



UNIVERSITAT_{DE}
BARCELONA

Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors

Selen Özkan

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors

Selen Özkan

DOCTORAL THESIS
2023



DOCTORAL THESIS 2023

Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors

Selen Özkan

Facultat de Biologia, Programa de Doctorat en Biomedicina, línia de recerca Bioinformàtica,
Universitat de Barcelona
Clinical and Translational Bioinformatics, Vall d'Hebron Institut de Recerca

Thesis submitted by:

Selen Özkan to qualify for the Doctorate degree by the Universitat de Barcelona



Thesis director:

Prof. Dr. Xavier de la Cruz

Thesis tutor:

Prof. Dr. Josep Lluís Gelpí Buchaca

*“The world was made of mysteries, and I was only another riddle
among the millions.”*

Circe, Madeline Miller

Acknowledgements

I would like to start by thanking my supervisor Prof. Xavier de la Cruz, whom I know since my master's studies. Thank you for giving me the opportunity to do this PhD. Your guidance has led me to be a more ambitious student, and your dedication to science has motivated me to become a better researcher. I cannot think of a better supervisor that I could do this PhD with.

I would like to also thank Prof. Josep Lluís Gelpí Buchaca and Prof. Núria López Bigas, my committee members, for their guidance and support throughout my PhD.

I would like to thank to my dearest friend Natàlia Padilla, who is not only an amazing colleague but also one of the kindest people I met. Thank you for helping me, answering my countless questions, teaching me how to code efficiently, and for your friendship.

I would like to thank to my dearest partner, Derya Buğra Yarkin, for his unconditional support and patience with me throughout these years. Thank you for helping me accomplish my dreams, always believing in me, and knowing how to cheer me up when I need the most.

I am extremely grateful to my sister, Ceren Pajanoja, who is hundreds of kilometers away physically, but has always been right by my side emotionally. Thank you for being my study buddy, programming with me, encouraging me, and understanding me.

I also would like to thank my family, Figen and Sırrı Özkan, for giving me the freedom and opportunity to pursue my dreams. It is all thanks to you

that I am here today, writing this thesis. And my dearest uncles, Derya and Cezmi Dıřpınar, who inspire me to me to leave my comfort zone.

I am so thankful to my amazing friends Derya Yavuz Zor and Dilřad Daętekin, for being exceptionally supportive to me. Thank you for all the laughter and your endless support.

Lastly, I would like to thank my dearest friends: İdil Batuman, for being my survival kit in Barcelona and my motivation source. And my moral support team; Eva Martin del Pico, Mübeccel Soysaraç, Cansu Utku, Merve Ön, Sena Duyar, Doęa Gündem, Melis Aygar, Arda Balkancı, Dani Peyrolon, İpek Aktan and Luz Porras, thank you all for your encouragements and friendship.

Hepinize ayrı ayrı ve hep birlikte, çok teşekkür ederim.

Abstract

Precision medicine aims to provide personalized patient care using genetic data. However, interpreting the phenotypic consequences of genetic variants remains a bottleneck in both clinical and research settings. Computational methods available mostly focus on predicting the binary impact of protein sequence variants (benign/pathogenic). Despite their high predictive power, there still are limitations that affect their reliability. New research efforts are made for approaching the variant prediction problem from a completely different angle: continuous prediction of the molecular impact of variants comparable to the functional assay values.

The work in this thesis belongs to this new paradigm. We start by reviewing the status of the pathogenicity predictors to understand their nature, progress, and limitations. We follow this review by focusing on prediction models for two important tumor suppressor genes involved in hereditary breast and ovarian cancers: BRCA1 and BRCA2. We developed two protein-specific predictors for BRCA1 and BRCA2 proteins trying several regression techniques. Using a small set of features and training sets, we show the potential of regression approaches for addressing the prediction of the quantitative impact of variants.

These studies led to the main goal of this thesis project: the development of a novel approach for predicting the clinical impact of missense variants based on a fundamental formulation of the problem. First, we train a series of 30 protein-specific regression models using the deep mutational scanning experiments available in the literature and a set of sequence- and structure (AlphaFold)-based input features. We show that our models' performances display significant predictive ability.

Second, we extend this protein-specific approach to the clinical genome by developing a general Ensemble Learning based-predictor. For a given variant, this model combines the predictions of a set of representative protein-specific predictors by simply taking their medians. When compared against predictors with similar nature, such as ENVISION, our model showed higher Pearson's correlations for most of the cases in our dataset.

Lastly, we address the final goal of the project: generating binary outcomes from our Ensemble model. The result was subject to a rigorous testing of the resulting methodology in an independent validation dataset against top ranking *in silico* tools in the field. We see that our methodology reaches a predictive level comparable to that of established binary methods, like REVEL or others. In summary, we show that our approach can be useful for predicting the molecular impact of variants and their potential binary status while keeping a high potential for interpretability.

Resumen

La medicina de precisión tiene como objetivo brindar atención personalizada al paciente usando datos genéticos. Sin embargo, interpretar las consecuencias fenotípicas de las variantes genéticas sigue siendo un obstáculo tanto en el ámbito clínico como en el de investigación. Los métodos computacionales actuales se centran principalmente en predecir el impacto binario (benigna/patógena) de las variantes de la secuencia de las proteínas. Pero a pesar de su alto poder predictivo, todavía existen limitaciones que afectan a su fiabilidad. Por ello, se están realizando nuevos esfuerzos de investigación para abordar el problema de la predicción de variantes desde un ángulo diferente: obtener una predicción continua del impacto molecular de variantes comparable a los valores de los ensayos funcionales.

El trabajo de esta tesis pertenece a este nuevo paradigma. Comenzamos revisando el estado de los predictores de patogenicidad para comprender su naturaleza, progreso y limitaciones. Continuamos esta revisión centrándonos en los modelos de predicción de dos importantes genes supresores de tumores implicados en el cáncer de mama y de ovario hereditario: BRCA1 y BRCA2. Finalmente, desarrollamos dos predictores específicos de proteínas para BRCA1 y BRCA2 probando varias técnicas de regresión. Utilizando un pequeño conjunto de características descriptivas y conjuntos de entrenamiento, mostramos el potencial de las técnicas de regresión para abordar la predicción del impacto cuantitativo de las variantes.

Estos estudios me prepararon para abordar el principal objetivo de esta tesis: el desarrollo de una nueva aproximación para predecir el impacto clínico de las variantes sin sentido basado en una formulación fundamental del problema. Primero, entrenamos una serie de 30

modelos de regresión específicos de proteínas utilizando los experimentos de escaneo mutacional profundo disponibles en la literatura y un conjunto de características descriptivas basadas en secuencia y estructura (AlphaFold). A continuación, mostramos que nuestros modelos tienen una mayor y significativa capacidad predictiva.

En segundo lugar, ampliamos esta aproximación específica de proteínas al genoma clínico mediante el desarrollo de un predictor general tipo Ensemble. Para una variante determinada, este modelo combina las predicciones de un conjunto de predictores específicos de proteínas representativos simplemente tomando sus medianas. En comparación con predictores de naturaleza similar, como ENVISION, nuestro modelo mostró mejores correlaciones de Pearson para la mayoría de los casos de nuestro conjunto de datos.

Por último, abordamos el objetivo final del proyecto: generar una predicción binaria a partir de nuestro modelo Ensemble. Evaluamos dichas predicciones con un riguroso test con un conjunto de datos de validación independiente y lo comparamos con las mejores herramientas *in silico* del campo. Los resultados nos muestran que nuestra metodología alcanza un nivel predictivo comparable al de los métodos binarios establecidos, como REVEL u otros. En resumen, mostramos que nuestra aproximación puede ser útil para predecir el impacto molecular de las variantes y su correspondiente estado binario, manteniendo un alto potencial de interpretabilidad.

Table of Contents

1. INTRODUCTION.....	1
1.1. PRECISION MEDICINE	3
1.2. MOLECULAR IMPACT OF PROTEIN SEQUENCE VARIANTS ON PROTEIN FUNCTION AND STRUCTURE	5
1.3. COMPUTATIONAL PREDICTION OF VARIANT EFFECTS	8
1.4. PATHOGENICITY PREDICTION AND INTERPRETATION OF GENETIC VARIANTS FOR THE BINARY OUTCOMES.....	10
1.4.1. <i>Defining the binary variant interpretation problem</i>	12
1.4.2. <i>Building the model for pathogenicity prediction</i>	12
1.4.3. <i>Parameters for evaluating classification performance</i>	14
1.4.4. <i>Limitations and challenges in pathogenicity prediction</i>	17
1.5. QUANTITATIVE PREDICTION OF THE MOLECULAR IMPACT OF GENETIC VARIANTS.....	18
1.5.1. <i>Defining the prediction problem</i>	20
1.5.2. <i>Building the model for continuous prediction</i>	21
1.5.3. <i>Parameters for evaluating the performance</i>	22
1.5.4. <i>Predictors in the field: sequence- vs. theory-driven predictors</i>	23
1.5.5. <i>Limitations and challenges</i>	25
2. OBJECTIVES	27
3. THE COMPUTATIONAL APPROACH TO VARIANT INTERPRETATION: PRINCIPLES, RESULTS, AND APPLICABILITY	31
3.1. INTRODUCTION	33
3.2. MATERIALS AND METHODS	34
3.2.1. <i>The set of protein stability predictors</i>	34
3.2.2. <i>The set of bioinformatics predictors</i>	36
3.2.3. <i>Performance assessment</i>	46
3.3. RESULTS AND DISCUSSION	47
3.3.1. <i>Pathogenicity predictors: Biophysical view</i>	47
3.3.1.1. Protein stability change upon mutation	48
3.3.1.2. Effect of variants on protein interactions	50
3.3.1.3. Applicability of biophysical models	52
3.3.2. <i>Pathogenicity predictors: Bioinformatics tools</i>	54
3.3.2.1. Development of a bioinformatics predictor	55
3.3.2.1.1. Building the variant dataset	55
3.3.2.1.2. Discriminant features.....	57
3.3.2.1.3. Model selection	59
3.3.2.1.4. Validation	59
3.3.2.2. Performance of bioinformatics tools	60
3.3.2.2.1. Variability of performance estimates	65
3.3.3. <i>Future development and challenges</i>	67

3.4. CONCLUSION	69
4. ENDOPHENOTYPE-BASED STRATEGY FOR PATHOGENICITY PREDICTION IN BRCA1 & BRCA2 MISSENSE VARIANTS	71
4.1. INTRODUCTION	73
4.2. MATERIALS AND METHODS	76
4.2.1. <i>Dataset retrieval</i>	76
4.2.2. <i>Discriminant features</i>	78
4.2.3. <i>Regression models</i>	80
4.2.4. <i>Binarization of the regression models</i>	81
4.2.5. <i>Binary classifiers</i>	82
4.2.6. <i>Performance assessment</i>	83
4.3. RESULTS	84
4.3.1. <i>Modelling the impact of variants on endophenotype</i>	84
4.3.1.1. Predicting the outcome of homology-directed repair (HDR) assays using multiple linear regression (MLR) models	86
4.3.1.2. Predicting the outcome of BRCA1/2 variants with other regression models	88
4.3.1.3. Effect of MSA on endophenotype estimates	91
4.3.2. <i>Performance of the regression models for pathogenicity prediction</i>	93
4.3.3. <i>Benchmarking the regression-based pathogenicity predictions against standard in silico predictions</i>	96
4.4. DISCUSSION	100
4.5. CONCLUSION	102
4.6. SUPPORTING INFORMATION	103
5. BINARY PATHOGENICITY CLASSIFICATION OF MISSENSE VARIANTS THROUGH DEVELOPMENT OF QUANTITATIVE PROTEIN-SPECIFIC PREDICTORS	107
5.1. INTRODUCTION	109
5.2. MATERIALS AND METHODS	112
5.2.1. <i>Dataset retrieval</i>	112
5.2.1.1. Collection of variant datasets from DMS experiments	112
5.2.1.2. Collection of clinically labeled variants	115
5.2.2. <i>Discriminant features</i>	116
5.2.2.1. Sequence-based features	118
5.2.2.2. Structure-based features	121
5.2.2.3. Normalization of the features	125
5.2.3. <i>Construction of protein-specific predictors</i>	126
5.2.3.1. Training technique	127
5.2.3.2. Undersampling	127
5.2.3.3. Cross validation	128
5.2.4. <i>Generating PSP predictions for the whole clinical genome: An Ensemble learning-based predictor</i>	129
5.2.5. <i>Obtaining binary (pathogenic/benign) predictions for any protein using the Ensemble-based predictor</i>	130
5.2.6. <i>Benchmarking</i>	132
5.2.7. <i>Performance metrics</i>	134
5.3. RESULTS	135

5.3.1. Feature importance and correlations	135
5.3.2. Performance of the protein-specific predictors.....	137
5.3.3. Ensemble learning model: Extending protein-specific predictions to clinical genome.....	140
5.3.4. Binarization of the Ensemble model	145
5.4. DISCUSSION.....	152
5.5. CONCLUSION	158
5.6. SUPPORTING INFORMATION.....	159
6. DISCUSSION	171
7. CONCLUSION	179
8. BIBLIOGRAPHY	183

1.

Introduction

1.1. Precision Medicine

Precision medicine (PM) is a novel approach in healthcare that has improved with recent technological advancements in medicine and biomedical research (Emmert-Streib and Dehmer, 2018; Maceachern and Forkert, 2021). The goal in PM is to make individually tailored, precise clinical decisions for patients to cover disease prevention, diagnosis, and treatment (Gomez-Lopez *et al.*, 2017; León and Pastor, 2021). It achieves this by integrating multiple sources of evidence from genetic, molecular, and clinical data to electronic health records and lifestyle information (Gomez-Lopez *et al.*, 2017; Carretero-Puche *et al.*, 2020; León and Pastor, 2021). The increasing amount of biological data that can be obtained for patients, as well as improvements in bioinformatics has opened the way for PM to gain more interest in both research and clinical applications (Ho *et al.*, 2019; Maceachern and Forkert, 2021). PM is expected to have growing impact in various areas of healthcare (Denny and Collins, 2021) due to its advantages such as medication effectiveness based on genetic profile, early diagnosis, better disease management, low cost with optimized therapeutic solutions, etc. (Fröhlich *et al.*, 2018)

Apart from using a broad spectrum of experimental techniques, the utilization of Artificial Intelligence (AI) techniques is an integral part of PM, as these computational methodologies are appropriate to handle the large amount of heterogeneous data required to aid in clinical decisions (Carretero-Puche *et al.*, 2020). Several AI-based tools are being studied for their application in various medical settings (Martínez-García and Hernández-Lemus, 2022). For instance, in image-based diagnosis, AI-assisted automated methods have shown high success rates in early identification of incorrectly placed endotracheal tubes (Lakhani, Flanders

and Gorniak, 2021); aided in the detection of pulmonary nodules from tomography scans (Ginneken, 2015); predicting disease progression (Baptista, Ferreira and Rocha, 2021); sensor-enabled tracking systems (wearables, GPS sensors) are utilized to enhance the physico-medical assistance to dementia patients and reducing health risks (Ray, Dash and De, 2019); deep convolutional neural networks are developed to distinguish between malignant and benign skin lesions to increase diagnostic accuracy (Esteva *et al.*, 2017), etc. These examples, and many others (Yu, Beam and Kohane, 2018; Ahmed *et al.*, 2020; Subramanian *et al.*, 2020) show how AI can effectively contribute to the PM approach, (Figure 1.1).

