/VASP. **Exp Cell Res**, 312, n. 11, p. 2154-2167, Jul 1 2006.

VERDONSCHOT, J. A. J.; VANHOUTTE, E. K.; CLAES, G. R. F.; HELDERMAN-VAN DEN ENDEN, A. *et al.* A mutation update for the FLNC gene in myopathies and cardiomyopathies. **Hum Mutat**, 41, n. 6, p. 1091-1111, Jun 2020.

VOYNOV, V.; CHENNAMSETTY, N.; KAYSER, V.; HELK, B. *et al.* Predictive tools for stabilization of therapeutic proteins. **mAbs**, 1, n. 6, p. 580-582, Nov-Dec 2009.

WADMORE, K.; AZAD, A. J.; GEHMLICH, K. The Role of Z-disc Proteins in Myopathy and Cardiomyopathy. **Int J Mol Sci**, 22, n. 6, Mar 17 2021.

WINN, M. D.; BALLARD, C. C.; COWTAN, K. D.; DODSON, E. J. *et al.* Overview of theCCP4 suite and current developments. **Acta Crystallographica Section D Biological Crystallography**, 67, n. 4, p. 235-242, 2011.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition. **Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition**, p. Cp1-621, 2017.

WORTH, C. L.; PREISSNER, R.; BLUNDELL, T. L. SDM—a server for predicting effects of mutations on protein stability and malfunction. **Nucleic Acids Research**, 39, n. suppl_2, p. W215-W222, 2011.

WU, Q.; SCHAPIRA, M.; ARROWSMITH, C. H.; BARSYTE-LOVEJOY, D. Protein arginine methylation: from enigmatic functions to therapeutic targeting. **Nat Rev Drug Discov**, 20, n. 7, p. 509-530, Jul 2021.

YANG, J.; ZHANG, Y. I-TASSER server: new development for protein structure and function predictions. **Nucleic Acids Res**, 43, n. W1, p. W174-181, Jul 1 2015.

ZHOU, Y.; CHEN, Z.; ZHANG, L.; ZHU, M. *et al.* Loss of Filamin C Is Catastrophic for Heart Function. **Circulation**, 141, n. 10, p. 869-871, Mar 10 2020.

ZOU, P.; PINOTSIS, N.; LANGE, S.; SONG, Y.-H. *et al.* Palindromic assembly of the giant muscle protein titin in the sarcomeric Z-disk. **Nature**, 439, n. 7073, p. 229-233, 2006/01/01 2006.