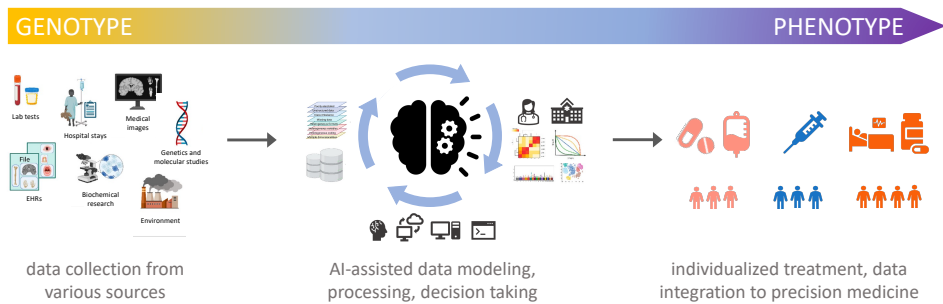


Figure 1.1: Progression from genotype to phenotype in precision medicine through Artificial Intelligence (AI) / Machine Learning (ML) modeling. Figure is adapted from Martínez-García *et al.* (2022).

Conceptually, PM is based on an in-depth exploration of the genotype-to-phenotype relationship, enabling medical decision-making based on the unique characteristics of each patient (Chen and Snyder, 2013; Dai and Shen, 2022). It utilizes genomic information to identify the factors that predispose individuals to specific disease phenotypes (Fernald *et al.*, 2011; Subramanian *et al.*, 2020). In hereditary diseases, the first step in understanding a disease phenotype is the characterization of the impact of genetic variants which can introduce significant alterations in protein

function (Wong *et al.*, 2021). Usually, this is done through a combination of experimental and computational evidence (e.g., functional assays, *in silico* predictions etc.) (Chen, Hu and Shen, 2015; Richards *et al.*, 2015; Vihinen, 2021). However, this annotation problem is complex and remains unsolved (Hoskinson, Dubuc and Mason-Suares, 2017). To aid in the identification of variants' clinical impacts, computational tools (also referred as *in silico* tools, or pathogenicity predictors) have been developed to tackle this problem in a fast and cost-efficient manner (Kumar, Henikoff and Ng, 2009; Adzhubei *et al.*, 2010; Kircher *et al.*, 2014), taking advantage of the growing advances in Machine Learning (ML). In the following sections, we will gradually go through the computational approaches developed to address this problem (known as the *in silico* variant interpretation problem), focusing on the case of missense variants, or variants that introduce an amino acid replacement in the protein sequence.

1.2. Molecular impact of protein sequence variants on protein function and structure

Single nucleotide variants (SNVs) are commonly observed changes in the coding or non-coding region of the genome's DNA (Petrosino *et al.*, 2021). Their phenotypical consequences follow different rules due to the different biological processes underlying them. Missense variants are the result of single base changes in the coding region that result in the substitution of an amino acid in the protein sequence (Ng and Henikoff, 2006), and they constitute 38% of the variants identified in single-gene disorders (Landrum *et al.*, 2016). Pathogenic missense variants can affect a broad variety of biological processes, including mislocalization, disruption of a protein's catalytic activity, aggregation, or degradation etc.

(Vihinen, 2021) (**Figure 1.2**). However, understanding their molecular impact is a difficult problem, as they can produce a wide range of possible effects (Raraigh *et al.*, 2018) and this interpretation relies on factors like specific location within the protein, the nature of the amino acid change, and other related properties of the variant (Arbiza *et al.*, 2006; Kumar *et al.*, 2009; Cooper *et al.*, 2010).

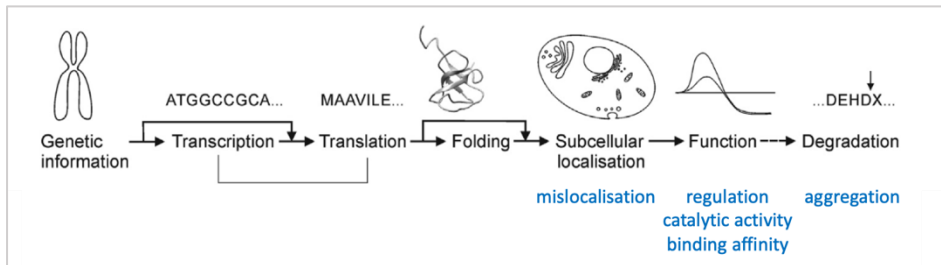


Figure 1.2: Examples of protein functional effects due to missense variants. Example mechanisms that protein function is affected are written in blue. Figure is adapted from Vihinen *et al.* 2021.

If the function of the variant carrier protein is known a priori, *in vitro* mutagenesis or functional assays can be employed to evaluate the effects of the variants (Antonarakis and Cooper, 2018). In some cases, these assays serve as a valuable tool to understand the molecular impact of variants. For example, in the study of Starita *et al.* (2015) they measure the ability of nearly 2000 missense variants in BRCA1 gene to rescue homology-directed repair assay, a mechanism essential for tumor suppression. Generally, these assays may offer insights into protein function, stability, evolutionary trajectories, and effect of the variants on intermediate level phenotypes (Gupta and Varadarajan, 2018). However, the reach of functional assays is limited, for the moment, to a small number of genes and variants (Starita *et al.*, 2017; Raraigh *et al.*, 2018).

Available data indicate that a fundamental approach to the analysis of variants could provide a solution to the variant interpretation problem (Bepler and Berger, 2021). Indeed, structural information can importantly contribute to understand the molecular impact of missense variants. For example, we know that sequence changes can alter protein function through various ways, including induction of abnormal protein-protein interactions (Schuster-Böckler and Bateman, 2008; Navío *et al.*, 2019), disruption of native interactions like hydrogen bonds, salt bridges or disulfide bonds (Steward *et al.*, 2003; Gong and Blundell, 2010), loss of post-translational modification sites (Shuyan *et al.*, 2010), etc. In general, we find that understanding these changes as well as their repercussions on the energetics and dynamics of the protein could then provide a better understanding on how they affect protein's cellular fate (Bhattacharya *et al.*, 2017). Important advances in this direction have taken place in the last years using calculations of physics-based energy functions or molecular dynamics simulations (Jordan, Ramensky and Sunyaev, 2010; Galano-Frutos, Garciá-Cebollada and Sancho, 2019). However, the computational costs involved still limit the use of this fundamental approach as well.

Starting in 2000s (Ng and Henikoff, 2001), a third alternative to the interpretation of missense variants has appeared, providing *in silico* pathogenicity predictors that are based on the combination of different sources of biological information and computational algorithms mostly borrowed from the ML field (Nicora *et al.*, 2022). The following sections are devoted to present the state-of-the art in this research area.

1.3. Computational prediction of variant effects

ML is an algorithmic approach that allows machines to learn from data to obtain statistical models and extract meaningful information without being explicitly programmed (Deo, 2015; Azodi, Tang and Shiu, 2020). It has been utilized in development of computational tools to provide rapid, scalable predictors of the effect of the genetic variations and, more recently, of their molecular impacts on the phenotypes (Goecks *et al.*, 2020; Livesey and Marsh, 2020). Two common approaches that ML encompasses are: (i) supervised learning, which involves labeled data for specific outcomes; and (ii) unsupervised learning, which focuses on identifying patterns or groups in unlabeled data (Wainberg *et al.*, 2018). In this thesis, we primarily concentrate on supervised learning, which is a widely used approach in missense variant interpretation, and also utilized in this thesis.

The workflow for developing a supervised ML model follows four conceptually simple steps (**Figure 1.3**). Each step is explained below briefly for the case of variant pathogenicity prediction.



Figure 1.3: Main steps of creating a supervised ML tool. The boxes describe the process followed in the creation of pathogenicity predictors.

- **Data retrieval:** Gathering relevant data for the pathogenicity prediction problem. Predictors are trained using datasets made of instances that reflect the classification task of interest, e.g. sets of pathogenic and benign variants for variant classification (Iqbal *et al.*, 2020).
- **Feature engineering:** Identifying features discriminating between variant types. During this step, the data are labeled with features that are relevant to the problem of interest. Pathogenicity predictors usually utilize properties such as evolutionary conservation, physicochemical properties, or structural properties of a geometric nature, like solvent accessibility, and others of a biophysical/biochemical nature, like stability, pattern of protein-protein interactions, (Gyulkhandanyan *et al.*, 2020; Yazar and Özbek, 2021), etc.
- **Model building:** Constructing the model using the feature-labeled dataset. During this step, an appropriate mathematical function is chosen and the model is trained using the dataset (Azodi, Tang and Shiu, 2020; Lo Vercio *et al.*, 2020). This choice of the function is determined by our understanding of the problem. In certain cases, such as variant pathogenic prediction, with comprehensive knowledge of the underlying biology, biophysics, and biochemistry of the problem, our decision-making is informed by this expertise. Subsequently, the ML tool facilitates the generation of a numerical model through the training process.
- **Performance estimation:** In this step, the success rate of the trained model is estimated on new instances. This is done by comparing observations and predictions for a set of variants not used in the model's derivation. The performance metrics used depends on the type of problem, e.g., in pathogenicity prediction most used metrics

are sensitivity, specificity, accuracy, precision (Niroula and Vihinen, 2016; Azodi, Tang and Shiu, 2020) etc. Usually, this performance is compared to that of other tools in the field (Yu, Beam and Kohane, 2018).

The previous four-step scheme is divided into two groups, depending on the prediction target (K. Mazandu *et al.*, 2019): (i) classification and (ii) regression. In classification, the prediction target is categorical (Goecks *et al.*, 2020): e.g., clinical outcome of a variant (pathogen or benign), a disease status (disease versus healthy) (Lo Vercio *et al.*, 2020), etc. On the other hand, in regression, the prediction target is a numerical, quantitative variable (Masica and Karchin, 2016; Azodi, Tang and Shiu, 2020): e.g., blood pressure (Hsieh *et al.*, 2016), the strength of response to therapy (Su *et al.*, 2020), medication dosage (Nguyen *et al.*, 2021), the functional assay value (Gray *et al.*, 2018) etc. Both classification and regression models are trained following the same framework described in **Figure 1.3**. The main difference is that the mathematical function used to build the model is different, as well as the performance estimators. In the following sections, we review the two approaches to pathogenicity prediction in more detail: first, binary classifiers and second, regression models.

1.4. Pathogenicity prediction and interpretation of genetic variants for the binary outcomes

Pathogenicity predictors (or classifiers) for the binary (benign/pathogenic) interpretation of missense variants have been rapidly progressing in recent years (Nicora *et al.*, 2022). They are relatively easy to develop, cost-efficient, can handle large amounts of complex data; and have

potential to identify which patients are at risk for certain diseases and may benefit from specific treatments (Vihinen, 2020), as desired in the precision medicine field (Ho *et al.*, 2019).

There are many pathogenicity predictors presently available (Ghosh, Oak and Plon, 2017; Hu *et al.*, 2019; Özkan *et al.*, 2021; Liu *et al.*, 2022). They vary in their methodologies (summarized in Katsonis *et al.*, 2022, Table 1): combination of different features (e.g., sequence conservation (Adzhubei *et al.*, 2010), multiple sequence alignment (Shihab *et al.*, 2013), structure (Yue and Moulton, 2006)); choice of algorithm (e.g., Random Forest (Niroula, Urolagin and Vihinen, 2015), Neural Network (Sundaram *et al.*, 2018), Bayes classifier (Schwarz *et al.*, 2014)), inclusion of multiple tools (also called ensemble predictors, or metapredictors (Ioannidis *et al.*, 2016)) etc.

The application of pathogenicity predictors is valuable in clinical and research settings, especially in cases where genetic or functional evidence is lacking (Pejaver *et al.*, 2022). According to the work of Amendola *et al.* (2016), 55% of the variants studied on a total of nine laboratories used computational tools for the interpretation of the variants, showing how handy these tools can be in the clinical settings. In fact, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines recommend the usage of pathogenicity predictors as supporting evidence for variant interpretation (Richards *et al.*, 2015). More effort has been made in this direction to improve the usage of these tools in the clinical setting (Azodi, Tang and Shiu, 2020; Cubuk *et al.*, 2021; Pejaver *et al.*, 2022).

In summary, the binary prediction of a variant's impact aims to provide a fast, accurate solution to the problem of assigning a clinical outcome to the variants. The accuracy of this approach has yet to improve, but there is a growing effort in this direction. Description of the pathogenicity predictors and their performances are discussed in depth in **The computational approach to the variant interpretation: principles, results, and applicability** chapter of this thesis. Here, we provide an overview on how they define the problem, commonly used algorithms, evaluation parameters and the limitations they face.

1.4.1. Defining the binary variant interpretation problem

According to the ACMG/AMP guidelines, genetic variants for Mendelian disorders are categorized into five classes regarding their clinical impacts: pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, or benign (Richards *et al.*, 2015). In the case of pathogenicity predictors, they usually address the variant interpretation problem as a classification task where the goal is to provide a binary estimate of the impact of variants as pathogenic or benign.

1.4.2. Building the model for pathogenicity prediction

In supervised learning, the pathogenicity prediction problem is implemented through classification models, in which the prediction output is a probability of being pathogenic or benign (Malley, Malley and Pajevic, 2011). This probability is subsequently discretized using a threshold (a value of 0.5 is normally used). The variants are characterized by several discriminant features (Azodi, Tang and Shiu, 2020). During training, the model learns the patterns and relationships between variant features and their corresponding binary labels. Some of the most frequently used classification algorithms are listed below:

- **Logistic Regression**, is a linear model that learns the weights associated with each feature, applies a logistic function to transform the linear combination of features into binary predictions (Malley, Malley and Pajevic, 2011).
- **Random Forest**, is an ensemble model that is based on classification and regression trees (Breiman, 2001). The algorithm grows trees by recursively splitting data, based on different features, to maximize the separation between two labels. In the end, the final decision is based on a majority vote of predictions from all trees.
- **Neural network**, is an algorithm that consists of interconnected layers of nodes, that process and transform the input data (Malley, Malley and Pajevic, 2011; Musolf *et al.*, 2022). The model learns the weights and biases associated with each connection between nodes and passes the data through the network to adjust parameters, and give predictions.
- **Support vector machines (SVM)**, draws a hyperplane between the two classes, aiming to separate them with the largest distance from the hyperplane (in the case of binary prediction) (Malley, Malley and Pajevic, 2011).

Overall, the classification algorithms work through complex technical calculations, except for the Logistic Regression, where the input enters and is transformed into the output without much transparency about how this has happened (Rudin, 2019; Iqbal *et al.*, 2020). To alleviate this problem, most of the algorithms provide feature importance values to clarify which features contribute more to the decisions the machine makes (Eraslan *et al.*, 2019).

1.4.3. Parameters for evaluating classification performance

The performance of pathogenicity predictors can be evaluated using numerous parameters (Vihinen, 2012). Given that the predictions are binary, the raw prediction results in a test sample are often displayed in a confusion matrix (Niroula and Vihinen, 2016) (**Figure 1.4**), which presents the following parameters:

- true positives (TP), the cases when the model correctly predicts that a variant is pathogenic.
- true negatives (TN), the cases when the model correctly predicts that a variant is benign.
- false positives (FP), the cases when the model incorrectly predicts a benign variant as pathogenic.
- false negatives (FN), the cases when the model incorrectly predicts a pathogenic variant as benign.

		PREDICTED	
		Pathogen	Benign
OBSERVED	Pathogen	TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
	Benign	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

Figure 1.4: Confusion matrix used in the evaluation of pathogenicity predictors. This matrix provides four descriptors: true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

The elements of the confusion matrix are then used for calculating more informative parameters. Below we describe the most frequently used performance measures, which are used to evaluate the performance of a model on a dataset of known pathogenic and benign variants.

Sensitivity, also called true positive rate (TPR) or recall, is the ratio of the number of correctly predicted pathogenic variants divided by total number of pathogenic cases in the dataset:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity, also called true negative rate (TNR), is the ratio of the number of correctly predicted benign variants divided by the total number of actual benign cases in the dataset:

$$Specificity = \frac{TN}{TN + FP}$$

Positive prediction value (PPV), also called precision, is the fraction of true pathogenic predictions among all pathogenic predictions made. It is a measure of how a pathogenic prediction is likely to be a true positive. Higher PPV indicates a pathogenic prediction is more likely to be a true positive, and low PPV indicates that a pathogenic prediction is more likely to be a false positive. The formula of PPV is:

$$Positive\ predictive\ value\ (PPV) = \frac{TP}{TP + FP}$$

Negative prediction value (NPV) is the fraction of true benign predictions among all benign predictions made. It is a measure of how a benign prediction result is likely to be a true negative. The formula of NPV is:

$$\text{Negative predictive value (NPV)} = \frac{TN}{TN + FN}$$

Accuracy is the ratio of correct predictions divided by the total number of predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Area Under the receiver operating characteristic Curve (AUC or ROC AUC), is an average measure of the success rate of a predictor. The ROC curve shows the tradeoff between sensitivity and specificity parameters, and AUC is the measure of the area under this curve. An example of ROC curve and AUC is shown in **Figure 1.5**.

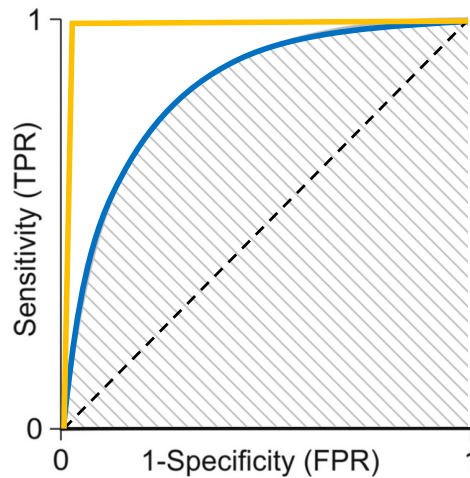


Figure 1.5: Area under the receiver operating characteristic curve (AUC).

The dashed area under the blue curve indicates the AUC for a given model. x and y axes show the false positive rate (FPR, also equal to 1-specificity) and true positive rate (TPR, also equal to sensitivity), respectively. The orange line shows how the ROC curve would look if a perfect model is obtained, whereas the black dashed line corresponds to a random model. Adapted from Niroula and Vihinen, 2016.

The parameters described above range from 0 to 1. In all measures, higher values indicate better performance, whereas the middle value 0.5 indicates a random behavior. These parameters are useful to evaluate the behavior of the model from different points of view. However, they should be carefully examined as they can give overoptimistic evaluations when the distribution of pathogenic and benign variants in the dataset are very different (Vihinen, 2012; Chicco and Jurman, 2020).

Matthews correlation coefficient (MCC), is a balanced measure of correlation between predictions and true cases. It is considered to be a good performance measure (Chicco and Jurman, 2020) because it only gives a higher score when all four descriptors give good results (high TP and TN, low FP and FN) (Chicco and Jurman, 2020). The formula of MCC is:

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

It ranges from -1 to 1; where -1 indicates complete disagreement, 0 indicates random predictions, and 1 indicates complete agreement of predictions with true cases.

1.4.4. Limitations and challenges in pathogenicity prediction

Despite their success rates and improvements, pathogenicity prediction tools have several limitations. One of them is their black-box nature (Rudin, 2019), which prevents our understanding of how a prediction is obtained. In other words, the lack of biophysical/biochemical interpretability of pathogenicity predictors hinders our ability to comprehend why predictions are made for a given variant (Iqbal *et al.*,

2020). Consequently, this situation decreases the trust in pathogenicity predictors among healthcare professionals (Fogel and Kvedar, 2018; Petch, Di and Nelson, 2022). Additionally, related to this issue is the fact that the results are difficult to relate to a mechanistic view of a variant's impact. By predicting variants as pathogenic or benign, these tools skip the complex pathway going from the molecular effect to the clinical phenotype (Gray *et al.*, 2018; Vihinen, 2020). Another important concern is the discrepancy between pathogenicity predictors which results in more difficult decision making, due to conflicting interpretations of variants from different tools (Ghosh, Oak and Plon, 2017; Sun and Yu, 2019). In summary, even with the vast amount of classification tools, we still need prediction tools for improving our understanding of how variants contribute to the molecular mechanisms of disease and lead to the clinical phenotype.

1.5. Quantitative prediction of the molecular impact of genetic variants

Functional assays are informative mutational analyses that allow us to understand the impact of individual variants at the biochemical level in an *in vitro* system (Botstein and Shortle, 1985), to study their effect on structure, general function, or specific interactions (Findlay, 2021). In line with this, the ACMG/AMP guidelines has approved the possibility of using robust and thoroughly verified experimental assays as strong evidence of pathogenicity (Richards *et al.*, 2015). However, given that these laborious experiments are done one-by-one for a limited number of variants, it is not a feasible method to study the effects of the large number of genetic variants that are regularly identified (Gasperini, Starita and Shendure, 2016). In recent years, to overcome this limitation, high-throughput

sequencing techniques like deep mutational scanning (DMS) assays (Fowler and Fields, 2014), have emerged as a powerful option to measure the impact of variants on protein function at a large scale. DMS experiments enable the comprehensive quantification of the functional consequences of all possible amino acid changes in a single experiment (**Figure 1.6**), thus providing significant insights on protein function.

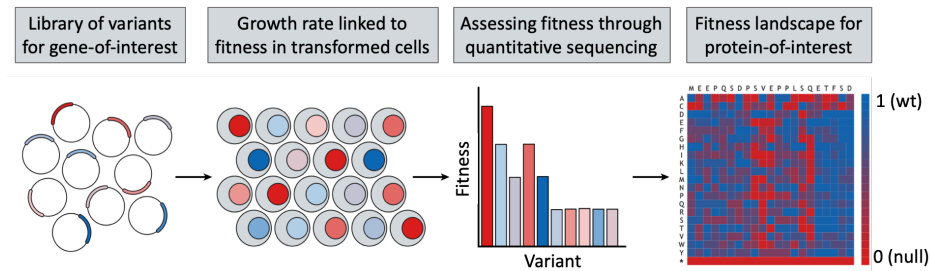


Figure 1.6: Workflow of deep mutational scanning assay experiment. A DMS experiment consists of four main steps: (i) generating library of variants cloning them into expression vectors; (ii) growth of vectors in host cells; (iii) measuring the variant fitness with quantitative sequencing; and (iv) calculation of the fitness values. wt: wild-type. Figure adapted from Livesey *et al.* (2022).

DMS assays are starting to gain importance in clinical applications for the interpretation of sequence variants (Gasperini, Starita and Shendure, 2016; Gelman *et al.*, 2019; Stein *et al.*, 2019). These assays can characterize variants more in accordance with the growing view of disease as a continuum that varies between patients (Sarkar, Yang and Vihinen, 2020); which goes beyond binary characterization of variants.

Yet, the analysis and interpretation of the outcomes of these large-scale assays is not always simple: it is influenced by factors such as the quality of the experimental design, selection criteria, and data processing methodologies (Faure *et al.*, 2020; Mehlhoff and Ostermeier, 2020). Additionally, if we are thinking in generalized clinical applications, we

cannot still comprehensively analyze all possible variants of all the proteins of the clinical genome using DMS. In these circumstances, ML methods have recently appeared as a promising option in the quantitative characterization of protein variants (Gray *et al.*, 2018). This new generation of tools leverage information from existing DMS experiments and combine it with different variant descriptors, in a way analogous to binary classifiers, to build predictive models of the molecular impact of variants. Due to their relatively low computational cost, these tools can provide impact predictions for massive numbers of variants in relatively little time. Continuous predictions can, a priori, enable more accurate assessments of pathogenicity and a deeper comprehension of the relationship between genotype and phenotype (Masica and Karchin, 2016). Furthermore, these estimates can always be binarized, when required, using cut-off values. This allows for the easy integration of the assay results into existing clinical decision frameworks.

In summary, the quantitative prediction of the molecular impact of missense variants is becoming a valuable and feasible goal nowadays, with important consequences for the clinical interpretation of variants. In this section, we explore the application of regression models to this problem, describing their performance and challenges.

1.5.1. Defining the prediction problem

The goal of quantitative (continuous) prediction is to provide a numerical (or non-categorical) estimate of the impact of variants on protein function. In other words, the goal is to obtain a continuous score estimate, which can vary within a certain range that can be normalized (e.g., restricting its variation between 0 and 1) to simplify the comparison or integration of results between different proteins.

1.5.2. Building the model for continuous prediction

When the dependent variable in a prediction problem is continuous, regression models are commonly employed (Lo Vercio et al., 2020). These models work by finding a line, or a plane (in multidimensional regressions), that minimizes the differences between the predicted and observed values of the dependent variable. During this minimization process, the weights of the different features in the model are estimated.

Regression models range from simplistic to more complex (Hastie *et al.*, 2009; Malley, Malley and Pajevic, 2011). Simplistic algorithms like **Linear regression** and **multiple linear regression** establish a linear relationship between the dependent variable (target) and the input features (independent variables) (Hastie *et al.*, 2009):

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \epsilon$$

where y is the dependent variable, β_0 is the intercept (slope of the line), β_i is the coefficient (or weight) for the i -th feature, n is the total number of independent variables, the x_i are the independent variables, and ϵ is the estimated error.

Complex regression models (Hastie *et al.*, 2009), such as Ridge, Lasso, Elastic, etc., introduce regularization techniques by adding penalty terms to control for overfitting, balancing model complexity, thus improving the robustness of the model. The choice of which regression approach works best usually depends on several factors in the dataset, such as the complexity of the target data and the number of features, etc.

In the context of variant impact prediction, even the simplistic regression models may provide interpretable predictions, where the contribution of

the different features or the relationship between them and the target is easier to understand. Therefore, these models may alleviate the black box problem observed in binary classifiers (Sruthi and Prakash, 2021).

1.5.3. Parameters for evaluating the performance

Two measures commonly used to evaluate the performance of regression models are the Pearson correlation coefficient and the Spearman rank correlation coefficient (Winter, Gosling and Potter, 2016), which can be described as follows:

The **Pearson correlation coefficient** is a measure of the linear relationship between observations and predictions. It ranges from -1 to 1: a value of -1 indicates an exact, decreasing, linear relationship; 0 indicates no correlation; and 1 indicates an exact, increasing, linear relationship.

The **Spearman rank correlation coefficient** is a nonparametric measure of the monotonic relationship between observed and predicted values. It is based on the rank order of the observations, rather than on their actual numerical values. It also varies between -1 and 1, with related interpretations to the Pearson correlation coefficient: -1 indicates that predicted ranks for the sample elements is opposite to that obtained with the observed values; 1 indicates that both rankings are same.

The two metrics have their own specific use cases and can be helpful for different types of machine learning problems (Winter, Gosling and Potter, 2016). We may use the Pearson correlation coefficient to evaluate how well a linear regression model fits the data, and Spearman rank correlation coefficient to evaluate the ranking power of the predictions.

1.5.4. Predictors in the field: sequence- vs. theory-driven predictors

There has been a growing interest in computational tools for quantitative variant effect prediction in recent years (Gray *et al.*, 2018; Riesselman, Ingraham and Marks, 2018; Luo *et al.*, 2021). Broadly, these tools can be categorized into two groups: sequence-based and theory-based.

Sequence-based predictors use only protein sequence information to derive predictive features. They utilize evolutionary information from related sequences to extract sequence-specific properties such as the composition, conservation, constraints, etc., to predict damaging or tolerable variant impacts. As they are (mostly) unsupervised, they do not rely on experimental or clinically labeled data and are applicable to a wide range of proteins and variants (Xu *et al.*, 2019). Several tools have been developed with this philosophy to tackle the sequence-to-function problem (Hopf *et al.*, 2017; Riesselman, Ingraham and Marks, 2018; Frazer *et al.*, 2021; Luo *et al.*, 2021). A recent example of a successful sequence-based prediction tool is EVE (Frazer *et al.*, 2021), a variant effect predictor using only the distribution of sequences across species. The model learns from the amino acid distributions, and the relationship between positions in a clinical label-free manner and predicts both a continuous score and binary class for the variants. Their predictor outperforms several pathogenicity predictors and has large Spearman correlations with high-throughput experimental data (Frazer *et al.*, 2021). Another example is ECNet, a predictor that also uses a “biologically motivated sequence modeling” approach to predict functional scores for the variants (Luo *et al.*, 2021). Their model incorporates both local and global evolutionary representations to capture the co-evolutionary relationship of the residues as well as structure and stability-related

properties. When they benchmarked their tool using DMS data, their model outperforms several ML tools (Luo *et al.*, 2021). Overall, although sequence-based tools show good results in cross-validated performances or in benchmarking studies with DMS data, the main concern is that as they are very difficult to interpret; they fail at capturing the complex functional relationships that cannot be covered from the sequence information alone; and it is unclear how some of them are going to be extended to other proteins beyond the data training set.

Theory-driven predictors, on the other hand, incorporate biophysical parameters, involve structure-to-function relationships, and domain-specific knowledge to derive predictive features and guide their predictors (Alber *et al.*, 2019; Stein *et al.*, 2019). Such features include evolutionary constraints, domain interactions, functional sites, interatomic distances, hydrophobicity, etc.; therefore, the resulting models are more interpretable. A unique example of a theory-driven predictor with regression approach is ENVISION (Gray *et al.*, 2018), which is the first quantitative predictor of variants' molecular impact that uses DMS data for training their models. This allows the model to learn the protein function information for large variant datasets and can generalize to variants from proteins that it had never seen before, with moderate performance.

In the end, both sequence-based and theory-driven models have their own strengths and weaknesses. The choice of the model depends on the requirements of the prediction problem, data availability and the desired interpretability from the predictor.

1.5.5. Limitations and challenges

DMS experiments provide comprehensive information about the functional consequences of variants, therefore this rich information can be used for training ML models. Although the predictors we covered in this section show promising performances, this is not the case for all proteins. This can be seen more clearly in observed vs. predicted plots, as shown in **Figure 1.7** for ENVISION and DeepSequence.

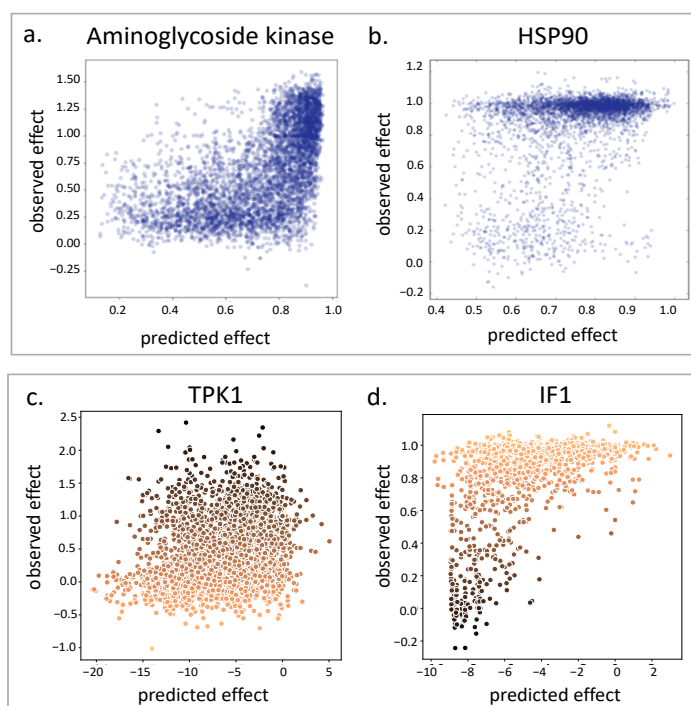


Figure 1.7: The performance of ENVISION and DeepSequence for the prediction of DMS assay values. Scatterplots showing the observed (y-axis) vs predicted (x-axis) values: performance of ENVISION is shown for (a) aminoglycoside kinase, (b) HSP90 proteins; performance of DeepSequence shown for (c) TPK1 and (d) IF1 proteins. Figures for ENVISION and DeepSequence are adapted from Gray et al. 2018 and Riesselman et al. 2018, respectively.

The examples for the poor performances indicate that these models may require additional descriptors of the impact of variants. Additionally, all predictors we mentioned in this section, except for EVE, have not been tested for the binary pathogenicity prediction, despite its immediate interest. This further shows the gap in this field when transforming quantitative predictions into binary classifications.

By leveraging the strengths of DMS experiments and addressing the limitations of the existing models, our goal in this study is to develop our own regression models that can advance our understanding of the impact of variants on protein function, and to transform our results into binary predictions (pathogenic/benign) to link the two approaches to variant interpretation.

2.

Objectives

The main objective of this thesis is to develop a novel pathogenicity predictor for the binary prediction of missense variants, grounded on the quantitative prediction of their molecular impacts.

To accomplish this goal, the thesis addresses the following objectives:

1. Understanding the current status of the pathogenicity predictors in the field: their progress, improvements and weaknesses
2. Testing the value of protein-specific regression models in a reduced version of the problem: the case of BRCA1 and BRCA2 proteins
3. Development of an ensemble-based predictor for the binary prediction of missense variants, through combining quantitative predictions from a set of representative protein-specific predictors.

This objective is divided into three parts:

- 3.1. Developing protein-specific predictors for the quantitative prediction of the molecular impact of missense variants, by taking advantage of the deep mutational scanning assays available in the literature
- 3.2. Generalizing the previous protein-specific predictors to the clinical genome with an ensemble-based predictor
- 3.3. Obtaining an ensemble-based binary (pathogenic/benign) classifier that is generalizable for missense variants in any protein.

3.

**The computational approach to variant
interpretation: principles, results, and
applicability**

The results presented in this part of the thesis are published as a review chapter in “Özkan et al. (2021). The computational approach to variant interpretation: principles, results, and applicability. In *Clinical DNA Variant Interpretation* (pp. 89-119). Academic Press” in Elsevier’s Clinical DNA Variant Interpretation book (Patrinos, 2021).

3.1. Introduction

The work presented in this chapter is focused on giving a unified view of the pathogenicity prediction process, starting from its fundamental (biophysical) aspects until reaching the most clinical approach. Our focus centers on computational tools designed to predict the clinical consequences of single amino acid variants and to evaluate their progress. More specifically, we concentrate on tools that treat this problem as a binary classification in which the impact of missense variants is considered as either benign or pathogenic.

The chapter is divided into two main blocks. In the first block, we give an overview of the biophysical approach to the pathogenicity prediction problem and the associated computational tools. These tools are fully interpretable because they are based on a theoretical comprehension of how two fundamental properties (protein stability and binding ability) underlying protein function change upon mutation. Next, we turn our attention to bioinformatics predictors (or pathogenicity predictors), the computational tools that are more commonly utilized in the clinical settings. These predictors follow a pragmatic philosophy, directly modeling the relationship between an arbitrary number of variant descriptors and a binary view of the variant's clinical impact: pathogenic or benign. No conceptual framework is favored; any feature with predictive power can be included in the final model: performance is their main goal.

This study required a review of the field's literature accompanied by a process of data retrieval and representations to support our view. It covers essentially all the material presented in our book chapter (Özkan

et al., 2021); however, the bibliography on some discussion points has been updated for this thesis.

3.2. Materials and Methods

Below, we describe our methodology for collecting information on the existing missense variant pathogenicity predictors including the literature search and filtering, and the grouping of tools based on technical similarities. For all the predictors described, our literature search focused only on reports that provided performance results with independent benchmarking studies.

The material used in the first block of the chapter (the protein stability predictors) is described in detail in **section 3.2.1**. The material used in the second block (bioinformatics predictors) is described in detail in **section 3.2.2**. Lastly, all the performance evaluation metrics mentioned throughout this chapter are briefly described with their formulas in **section 3.2.3**.

3.2.1. The set of protein stability predictors

The collection of stability predictors and their performances (expressed using Observed vs. Predicted correlations) are shown in **Table 3.1**.

Table 3.1 Protein stability predictors. Protein stability predictors used in Figure 3.1. Table is adapted from Özkan et al. 2021.

Predictor (stability)	Correlation	Stability / protein interaction	Reference (DOI)
mCSM	0.74	$\Delta\Delta G$	10.1093/bioinformatics/btt691
DUET	0.71	$\Delta\Delta G$	10.1093/nar/gku411
DynaMut	0.70	$\Delta\Delta G$	10.1093/nar/gky300
MAESTRO	0.69	$\Delta\Delta G$	10.1186/s12859-015-0548-6
INPS	0.68	$\Delta\Delta G$	10.1093/bioinformatics/btv291
FEP+	0.65	$\Delta\Delta G$	10.1016/j.jmb.2016.12.007
SAAMBE	0.62	$\Delta\Delta G$	10.1371/journal.pcbi.1004276
SDM2	0.48	$\Delta\Delta G$	10.1093/nar/gkx439
FEP	0.78	$\Delta\Delta G_{bind}$ –PPI	10.1016/j.jmb.2019.02.003
iSEE	0.73	$\Delta\Delta G_{bind}$ –PPI	10.1002/prot.25630
SAMPDI	0.72	$\Delta\Delta G_{bind}$ –PDI	10.1093/bioinformatics/btx698
MM-PBSA	0.69	$\Delta\Delta G_{bind}$ –PPI	10.1021/ct401022c
mCSM-NA	0.68	$\Delta\Delta G_{bind}$ –PRI	10.1093/nar/gkx236
Flex ddG	0.63	$\Delta\Delta G_{bind}$ –PPI	10.1021/acs.jpcb.7b11367
BindProfX	0.63	$\Delta\Delta G_{bind}$ –PPI	10.1016/j.jmb.2016.11.022
mCSM	0.56	$\Delta\Delta G_{bind}$ –PPI	10.1093/bioinformatics/btt691
MutaBind	0.56	$\Delta\Delta G_{bind}$ –PPI	10.1093/nar/gkw374
mCSM-AB	0.45	$\Delta\Delta G_{bind}$ –PPI	10.1093/nar/gkw458
EasyE	0.41	$\Delta\Delta G_{bind}$ –PPI	10.1093/bioinformatics/bty092
ProAffiMuSeq	0.27	$\Delta\Delta G_{bind}$ –PPI	10.1093/bioinformatics/btz829

3.2.2. The set of bioinformatics predictors

The predictors used in this study were collected through a literature search with the following criteria: publication date from 2014 to 2019; description of the training set, availability of performance estimations obtained in validation test sets that are independent from the training dataset. Below, we list the bioinformatics predictors retrieved which are organized in separate tables according to the analyses we presented. The list of all predictors collected is given in **Table 3.2**.

Table 3.2 Bioinformatics predictors retrieved for the 2014 – 2019 period.
Predictors and parameters used in Figures 3.3., 3.5, 3.6, 3.7, 3.8. MCC: Matthew’s correlation coefficient, AUC: Area under the ROC curve. This table is adapted from Özkan et al. 2021.

YEAR	SOURCE	PREDICTOR	TYPE	ALGORITHM	TRAINING SET	MCC	AUC	ACCURACY	SENSITIVITY	SPECIFICITY
2014	Kircher et al. 2014	CADD	Metapredictor	support vector machine	Simulated, Swissvar, HumVar	-	0.918	-	-	-
2014	Bendi et al. 2014	PredictSNP	Metapredictor	majority vote	SwissProt, HGMD (P), dbSNP (N), Phenocode(swissprot+db) (P), others (P)	0.494	0.808	0.747	0.741	0.753
2014	Rapakoulia et al. 2014	EnsembleGASVR	Metapredictor	adaptive genetic algorithm and nu-support vector regression	SwissProt (P), dbSNP (N)	0.642	0.902	0.818	0.762	0.877
2014	Wong et al. 2014	SNPdyad	Fundamental	random forest	HumDiv, HumVar	-	0.7	-	-	-
2014	Schwarz et al. 2014	MutationTaster2	Fundamental	Bayes classifier	ClinVar (P), HGMD (P), 1000G(N)	0.8	-	0.9	0.93	0.87
2014	Yates et al. 2014	SuSpect	Fundamental	support vector machine	HumSavar	0.65	0.9	-	-	-
2015	Leong et al. 2015	KCNH2-specific	Metapredictor	Kruskal-Wallis test	no training data	0.62	-	0.911	0.915	0.875
2015	Leong et al. 2015	KCNQ1-specific	Metapredictor	Kruskal-Wallis test	no training data	0.7	-	0.927	0.931	1.0
2015	Dong et al. 2015	MetaSVM	Metapredictor	support vector machine	UniProt	0.7	0.93	0.85	0.87	0.85
2015	Dong et al. 2015	MetaLR	Metapredictor	logistic regression	UniProt	0.71	0.94	0.86	0.89	0.85
2015	Quang et al. 2015	DANN	Metapredictor	deep neural network	Simulated	-	0.722	0.661	-	-
2015	Niroula et al. 2015	PON-MMR2	Fundamental	random forest	VarBench	0.69	-	0.866	0.88	0.85

continued

YEAR	SOURCE	PREDICTOR	TYPE	ALGORITHM	TRAINING SET	MCC	AUC	ACCURACY	SENSITIVITY	SPECIFICITY
2015	Hetch et al. 2015	SNAP2	Fundamental	neural network	PMD, Literature, SwissProt (N), OMIM (P), HumVar (P)	0.24	-	0.688	-	-
2015	Niroula et al. 2015	PON-P2	Fundamental	random forest	VarBench	0.71	-	0.86	0.85	0.86
2015	Vaser et al. 2015	SIFT 4G	Fundamental	alignment score	No training data	0.53	-	0.774	0.803	0.732
2016	Wang et al. 2016	iFish	Metapredictor	Support vector machine	HGMD (P), 1000G (N) and dbSNP (N)	0.54	0.85	0.769	0.81	0.73
2016	Bendi et al. 2016	PredictSNP2	Metapredictor	majority vote	ClinVar (P), VarSNP (N)	0.533	0.803	0.761	0.863	0.660
2016	Alvarez et al. 2016	PRDIS	Metapredictor	neural network	UniProt (P), Evolutionary model (N)	0.739	-	0.873	0.933	0.846
2016	Knecht et al. 2016	IMHOTEP	Metapredictor	random forest	HGMD (P), UCSC(N)	0.93	0.99	0.97	0.97	0.96
2016	Laza et al. 2016	Eigen	Metapredictor	unsupervised	dbSNP	-	0.864	-	-	-
2016	Ioannidis et al. 2016	REVEL	Metapredictor	random forest	HGMD (P), ESP (N), literature (N), 1000G (N)	-	0.957	-	-	-
2016	Jagadeesh et al. 2016	M-CAP	Metapredictor	gradient tree boosting	HGMD (P), Exac (N)	-	0.9	-	-	-
2016	Adebali et al. 2016	SAVER	Fundamental	phylogenetic and protein sequence analyses	Literature search	0.59	-	0.91	0.95	0.67

continued

YEAR	SOURCE	PREDICTOR	TYPE	ALGORITHM	TRAINING SET	MCC	AUC	ACCURACY	SENSITIVITY	SPECIFICITY
2016	Riera et al. 2016	PSP	Fundamental	neural network	UniProt (P), Evolutionary model (N)	0.66	-	0.853	0.854	0.847
2016	Baugh et al. 2016	VIPUR	Fundamental	logistic regression model	HumDiv, UniProt	0.432	-	0.748	0.745	0.759
2016	Thomas et al. 2016	PANTHER-PSEP	Fundamental	phylogenetic tree	no training data	0.38	0.762	0.69	0.67	0.71
2016	Salgado et al. 2016	UMD-Predictor	Fundamental		no training data	0.7	0.91	0.85	0.88	0.84
2017	Velde et al. 2017	GAVIN	Metapredictor	gene calibration	ClinVar (P), ExAc (N)	-	-	-	0.91	0.77
2017	Feng et al. 2017	BayesDel	Metapredictor	Bayesian framework	ClinVar (P), UniProt (P), dbSNP(N), 1000G(N), ExAc(N), others (N)	-	0.8	-	-	-
2017	Raimondi et al. 2017	DEOGEN2	Metapredictor	random forest	HumSavar	0.73	-	0.87	0.87	0.86
2017	Capriotti et al. 2017	PHD-SNPg	Metapredictor	Gradient tree boosting	ClinVar	0.65	0.91	0.85	0.94	0.87
2017	Hopf et al. 2017	Evmutation	Fundamental	unsupervised	literature search	-	0.89	-	-	-
2017	Rogers et al. 2018	FATHMM-XF	Fundamental	Gradient tree boosting	HGMD (P), 1000G (N)	0.76	0.96	0.88	0.84	0.92
2017	Pejaver et al. 2017	MutPred2	Fundamental	neural network	HGMD (P), UCSC (N), SwissProt, dbSNP	0.379	0.848	0.59	0.44	0.95

continued

3. The computational approach to variant interpretation: principles, results, and applicability

YEAR	SOURCE	PREDICTOR	TYPE	ALGORITHM	TRAINING SET	MCC	AUC	ACCURACY	SENSITIVITY	SPECIFICITY
2017	Pan et al. 2017	PredSav	Fundamental	gradient tree boosting	UniProt (P), Ensemble human variation db (N)	0.651	0.9	0.826	0.814	0.838
2017	Lopez-Fernando et al. 2017	PMut	Fundamental	random forest	SwissVar	0.42	0.71	0.71	0.76	0.65
2018	Zhou et al. 2018	ADME optimized model	Metapredictor	Youden index	Literature search	0.85	-	0.92	0.926	0.929
2018	Alirezale et al. 2018	ClinPred	Metapredictor	random forest and decision tree	ClinVar	0.85	0.98	0.94	0.93	0.94
2018	Shrestha et al. 2018	DPYD-Verifier	Fundamental	random forest	Literature search	0.67	-	0.84	-	-
2018	Ponzoni et al. 2018	RAPSODY	Fundamental	random forest	HunVar, ExoVar, VariBench, PredictSNP (SwissProt, HGMD (P), dbSNP (N), Phenocode (P), others (P)), SwissVar	0.42	0.794	0.744	0.766	0.686
2018	Gray et al. 2018	Envision	Fundamental	decision tree	Mutagenesis data from literature	-	0.72	-	-	-
2019	Hart et al. 2019	BRCA1-NVM	Metapredictor	Naïve Voting Method	Literature search	0.7	0.93	0.881	0.83	0.898
2019	Hart et al. 2019	BRCA2-NVM	Metapredictor	Naïve Voting Method	Functional assay	0.659	0.9	0.844	0.788	0.874
2019	Padilla et al. 2019	MLR-BRCA1	Fundamental	multiple linear regression	Literature search	0.502	-	0.75	0.615	0.867
2019	Padilla et al. 2019	MLR-BRCA2	Fundamental	multiple linear regression	Literature search	0.714	-	0.857	0.862	0.852

The metapredictors retrieved, together with the combination of pre-existing predictors used to build them is listed in Table 3.3.

Table 3.3 Metapredictors. Predictors used in Figure 3.4. Table is adapted from Özkan et al. 2021.

Metapredictor	Preexisting predictors	Reference (DOI)
iFish	SIFT, PolyPhen2, MutationAssessor, GERP, PhyloP	10.1038/srep31321
ADME optimized model	LRT, MutationAssessor, PROVEAN, VEST3, CADD	10.1038/s41397-018-0044-2
BRCA1-NVM	Vest3, Perch, AlignGVGD, PolyPhen2, MetaSVM, SIFT, GERP, LRT, Provean, MutationAssessor, MetaLR, PhastCons	10.1038/s41436-018-0018-4
BRCA2-NVM	Perch, MetaSVM, MetaLR, PolyPhen2, MutationAssessor, Vest3	10.1038/s41436-018-0018-4
KCNH2 - specific	SIFT, PROVEAN, SNPs&GO	10.1186/s12881-015-0176-z
KCNQ1 - specific	SIFT, PROVEAN, SNPs&GO	10.1186/s12881-015-0176-z
GAVIN	CADD	10.1186/s13059-016-1141-7
BayesDel	PolyPhen2, SIFT, FATHMM, LRT, MutationTaster, MutationAssessor, PhyloP, GERP++, SiPhy	10.1002/humu.23158
CADD	SIFT, PolyPhen, GERP, phastCons, phyloP, Grantham	10.1038/ng.2892
PredictSNP2	CADD, DANN, FATHMM, FunSeq2, GWAVA	10.1371/journal.pcbi.1004962
PredictSNP	PANTHER, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP	10.1371/journal.pcbi.1003440
MetaSVM	SIFT, PolyPhen2, GERP++, MutationTaster, MutationAssessor, FATHMM, LRT, SiPhy, PhyloP	10.1093/hmg/ddu733
MetaLR	SIFT, PolyPhen2, GERP++, MutationTaster, MutationAssessor, FATHMM, LRT, SiPhy, PhyloP	10.1093/hmg/ddu733
EnsembleGASVR	PANTHER	10.1093/bioinformatics/btu297
PRDIS	combinations of SIFT, PolyPhen2, CADD, PON-P2, MutationTaster2	10.1186/s12864-017-3914-0
IMHOTEP	SIFT, PolyPhen2, SNPs&GO, MutPred, MutationTaster2, FATHMM, MutationAssessor, Grantham, PhyloP	10.1093/nar/gkw886
ClinPred	SIFT, PolyPhen2, LRT, MutationAssessor, PROVEAN, CADD, GERP, DANN, PhasCons, fitCons, PhyloP, SiPhy	10.1016/j.ajhg.2018.08.005
DEOGEN2	PROVEAN	10.1093/nar/gkx390
Eigen	SIFT, PolyPhen2, MutationAssessor	10.1038/ng.3477
REVEL	MutPred, FATHMM, VEST3, PolyPhen2, SIFT, PROVEAN, LRT, MutationAssessor, MutationTaster, GERP, SiPhy, phyloP, phastCons	10.1016/j.ajhg.2016.08.016
M-CAP	SIFT, PolyPhen2, CADD, MetaLR, MetaSVM, FATHMM, PhastCons, MutationTaster, MutationAssessor, PhyloP, LRT, GERP, SiPHY	10.1038/ng.3703
DANN	CADD	10.1093/bioinformatics/btu703
PhD-SNPg	PhyLoP	10.1093/nar/gkx369

The references (DOI) of the pre-existing predictors cited in Table 3.3 are listed in Table 3.4.

Table 3.4 List of pre-existing predictors that are used in metapredictors given in Table 3.3. Table is adapted from Özkan et al. 2021.

Pre-existing predictors	Reference (DOI)	Pre-existing predictors	Reference (DOI)
SIFT	10.1038/nprot.2009.86	SNPs&GO	10.1002/humu.21047
PolyPhen2	10.1038/nmeth0410-248	PERCH	10.1002/humu.23158
MutationAssessor	10.1093/nar/gkr407	PolyPhen1	10.1093/nar/gkf493
PhyloP	10.1007/11732990_17	Grantham	10.1126/science.185.4154.862
GERP++	10.1371/journal.pcbi.1001025	DANN	10.1093/bioinformatics/btu703
LRT	10.1101/gr.092619.109	PANTHER	10.1101/gr.772403
PROVEAN	10.1371/journal.pone.0046688	MutationTaster2	10.1038/nmeth.2890
CADD	10.1038/ng.2892	MutPred	10.1093/bioinformatics/btp528
FATHMM	10.1002/humu.22225	AlignGVGD	10.1136/jmg.2005.033878
SiPhy	10.1093/bioinformatics/btp190	FunSeq2	10.1186/s13059-014-0480-5
phastCons	10.1101/gr.3715005	GWAVA	10.1038/nmeth.2832
MutationTaster	10.1038/nmeth0810-575	PhD-SNP	10.1093/bioinformatics/btl423
VEST3	10.1186/1471-2164-14-S3-S3	SNAP	10.1093/nar/gkm238
MetaSVM	10.1093/hmg/ddu733	PON-P2	10.1371/journal.pone.0117380
MetaLR	10.1093/hmg/ddu733	fitCons	10.1038/ng.3196

Lastly, we did a literature search to collect independent benchmarking studies of the retrieved bioinformatics predictors. We limited our search to those studies that used Matthew's correlation coefficient (MCC)

parameter as one of their performance estimators. The list of the benchmarking studies collected is shown in Table 3.5.

Table 3.5 Benchmarking studies collected from literature. List of studies used in Figure 3.9. Table is adapted from Özkan et al., 2021.

Source	Predictors used for benchmarking
Alirezai et al. 2018	ClinPred, REVEL, M-CAP, MetaLR, FATHMM-MKL
Alvarez et al. 2016	SIFT, PolyPhen2, MutationTaster2, CADD, PON-P2
Baugh et al. 2016	SIFT, PolyPhen2, CADD, PROVEAN, VIPUR
Bendl et al. 2014	CONDEL, Meta-SNP, PredictSNP, PON-P
Bendl et al. 2016	SIFT, PolyPhen2, PolyPhen1, MAPP, nsSNPAnalyzer, PANTHER, SNAP, PhD-SNP, PreditSNP
Dong et al. 2015	SIFT, PolyPhen2, MetaSVM, MetaLR, LRT, MutationTaster, MutationAssessor, FATHMM, GERP++ RS, PhyloP, siPhy, KGSSeq, CONDEL, CADD phred
Grimm et al. 2015	SIFT, PolyPhen2, MutationTaster2, MutationAssessor, LRT, FATHMM-U, FATHMM-W
Hetch et al. 2015	SIFT, PolyPhen2, SNAP, SNAP2
Knetch et al. 2016	IMHOTEP, Condel, CoVEC, CAROL, CADD, MetaSVM, MetaLR, MutatonTaster2
Konig et al. 2016	SIFT, PON-P2, PROVEAN
Li et al. 2018	SIFT, PolyPhen2, FATHMM, fitCons, LRT, MutationAssessor, MutationTaster, PROVEAN, VEST3, GERP++, phastCons, phyloP, SiPhy, CADD, DANN, Eigen, FATHMM-MKL, GenoCanyon, M-CAP, MetaLR, MetaSVM, REVEL
Lopez-Fernando et al. 2017	SIFT, PolyPhen2, PROVEAN, FATHMM, PON-P2, CADD, M-CAP, Condel, Pmut
Mahmood et al. 2017	SIFT, PolyPhen1, CADD, Condel, REVEL, FATHMM, GERP++
Niroula et al. 2015	SIFT, PolyPhen2, Condel, PROVEAN, SNAP, PON-P, PON-P2
Pan et al. 2017	SIFT, PolyPhen2, PredSav, FunSAV, SusPect, SNAP, nsSNPAnalyzer
Ponzoni et al. 2018	SIFT, PolyPhen2, FATHMM-W, MutationTaster, MutationAssessor, LRT, FATHMM-U, Condel, Logit, SEQ+DYN
Raimondi et al. 2017	SIFT, PolyPhen2, LRT, MutationTaster, MutationAssessor, FATHMM, GERP++, PhyloP, SNAP, SNPs&GO, MutPred, CONDEL, CADD phred, M-CAP, DEOGEN, DEOGEN2
Rapakoulia et al. 2014	SIFT, PolyPhen2, MutationAssessor, PANTHER, EnsembleGASVR
Rogers et al. 2018	FATHMM-XF, FATHMM-MKL, GAVIN, CADD, DANN
Salgado et al. 2016	SIFT, PolyPhen2, PROVEAN, MutationAssessor, CONDEL, MutationTaster, CADD, UMD-Predictor
Thomas et al. 2016	SIFT, PolyPhen2, PolyPhen1, PANTHER-PSEC, PANTHER-PSEP, FATHMM-U, PhD-SNP, SNPs&GO, nsSNP Analyzer, SNAP, MutPred, FATHMM-W
Yates et al. 2014	SIFT, PolyPhen2, SusPect, SNPs&GO, MutPred, PhD-SNP, MutationAssessor, SNAP, FATHMM, Condel, SNPAnalyzer, PANTHER

For the predictors used in the benchmarking studies listed in Table 3.5, if the predictor's original publication provided an MCC value, we collected this information as well. Predictors and the MCC values published in their original publication are shown in **Table 3.6**.

Table 3.6 List of predictors used in benchmarking studies. Matthew's correlation coefficient (MCC) used in Figure 3.9. Table is adapted from Özkan et al., 2021.

Predictor	MCC (from original publication)	Reference (DOI)
MutationTaster2	0.8	10.1038/nmeth.2890
FATHMM-W	0.72	10.1002/humu.22225
FATHMM	0.72	10.1002/humu.22225
PON-P2	0.71	10.1371/journal.pone.0117380
MetaLR	0.71	10.1093/hmg/ddu733
MutationTaster	0.71	10.1038/nmeth0810-575
MetaSVM	0.7	10.1093/hmg/ddu733
SusPect	0.65	10.1016/j.jmb.2014.04.026
SNPs&GO	0.63	10.1002/humu.21047
LRT	0.61	10.1101/gr.092619.109
SNAP	0.58	10.1093/nar/gkm238
PON-P	0.5	10.1002/humu.22102
PredictSNP	0.5	10.1371/journal.pcbi.1003440
MutPred	0.44	10.1093/bioinformatics/btp528
PhD-SNP	0.39	10.1093/bioinformatics/btl423
FATHMM-U	0.38	10.1002/humu.22225
SIFT	0.35	10.1038/nprot.2009.86
REVEL	-	10.1016/j.ajhg.2016.08.016
PROVEAN	-	10.1371/journal.pone.0046688
CADD	-	10.1038/ng.2892
MutationAssessor	-	10.1093/nar/gkr407
PhyloP	-	10.1007/11732990_17
POLYPHEN2	-	10.1038/nmeth0410-248
GERP++	-	10.1371/journal.pcbi.1001025
FATHMM-MKL	-	10.1093/bioinformatics/btv009
CONDEL	-	10.1016/j.ajhg.2011.03.004
PolyPhen	-	10.1093/nar/gkf493
SiPhy	-	10.1093/bioinformatics/btp190
M-CAP	-	10.1038/ng.3703
DANN	-	10.1093/bioinformatics/btu703
nsSNPAnalyzer	-	10.1093/nar/gki372
PANTHER	-	10.1101/gr.772403

3.2.3. Performance assessment

To describe the performance of the bioinformatics predictors, we used the most frequently used parameters (Niroula and Vihinen, 2016): sensitivity, specificity, accuracy, MCC and area under the ROC curve. These parameters are described in detail in the Introduction, section **1.4.3. Parameters for evaluating the performance**; here we only give the formulas as a reminder.

Sensitivity, also called true positive rate (TPR) or recall:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity, also called true negative rate (TNR):

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Matthew's correlation coefficient (MCC):

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

In all formulas: TP is true positive; TN is true negative; FP is false positive; FN is false negative. Note on terminology: “positive” indicates pathogenic and “negative” indicates benign variants.

Area Under the receiver operating characteristic Curve (AUC or ROC AUC), a measure of goodness of prediction based on the ROC curve (see **1.4.3. Parameters for evaluating the performance, Figure 1.5**).

3.3. Results and Discussion

The pathogenicity prediction tools described in this chapter are presented in two separate blocks. In **3.3.1. Pathogenicity predictors: biophysical view**, we describe the biophysical approach to the pathogenicity prediction problem and the associated computational tools. In **3.3.2. Pathogenicity predictors: bioinformatics tools**, we present the bioinformatics tools that treat pathogenicity prediction as a binary classification problem. It should be noted that the original publication (Özkan *et al.*, 2021) also includes a separate block dedicated to pathogenicity prediction on splicing variants, which is the work of the collaborating authors to the paper, and for this reason is not included in this thesis.

3.3.1. Pathogenicity predictors: Biophysical view

The biophysical approach to the pathogenicity prediction problem aims to explain the connection between genetic variants and disease based on our understanding of the physics and chemistry of living organisms. Conceptually, we can distinguish two components in the relationship between a variant and its clinical phenotype: first, the variant's direct effect on the protein molecule (Kucukkal *et al.*, 2015), and second, the propagation of this effect through various levels of the biological hierarchy (cell, tissue, organ, etc.) (Yeger-Lotem and Sharan, 2015;

Zitnik and Leskovec, 2017), a process regulated by genetic background and environment. In this work, we concentrate on the first aspect, detailing the biophysical approach to analyze how variants affect protein structure and function.

3.3.1.1. Protein stability change upon mutation

Proteins exist in equilibrium between the native (correctly folded, biologically active conformation) and a range of non-native states, including unfolded or misfolded forms (Sanavia *et al.*, 2020). Protein stability ($\Delta\Delta G$) is a thermodynamic property that measures the separation between native and competing states (Alan Fersht, 1985). Depending on its location in the structure, a single amino acid replacement might result in alterations in protein stability, leading to functional disruptions (Hartl, 2017). Given this relationship, an effort has been made in assessing what would be the functional consequences of such amino acid replacements on protein stability (Stefl *et al.*, 2013).

This topic has been addressed, from a general point of view, through comparison of pathogenic and benign variants in terms of properties related to protein stability (Ferrer-Costa, Orozco and De La Cruz, 2002; Yue, Li and Moulton, 2005; Kucukkal *et al.*, 2015). These properties can be grouped into two broad classes. First, those that are intrinsic to the amino acid substitution, such as changes in hydrophobicity indexes, amino acid volume, or charge. And second, those properties that depend on the structural location of the affected residue, such as solvent accessibility, backbone strain, etc. In all cases, the properties show differences between pathogenic and benign variants, but they are not enough to discriminate them completely. In other words, the effect of pathogenic variants on protein stability is observed in varying degrees.

There are several tools available in the field devised to estimate the impact of a variant on protein stability (Yin, Ding and Dokholyan, 2007; Dehouck *et al.*, 2011; Steinbrecher *et al.*, 2017). Mostly, these tools are based on physicochemical models and generate $\Delta\Delta G$ values using weighted sums of different structure-based properties. Others, such as I-Mutant (Capriotti, Fariselli and Casadio, 2005), use sequence information when no experimental structure is available for a given protein. To illustrate the performance range of stability predictors, we scanned the field's literature to collect the observed-predicted correlation values for available predictors (**Figure 3.1**). These values were only kept if they were obtained on independent validation sets.

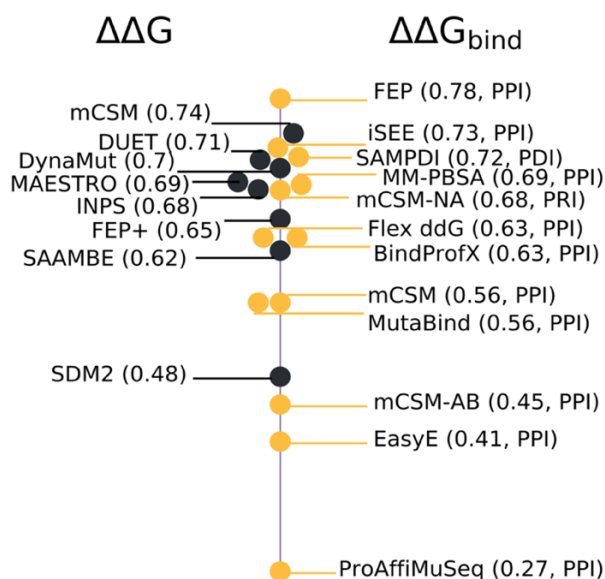


Figure 3.1. The correlation between observed and predicted values for protein stability ($\Delta\Delta G$, left side, black dots) and binding free energies ($\Delta\Delta G_{bind}$, right side, orange dots). The correlation values are shown in parenthesis next to the predictors' names. For the $\Delta\Delta G_{bind}$ predictions, the interaction type is also given inside parenthesis (PPI: protein-protein interaction; PDI: protein-DNA interaction; PRI: protein-RNA interaction).

It can be seen from **Figure 3.1 (left side)**, that these correlations fluctuate between moderate and high, ranging from 0.48 to 0.74. These values reflect the approximations applied to the modeling of $\Delta\Delta G$ as well as the quality and heterogeneity of the training datasets employed to fit the models (Yang *et al.*, 2018). These moderate correlations limit the value of stability predictors as standalone tools for the identification of pathogenic variants, indicating that they should be used with evidence from other independent sources of information (Petukh, Kucukkal and Alexov, 2015). It is also important to note that, even if a protein stability tool was 100% accurate, it does not constitute a direct measure of the protein function levels available to the cell. Unstable proteins are degraded by a complex quality control machinery that is triggered at different $\Delta\Delta G$ thresholds, depending on the protein (Sontag, Samant and Frydman, 2017). Obeying a better-safe-than-sorry principle, this machinery degrades proteins with even small stability losses (Stein *et al.*, 2019). Therefore, low $\Delta\Delta G$ predictions are inconclusive and additional experiments are needed to estimate the true impact of a variant.

3.3.1.2. Effect of variants on protein interactions

Although initial analyses suggested stability changes as the main consequence of pathogenic variants (Ferrer-Costa, Orozco and De La Cruz, 2002; Yue, Li and Moulton, 2005), they did not exclude other mechanisms. For instance, complete functional loss and the resulting clinical phenotype may only require a substitution of a catalytic side, without further affecting the protein structure (Schuster-Böckler and Bateman, 2008; Gao, Zhou and Skolnick, 2015). Additionally, pathogenic variants located on the protein surface might alter the native protein interactions (Fernández-Recio, 2011), e.g., protein-protein, protein-DNA, protein-ligand. To evaluate the effect of such variants on protein

interactions, several groups used structures of known protein complexes to focus on the fraction of pathogenic variants at the interaction interface. Initial studies suggested that the effect of pathogenic mutations on protein interactions is small (Ferrer-Costa, Orozco and De La Cruz, 2002; Schuster-Böckler and Bateman, 2008). However, these results have been challenged by the work of Sahni et al (2015), who analyzed an unbiased sample of 2890 disease-associated variants finding that they primarily affect protein interactions, not stability.

Structural analyses have been paramount in identifying the most relevant features of protein-protein interfaces (Fernández-Recio, 2011) and the effects of disruptive variants on the pathogenicity mechanisms (David and Sternberg, 2015; Navío *et al.*, 2019). This knowledge is used in combination with physicochemical principles to develop tools for predicting the effects of variants on binding affinity changes ($\Delta\Delta G_{\text{bind}}$) (Geng *et al.*, 2019). The observed-predicted correlations of these tools are shown in **Figure 3.1 (right side)**. The behavior of $\Delta\Delta G_{\text{bind}}$ predictors is similar to that of $\Delta\Delta G$ predictors, with correlations between very low (0.27) and moderately high values (0.78). As in the case of protein stability, it is important to note that $\Delta\Delta G_{\text{bind}}$ estimates are obtained using parameters intrinsic to the interacting molecules. They do not include the modulatory effect of the protein's environment, characterized by a high concentration of macromolecules known as macromolecular crowding (Rivas and Minton, 2016). Therefore, when small, the predicted $\Delta\Delta G_{\text{bind}}$ may not be relevant in vivo.

In summary, although the previous data show important advances in the characterization of variants with biophysical models, these models should be evaluated cautiously when using them for the identification of pathogenic variants. Given that disease variants can affect either protein

stability, protein interactions, or both, the estimates of $\Delta\Delta G$ and $\Delta\Delta G_{\text{bind}}$ should be examined simultaneously for a more accurate assessment of the mechanism of a variant's impact (Pires, Ascher and Blundell, 2014). Additionally, checking some simple structural properties of the variant, e.g., whether it happens at buried locations, binding hot spots, etc., can reinforce the conclusions obtained from $\Delta\Delta G$ and $\Delta\Delta G_{\text{bind}}$ values.

The focus of this block was concentrated on functional interactions; but this is only part of the picture. A protein is in continuous physical contact with a highly dense environment (Rivas and Minton, 2016), and its properties are shaped by the macromolecular crowding. From a biochemical point of view, crowding is important because it favors protein aggregation (Ellis and Minton, 2006) and plays a key role in what are known as "conformational diseases" (Carrell and Lomas, 1997). Unfortunately, theoretical approaches to crowding are not mature enough to result in an easy-to-use computational model (Minton, 2000; Rivas and Minton, 2016). At the application level, either sequence-based approaches (using physicochemical properties of the amino acids) or costly computationally structure-based simulations are applied to the study and prediction of aggregation propensity (Graña-Montes R *et al.*, 2017). Bioinformatics tools like Aggrescan (Conchillo-Solé *et al.*, 2007), Tango (Fernandez-Escamilla *et al.*, 2004) or Pasta (Walsh *et al.*, 2014) offer an alternative to infer the resulting changes upon mutation of the aggregation profile of specific regions of the protein sequence.

3.3.1.3. Applicability of biophysical models

Despite their limited accuracies, biophysical models have a clear value in identifying and understanding the role of pathogenic variants. In Ancien *et al.* (2018), the stability-based classifier SNPMusic predicts the

pathogenicity of destabilizing or stabilizing mutations with 71% accuracy in cross-validation tests. Another predictor, mCSM, was tested on variants of proteins with kinase activity, and it was able to predict the direction of the change in stability of proteins and protein affinities with correlations around 83% (Pires, Ascher and Blundell, 2014). Riera et al. (2015) utilized FoldX (Guerois, Nielsen and Serrano, 2002) to predict stability changes of variants in α -galactosidase with classification accuracy of 71%. This study also found that two conservation measures derived from multiple sequence alignments, Shannon's entropy, and position-specific scoring matrix, delivered slightly better accuracies of 73% and 76%, respectively. This last example illustrates an important point: conservation-based measures can discriminate pathogenic from benign variants with a success rate comparable to or better than that of structure-based properties. This result is relevant because, generally, we cannot apply biophysical methods to variants in proteins with unknown experimental structure. While the appearance of AlphaFold (Jumper et al., 2021) has revolutionized this situation, it is true that for disordered parts of proteins, we cannot use structure-based properties to predict the pathogenicity of variants. To provide an answer to these variants, we may turn to the use of conservation-based descriptors (Riera, Lois and De la Cruz, 2014). The rationale behind this is that functional constraints constitute an essential factor in the sequence divergence between species (Echave, Spielman and Wilke, 2016).

Independently from theoretical evolutionary considerations, the relationship between conservation and protein stability, and related structure/function features, has been established in several works. For example, different studies have unveiled a correlation of conservation measures with residue packing density (Echave, Spielman and Wilke, 2016), and with protein stability changes (Sánchez *et al.*, 2006). The

combination of technical circumstances (lack of experimental structures and good predictive value of sequence-based evolutionary information) and the pressing need for fast tools for annotating variants have led to the development of a different approach to the prediction of pathogenic variants. The resulting tools, to which we will refer as bioinformatics (or pathogenicity) predictors in the following section, were initially inspired by but not restricted to the principles of protein biophysics. Their main features and performances are discussed below.

3.3.2. Pathogenicity predictors: Bioinformatics tools

Pathogenicity predictors are widely used in bioinformatics to identify which genetic variants are likely to have a significant impact on an individual's health (Ho *et al.*, 2019). A substantial number of these tools have been published since the development of the SIFT (Ng and Henikoff, 2001), the first genome-wide predictor developed, using different methodologies. These tools vary in their predictive features, algorithms, training methods etc. However, they all address the prediction problem in a straightforward manner, skipping all the intermediate levels between molecular impact and clinical outcome. They treat pathogenicity prediction as a two-category classification problem in which variants are either pathogenic or benign and solve it using standard machine learning tools (Riera, Lois and De la Cruz, 2014; Niroula and Vihinen, 2016). This approach comprises a series of steps that do not necessarily follow a theoretical framework; they are arbitrarily established by developers according to their view of the prediction problem. Consequently, when looking for a predictor for a specific application, in addition to comparing performances, one should consider several aspects related to the choices made by the developer.

3.3.2.1. Development of a bioinformatics predictor

The development of bioinformatics predictor follows, with small variations, a four-step workflow commonly used in supervised learning (**Figure 3.2**): (i) building training datasets of benign and pathogenic variants; (ii) identifying the properties that discriminate between the two variant types; (iii) training the classification tool; and (iv) estimating its performance. Below we describe these steps are in more detail.

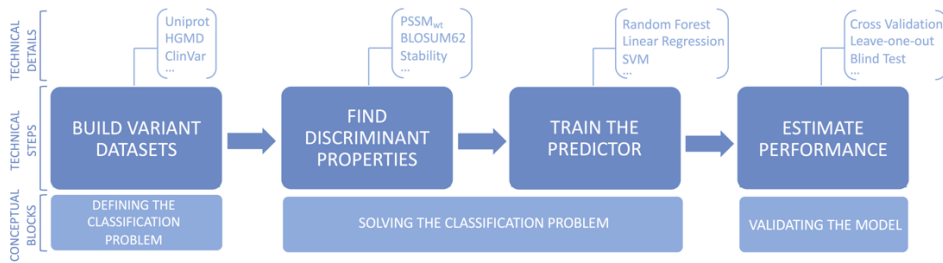


Figure 3.2. The four steps in the development of a bioinformatics predictor.

The figure illustrates the correspondence between the four technical steps (central, darker blue boxes); the main conceptual parts of the prediction problem (bottom, light blue boxes) and technical details related to these steps (top, in parentheses). Figure is taken from Özkan et al. 2021.

3.3.2.1.1. Building the variant dataset

The training set is a collection of pathogenic and benign variants that the model uses to learn about the classification task. Thus, the training set defines the prediction problem addressed and consequently, the applicability range of the resulting tool, e.g., predictors trained with variants associated to monogenic disorders should not be utilized for scoring variants in complex disease.

The pathogenic and benign variants used to derive predictors come from different sources (**Figure 3.3**). For example, the Exome Aggregation Consortium (ExAC) (Lek *et al.*, 2016) and the Genome Aggregation Database (gnomAD) (Karczewski *et al.*, 2020) are large databases of exome and genome sequencing data that can be utilized to find benign variants and information about the frequency and functional impact of variants in different populations. In the case of pathogenic variants, developers turn to large databases like UniProt (Bateman *et al.*, 2017), Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2012) and ClinVar database (Landrum *et al.*, 2016). These databases are periodically updated and apply different levels of manual curation.

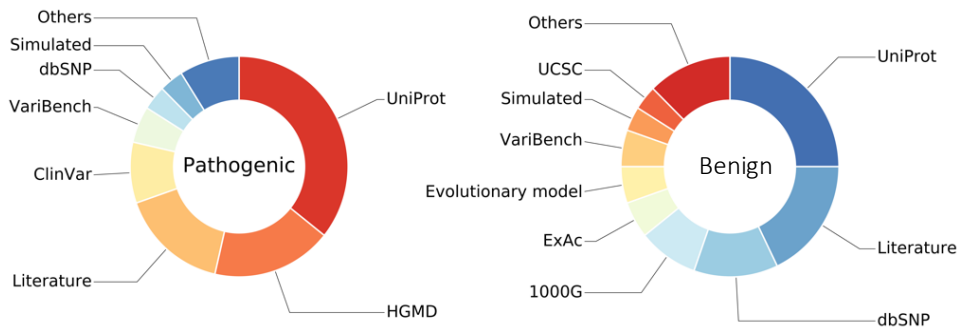


Figure 3.3. Pie charts showing the dataset source distributions of pathogenic and benign variants used to train pathogenicity predictors, respectively. Figure is adapted from Özkan *et al.* 2021.

Ensuring correctly annotated variants is essential for the development of ML models (Vihinen, 2012). Clinical data exhibit several challenges including heterogeneity, complex formats, poor annotations, class imbalance etc. that can affect the performance of the pathogenicity predictors (Hasanin *et al.*, 2020; Martínez-García and Hernández-Lemus, 2022). However, in spite of the quality of the databases or the manual curation effort, there are still several incorrect annotations (MacArthur *et*

al., 2014). For predictors trained with thousands of instances, the consequences may not be too severe. However, when the training set sizes are much smaller, such as the case of protein-specific predictors, this effect can increase the model error significantly. For these reasons, the quality and origin of the data should be examined carefully before using a pathogenicity predictor or trusting its predictive power.

3.3.2.1.2. Discriminant features

Discriminant features are used by pathogenicity predictors to distinguish between pathogenic and benign variants. Many predictors, to which we will refer as fundamental predictors, use a mixture of properties inspired in the biophysical or evolutionary view of the problem: conservation, structure or both (Tang and Thomas, 2016). In general, methods using evolutionary conservation-based features have better success rates (Niroula and Vihinen, 2016), and for this reason they tend to predominate in the field (Livesey and Marsh, 2022).

In recent years, as fundamental predictors have reached a performance plateau (Riera, Lois and De la Cruz, 2014), a new generation of methods has appeared. These new predictors, called metapredictors (also known as ensemble (Ioannidis *et al.*, 2016) or consensus methods (Vihinen, 2014)), are conceptually different than fundamental predictors: they combine the scores of preexisting pathogenicity predictors to generate the final predictions (**Figure 3.4,a**).

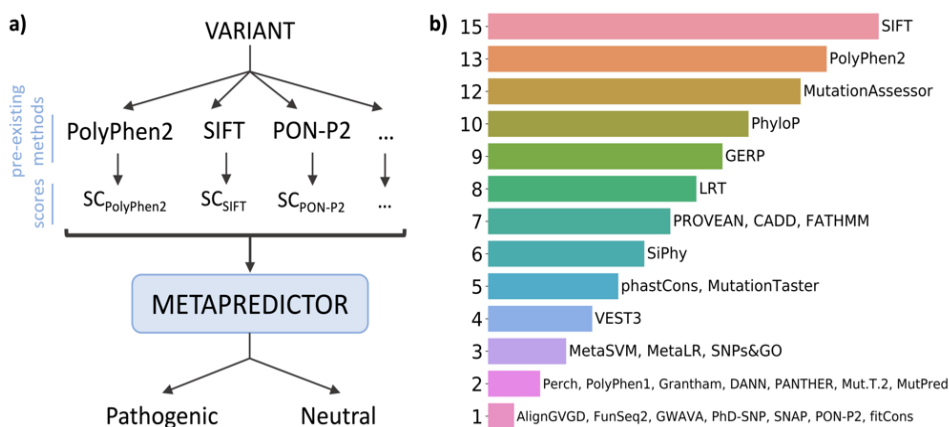


Figure 3.4. Description of Metapredictors. (a) Workflow for obtaining predictions from a metapredictor. For a given variant, metapredictors use the output (scores) of preexisting methods (SC_{SIFT} , $SC_{PolyPhen2}$, SC_{PON-P2} etc.) to generate their predictions. (b) The figure shows how many times a preexisting method has been used to build a metapredictor in the literature, obtained from a set of 23 metapredictors (Table 3.3). Figure is taken from Özkan *et al.*, 2021.

There is an increasing number of studies that show metapredictors may surpass the plateau of fundamental predictors (Ghosh, Oak and Plon, 2017; Tian *et al.*, 2019; Wilcox *et al.*, 2022). But their use in biomedical applications may entail some problems. First, they can induce unwanted biases toward the predictions of certain tools (Vihinen, 2014), like SIFT (Kumar, Henikoff and Ng, 2009) or PolyPhen2 (Adzhubei *et al.*, 2010) that are frequently used to build metapredictors (**Figure 3.4,b** and **Table 3.4**). Second, in most of the cases, metapredictors sacrifice interpretability in exchange for improved performance. They are harder to interpret because they are not based on explicit descriptors of the molecular impact of the variant. In addition, aggregating or taking majority vote of the individual predictors makes it more difficult to trace back the contribution of each model to the final decision (Vihinen, 2020).

3.3.2.1.3. Model selection

Classification algorithms are the technical core of pathogenicity predictors. These algorithms address the classification task combining a set of examples of both classes, labeled with the chosen discriminant features. The parameters of the model depend on the classification algorithm (Hastie *et al.*, 2009; Bishop, 2011). A considerable variety of algorithms have been used to build bioinformatics predictors (Neural Networks, Support Vector Machines, Random Forests, etc.), although in recent years the developer community has shown a preference for Random Forests (Breiman, 2001; Musolf *et al.*, 2022). It has to be noted that, in spite of their good performance, these algorithms have a black box nature due to the complexity of their architectures or way of capturing interactions between features, making them difficult to interpret or trace (Azodi, Tang and Shiu, 2020; Petch, Di and Nelson, 2022). Presently, improving the interpretability of such models is an important research topic in Artificial Intelligence (Rudin, 2019), particularly for biologists who are interested in understanding the numerical reasons behind predictions rather than solely seeking accurate predictions (Greener *et al.*, 2022).

3.3.2.1.4. Validation

The final step in the development of a pathogenicity predictor is validating its performance. Validation refers to the process of estimating the performance of a pathogenicity predictor on unseen data. In other words, it is a way of assessing how much the model has learned from the training data, and how efficiently it can make predictions on new, independent samples. Some of the most common are: Accuracy, MCC, Sensitivity, Specificity, Negative Predictive Value (NPV), Positive

Predictive Value (PPV), AUC-ROC, etc. There is not a unique parameter to describe the success rate of a predictor (Baldi *et al.*, 2000; Vihinen, 2012), the best option to assess performance is to consider several success rate descriptors simultaneously. However, in this comparison, it should be considered that their estimates will depend on dataset size and composition. This is particularly true for Accuracy, NPV, and PPV, whose values show substantial biases depending on the relative numbers of pathogenic and benign variants in the validation sample (Pepe, 2003).

As independent benchmark studies are regularly published, it is always a good practice to check how the candidate tools perform in these datasets. In general, a dataset of a benchmarking study should provide non-redundancy with the training set, and high quality of clinical labels (Sarkar, Yang and Vihinen, 2020). An interesting initiative to provide validation datasets meeting the previous conditions is the international contest Critical Assessment of Genome Interpretation (CAGI). (Hoskins *et al.*, 2017; Andreoletti *et al.*, 2019). In these contests, they release challenges to assess the performance of prediction methods in specific scenarios on set of variants for predicting their phenotypes. In parallel, new efforts are being made to generate nonredundant, comprehensive benchmark datasets in performance estimation (Pejaver *et al.*, 2022).

3.3.2.2. Performance of bioinformatics tools

While the main steps in the development of a predictor are essentially the same (**Figure 3.2**), their specific nature admits variations: training sets (**Figure 3.3**), discriminant properties, and algorithms differ between methods. These differences are behind the large number of predictors that have been presented in recent years. For this chapter we collected a

set of pathogenicity predictors (see section **3.2.2 The set of bioinformatics predictors**), focusing on methods published between the years 2014–2019. At the end of this process, we retrieved the performance of 45 tools (**Table 3.2**). A view of the performance of 32 predictors (those for which performance of MCC, accuracy, sensitivity and specificity metrics were available) is shown in **Figure 3.5**.

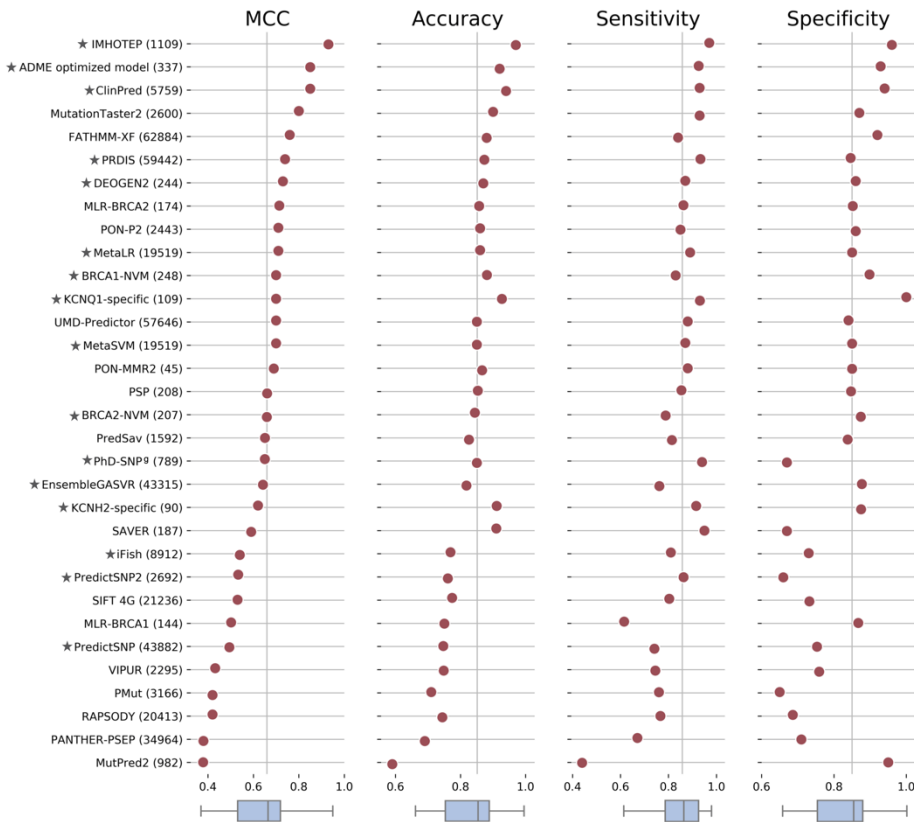


Figure 3.5. Performance of pathogenicity predictors. The figure provides four performance measures (MCC, Accuracy, Sensitivity, Specificity) for a set of 32 tools (Table 3.2) described in the literature for which the values of these four parameters are given. Asterisks are utilized to identify metapredictors. Figure is taken from Özkan et al. 2021.

It can be observed from **Figure 3.5** that, despite their simplified representations of the impact of variants, bioinformatics predictors work surprisingly well. Their performances are above random, and go from moderate to high, indicating that these tools capture the difference between benign and pathogenic variants in varying degrees. A second aspect of interest is that the ranking of the top predictors does not change, regardless of the performance descriptor used. This coherence is partly lost as we progress down the list, reinforcing the idea that we should use more than one parameter when comparing predictors. The figure also allows us to compare the performance of fundamental tools and metapredictors (marked with asterisks). We see that although the three top predictors (IMHOTEP (Knecht *et al.*, 2017), ADME-optimized model (Zhou *et al.*, 2019), and ClinPred (Alirezaie *et al.*, 2018)) are metapredictors, as we progress down the classification, there is a mixture of methods of both types.

This comparison can be refined by grouping the predictors according to their nature (fundamental vs. metapredictor). As shown in **Figure 3.6** for the same four metrics, and in **Figure 3.7** for the AUC, we can see that although metapredictors tend to outperform fundamental predictors visually, the trend is not statistically significant. Thus, we believe that in professional settings it is advisable to combine the results of both metapredictors and fundamental predictors to both optimize accuracy and provide a certain level of interpretability.

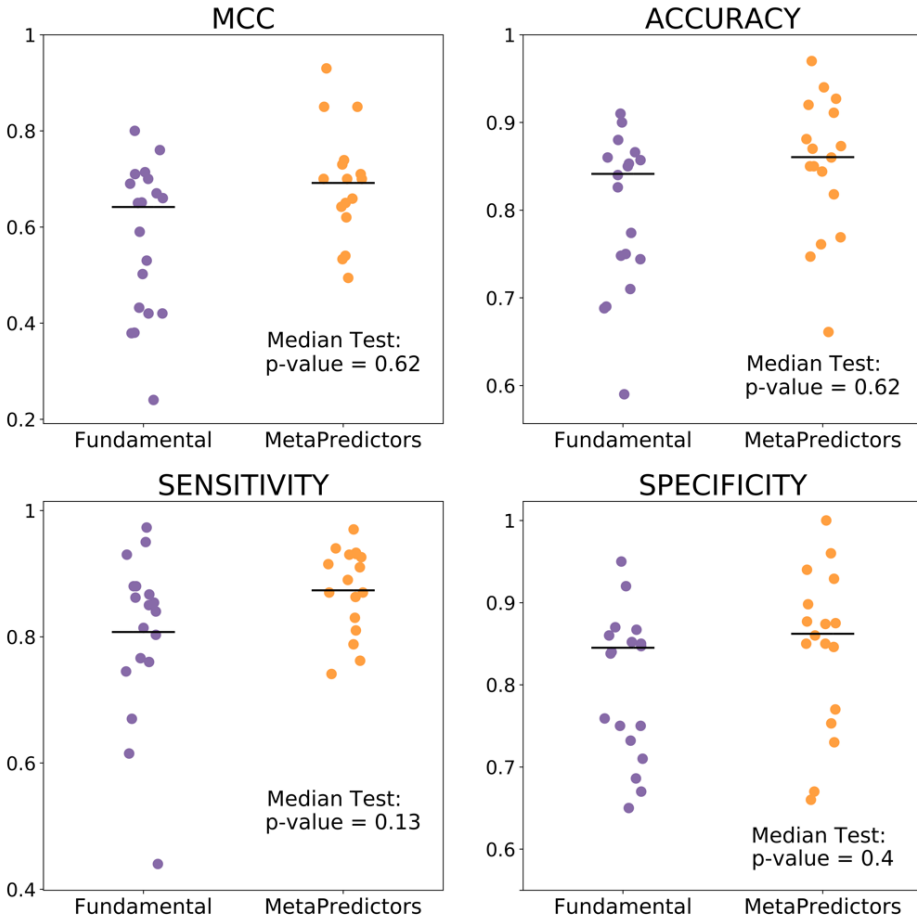


Figure 3.6. Comparison of Fundamental predictors versus Metapredictors.

The figures show the distribution of values for Fundamental (lilac dots) and Metapredictors (orange dots) for (from top to bottom and left to right): Matthew correlation coefficient (MCC), Accuracy, Sensitivity, and Specificity. Within each figure we give the results of the comparison between distributions using the Median Test. Figure is taken from Özkan et al. 2021.

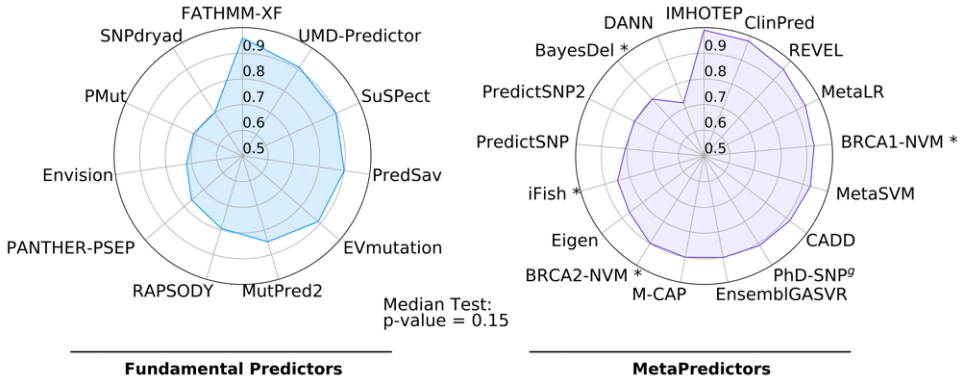


Figure 3.7. Area under the curve (AUC) values of fundamental versus metapredictors. The figure shows comparison of AUC values for those predictors for which AUC value was available. We also give the results of the Median Test for the comparison between both sets of values. Asterisks are utilized to identify protein-specific methods. Figure is taken from Özkan *et al.* 2021.

In fact, the idea of combining predictors has been proposed to increase the reliability of *in silico* evidence in the clinical setting (Richards *et al.*, 2015). However, when individual predictors do not coincide for the prediction of same variants, this approach becomes problematic. For instance, in **Figure 3.8**, the correct predictions of a selected set of methods coincide only for 12 (10 pathogenic, 2 benign) out of 50 manually annotated BRCA1 variants (Padilla *et al.*, 2019). The lack of consensus between different predictors limits the generalizability as well as reliability of the tools. The severity of the problem can be reduced using a certain combination of predictors with increased concordance to improve the reliability of *in silico* tools for the evaluation of clinically relevant variants (Ghosh, Oak and Plon, 2017).

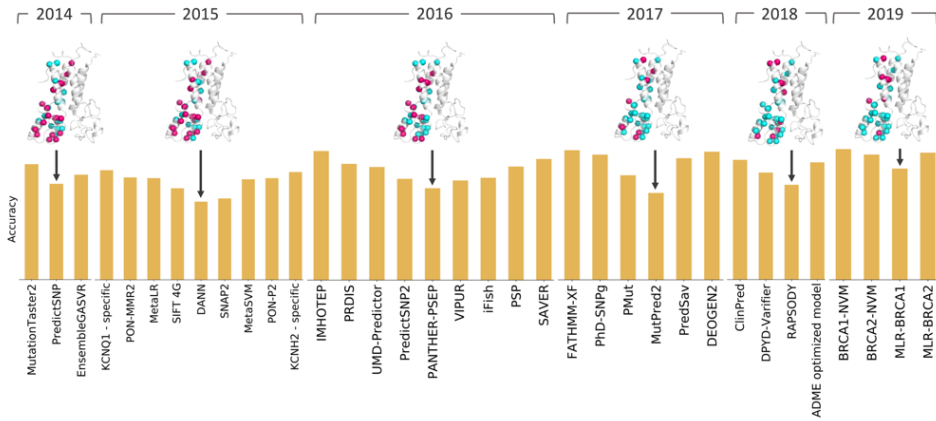


Figure 3.8. Discrepancy between pathogenicity predictors. The figure shows the disagreement, for 50 known BRCA1 variants, between the predictions of a group of selected classifiers. The structures on top correspond to the BRCA1-BARD1 complex (Brzovic et al., 2001); we utilize spheres to indicate the location of the variants correctly (blue) and incorrectly (pink) predicted by the method signaled by an arrow. The bar plot (orange) at the bottom shows the accuracies of the methods in this work, ordered according to publication year. Figure is taken from Özkan et al., 2021.

3.3.2.2.1. Variability of performance estimates

As mentioned earlier, an important aspect when validating the performance of pathogenicity predictors is the dependence of the results on the dataset used. Here, we explore this aspect for the 32 predictors, by collecting their MCC estimates found in different benchmark studies published after the original manuscript (**Table 3.5**). The summary of this analysis is shown in **Figure 3.9**.

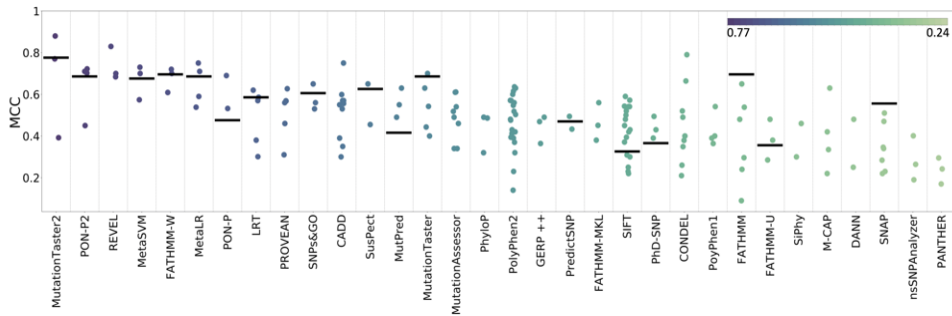


Figure 3.9. The performance of pathogenicity predictors in independent benchmarks. For a set of 32 predictors, the MCC values were obtained in a set of benchmarks generated by different authors (Table 3.5). The black horizontal lines correspond to the MCC value of the method provided by the authors in their original publication (Table 3.6). The predictors are ranked (highest to lowest) and colored (darker to lighter color) by the median of MCC values (color scale shown in the upper right of the plot). Figure taken from Özkan et al. 2021.

Figure 3.9 clearly illustrates the variation between performance estimates obtained in distinct datasets as well as the difference relative to the performance of the original study. From this result, we make three main observations. First, independent estimates can be better than the authors' estimates, e.g., this is the case for 13 out of 18 benchmarks of the SIFT predictor (Kumar, Henikoff and Ng, 2009). Second, for a given method, the range of MCC values may be large, suggesting that, to obtain a fair idea of the performance of a predictor, several independent tests should be conducted. And third, in many cases, we see an overlap between the range of values for different predictors, e.g., between SIFT and PolyPhen2, CADD and MutationTaster, etc. In this situation, it is desirable to analyze the datasets utilized in the different benchmark and how well they adjust to our problem before comparing performance estimates. Finally, it is worth mentioning that the performance of fundamental methods remains in the range of that of metapredictors,

reinforcing the idea that fundamental methods still show competitive success rates.

3.3.3. Future development and challenges

Advances in the design of bioinformatics predictors can be inspired in biophysics (in the different approaches to the genotype-to-phenotype problem) or may result from transferring successful strategies from other areas of bioinformatics or machine learning. An example of the former is the work of Sancho and his group (Angarica, Orozco and Sancho, 2015; Galano-Frutos, Garcíá-Cebollada and Sancho, 2019), who have introduced the use of molecular dynamics simulations, to find whether unfolding events may contribute to discriminate between pathogenic and benign variants. The relevance of this approach, apart from providing an original answer to the pathogenicity prediction problem, is that it opens the way to routinely complement predictions with mechanistic analyses.

Bioinformatics predictors continue to advance with the introduction of technical improvements. An example of a recent approach is the development of bioinformatics tools for specific proteins. Protein-specific predictors can be obtained using different strategies, e.g., following a *de novo* approach (Riera, Padilla and de la Cruz, 2016), adapting preexisting tools using protein-specific prediction cutoffs (Itan *et al.*, 2016), etc. Comparative studies show that specific predictors can compete in performance with generalized, non-specific predictors (Crockett *et al.*, 2012; Riera, Padilla and de la Cruz, 2016).

Approaching prediction from a different angle, Masica and Karchin (2016) propose enhancing the performance of bioinformatics methods using endophenotypes (quantitative measurements of biological properties, like catalytic activity, functional assays, etc.) as prediction targets. In this

approach, rather than simply categorizing variants as either pathogenic or benign, a quantitative prediction presents a spectrum of variant effects that allows a finer interpretation of the impact of variants. This is particularly interesting in the case of variants with moderate effects, for which pathogenicity predictors are more likely to give unreliable predictions. In this case, continuous predictions can provide a better understanding of the molecular impact of variants and conjecture a possible role for the genetic background. Increasing interest in this direction indicates an ongoing change in paradigm, resulting from a growing awareness of the continuous nature of disease (Gray *et al.*, 2018; Sarkar, Yang and Vihinen, 2020).

The abundance in new developments from widely different areas leaves predictive performance as the unique, unifying criterion for the selection of pathogenicity predictors by users. As a consequence, researchers are devoting their efforts to the validation of pathogenicity predictors from different angles. Grimm *et al.* (2015) described how the complexity of predictors and training procedures results in unnoticed overlaps between training and validation datasets. These authors refer to these overlaps as circularities and identify two types of them. Circularity of Type 1 corresponds to the coincidence of variants in the training and validation datasets and is undesirable because it results in inflated performance estimates. It is unclear to which extent circularity of Type 2, which happens when a protein contributes different variants both to the training and validation datasets, is also undesirable: in certain circumstances, it is standard practice.

3.4. Conclusion

Bioinformatics predictors are relatively straightforward to use, cost-effective, and readily available. While they offer valuable benefits, their applicability in clinical settings far from trivial. Evaluations on diverse datasets show both capabilities and limitations of these models, showing the importance of interpreting the results of such models cautiously. In general, establishing a reliable procedure for estimating the performance of bioinformatics predictors may prove to be one of the most critical challenges in the field, which, if unsolved, may severely limit the applicability of these tools.

Looking ahead, it is important to recognize that the development of quantitative predictors for estimating the molecular impact of variants holds the potential of a more nuanced solution to the variant interpretation problem. We devote the next two chapters of this thesis to describe our efforts in this direction, with overall the goal of generating clinically useful pathogenicity predictors based on quantitative predictions.

4.

**Endophenotype-based strategy for
pathogenicity prediction in BRCA1 & BRCA2
missense variants**

The results presented in this part of the thesis have been published in "Özkan S, Padilla N and de la Cruz X (2021). Towards a New, Endophenotype-Based Strategy for Pathogenicity Prediction in BRCA1 and BRCA2: In Silico Modelling of the Outcome of HDR/SGE Assays for Missense Variants. *International journal of molecular sciences*, 22(12), 6226."

4.1. Introduction

BRCA1 and *BRCA2* (*BRCA1/2*) are high-risk breast and ovarian cancer predisposition genes that play critical roles in the management of the DNA repair mechanism (Roy, Chun and Powell, 2012). Germline pathogenic variants in these genes are responsible for 20-25% of hereditary breast and ovarian cancers (HBOC) (Hart *et al.*, 2018). Most of the well-known pathogenic variants in these genes lead to truncated proteins, easy to interpret as disruptive/pathogenic, given their large molecular impact. Depending on ethnicity (So *et al.*, 2019; Andreis *et al.*, 2023), 5-10% of patients with *BRCA1/2* germline variants carry missense variants of unknown significance (VUS) (Tavtigian *et al.*, 2006); that is, variants for which no molecular nor clinical phenotype can be assigned. Considering that variants labeled as VUS are not useful for medical purposes, there is a pressing need to devise effective models to estimate the cancer risk or clinical phenotype associated to *BRCA1/2* variants.

In genetic research studies, *in silico* tools have emerged as valuable resources for addressing the challenge of identifying the impact of genetic variants. In general, these tools work by providing binary predictions (pathogenic/benign) and can be applied to variants of any human protein. In the case of *BRCA1/2*, several studies have evaluated the ability of such tools on predicting variants of these proteins (Hart *et al.*, 2020; Cubuk *et al.*, 2021; Aljarf *et al.*, 2022; Qorri *et al.*, 2022). Ernst *et al.* (2018), after evaluating the performance of four prediction tools (Align-GVGD (Tavtigian *et al.*, 2006), SIFT (Kumar, Henikoff and Ng, 2009), PolyPhen2 (Adzhubei *et al.*, 2010), MutationTaster2 (Schwarz *et al.*, 2014)) on a set of 161 *BRCA1/2* missense variants, found sensitivities ranging from 0.67 to 1.00 and specificities ranging from 0.67 to 0.92. While useful insight on the impact of variants was obtained, the

study concluded that *in silico* tools are not suitable for prediction of VUS in BRCA1/2. Similarly, in the study of Sadowski et al. (2017), for a set of BRCA1/2 variants they compare the classifications given in the German Consortium of Hereditary Breast and Ovarian Cancer (GC_HBOC) database to the predictions obtained from three *in silico* tools (SIFT, PolyPhen2, MutationTaster2). They observed high number of misclassifications, affecting 417 patients with potential under or overestimation of clinical relevance. More recently, Aljarf et al. (2022), assessed the performance of BRCA1 and BRCA2-specific tools in independent variant datasets (2882 for BRCA1 and 229 for BRCA2). They reported achieving MCCs up to 0.89, a reasonably strong predictive ability, although it is worth noting that the over-optimistic results may be due to the using the same dataset for both feature selection and training the model.

Overall, *in silico* approaches yield valuable results, but are still considered unsuitable as a source of standalone evidence for diagnostic purposes for several reasons. First, they generally integrate heterogeneous information from different sources that may bias the predictions due to unknown redundancies (Martínez-García and Hernández-Lemus, 2022). Second, the results of most *in silico* tools are difficult to interpret due to their black box nature (Azodi, Tang and Shiu, 2020). This becomes problematic as interpretation will play an increasingly crucial role in the design and use of these tools in healthcare applications (Rudin, 2019).

Different approaches are being explored in the field to improve the performance and interpretability of pathogenicity predictors for clinical applications. One that stands among them, introduced by Masica and Karchin (2016), replaces the traditional objective of binary classification

by that of predicting continuous effects of variants at the molecular level. In their seminal work, the authors use the term “endophenotype” to describe this continuous data, which is “...quantitative measurements that are correlated with phenotypes via shared genetic causes (e.g., enzyme catalytic activity, serum cholesterol or glucose level, volumetric lung capacity)”. This reformulation of the prediction objective offers several advantages over the binary approach. Endophenotype prediction is closer to the genotype, making it easier to predict than clinical phenotypes, where genetic background and environmental effects can play a modulatory role, even in Mendelian disease. Additionally, replacing binary classifications by quantitative assessments is coherent with the continuous nature of the disease (Niroula and Vihinen, 2017). This innovative approach has been recently employed in our group (Padilla *et al.*, 2019) by building multiple linear regression model to predict the results of the homology-directed repair (HDR) assays on BRCA1/2 proteins. These good results obtained allowed our group to reach the second place at the prestigious international challenge CAGI5-Enigma (Cline *et al.*, 2019) for the annotation of *BRCA1/2* missense variants.

The primary objective of the study presented in this chapter is to investigate the power of regression models in predicting the molecular impact of missense variants in BRCA1/2 by building BRCA1- and BRCA2-specific models. Our starting point is the in-house multiple linear regression (MLR) model developed in our group (Padilla *et al.*, 2019). To train our models, we used the HDR (Starita *et al.*, 2017; Richardson *et al.*, 2021) and saturation genome editing (SGE) assays (Findlay *et al.*, 2018) as prediction targets and we characterized the variants using five input features. We explored the predictive performance of five regression algorithms: multiple linear regression, ridge, lasso, elastic net, and

kernel. Next, we binarized the predictions of our regression models and compared their performance in the prediction of the clinical phenotype (pathogenic/benign). Lastly, we benchmarked our models against a set of *in silico* tools.

In summary, our goal is to build regression models trained following the endophenotype prediction paradigm (Masica and Karchin, 2016) for the binary prediction of BRCA1/2 variants and see if they are comparable to those standard tools trained with clinically labeled data. The main results explained in this chapter have been published in Özkan, Padilla and de la Cruz, 2021. Selected figures, discussion points and the bibliography have been updated for this thesis.

4.2. Materials and methods

This section is described in six parts: (i) dataset retrieval; (ii) discriminant features; (iii) regression models; (iv) binarization of the regression models; (v) binary classifiers and (vi) performance assessment.

4.2.1. Dataset retrieval

Dataset used in this study was collected from two available sources in the literature: HDR assays for BRCA1/2 and SGE assay performed on BRCA1.

HDR is an assay that measures the change upon mutation in the ability of BRCA1/2 to repair DNA double-stranded breaks (Moynahan *et al.*, 1999; Towler *et al.*, 2013). For BRCA1, we retrieved HDR values for 44 variants (17 pathogenic and 27 benign) (Starita *et al.*, 2015). For BRCA2,

we retrieved HDR values for 252 variants (90 pathogenic and 162 benign) (Richardson *et al.*, 2021).

SGE is a saturation genome editing experiment whose results for BRCA1 were presented in Findlay *et al* (2018). From this study, a total of 1837 missense variants with associated SGE values were retrieved. Using the functional classification thresholds defined by the authors, variants were labeled as pathogenic (score < -1.328), intermediate (-1.328 < score < -0.748) and benign (score > -0.748). The final set constituted by 393 pathogenic, 1276 benign and 168 intermediate variants.

We also used a third variant dataset, which we called clinical phenotypes (CP), constituted by variants with known binary clinical phenotype (pathogen/benign). This dataset is a combination of two sources. One is a manually curated set of missense variants used in a previous study of our group (Padilla *et al.*, 2019). The second source is a collection of missense variants proposed in the CAGI 5-ENIGMA challenge (Cline *et al.*, 2019; Parsons *et al.*, 2019) for BRCA1/2. The resulting CP dataset consists of 350 variants (81 pathogenic and 269 benign) for BRCA1, and 308 variants (39 pathogenic and 269 benign) for BRCA2. This dataset was used for testing the predictive performance of binarized endophenotype estimates. The distribution of these datasets over the protein sequences is shown in **Figure 4.1**.

The datasets described above were used in different train-test combinations to characterize the performance of our computational models.

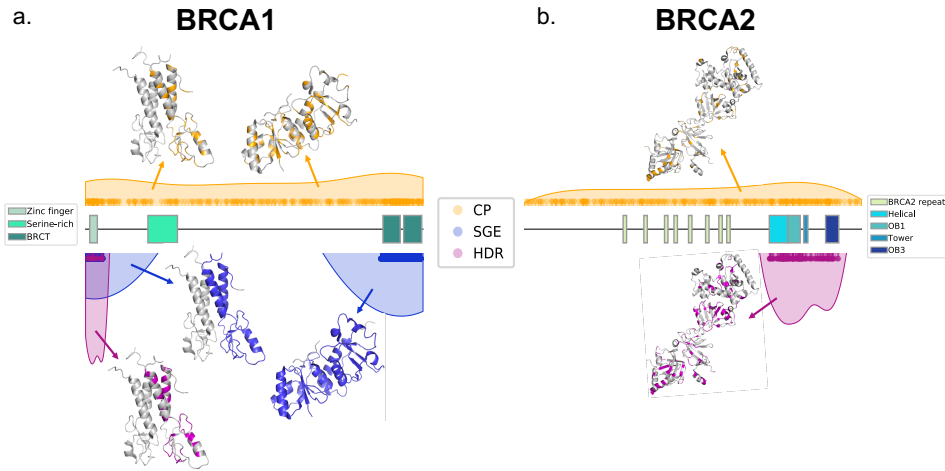


Figure 4.1: Sequence distribution of the BRCA1 and BRCA2 variants used in this work. The distributions of the three datasets mentioned in the text for (a) BRCA1 and (b) BRCA2 are shown in different colors: homology-directed repair assay (HDR, pink), saturation genome editing (SGE, blue), and clinical phenotype (CP, orange). The functional domains of the proteins are shown using rectangular boxes, and the domain names are given on the side of each protein sequence. The three-dimensional location of the variants is also shown using the three-color codes assigned to each dataset. (BRCA1: PDB 1jm7 and 1t15; BRCA2: PDB 1iyj). Figure is taken from Özkan, Padilla and de la Cruz, 2021.

4.2.2. Discriminant features

In our modeling approach, we have incorporated five features previously utilized in research conducted by our group (Riera, Padilla and de la Cruz, 2016; Padilla *et al.*, 2019). These features correspond to properties relevant to the variant function and are described below:

- **Shannon's entropy** (Cover and Thomas, 2005), a conservation descriptor derived from the multiple sequence alignment (MSA) of the

target protein's family. It is equal to: $-\sum_i p_i \cdot \log(p_i)$, where p_i is the frequency of an amino acid with index i at the variant's location in the MSA. The index i goes over all natural amino acid types present at that location.

- **Position-specific scoring matrix** (*pssm-native*) (Ng and Henikoff, 2001; Ferrer-Costa, Orozco and de la Cruz, 2004), also a conservation measure derived from MSA. It reflects the conservation of the native residue. It is equal to: $\log\left(\frac{f_{nat,i}}{f_{nat,MSA}}\right)$ where $f_{nat,i}$ and $f_{nat,MSA}$ are the frequencies of the native amino acid at the variant's location, i , and in the whole MSA respectively.

To study the impact of the MSAs on these properties, we repeated all the computations using two different MSAs per protein. One, to which we refer as **Omsa**, was retrieved from the Align-GVGD server (Tavtigian *et al.*, 2008) (<http://agvgd.hci.utah.edu/about.php> accessed on March 2018). The second MSA, to which we refer as **Hmsa**, was built following the automatic procedure described in Riera *et al.* (2016). This procedure includes: recovery of all BRCA1/2 homologs using a sequence similarity search of Uniref100 (Suzek *et al.*, 2007); filtering out distant homologs (< 40% sequence identity); and alignment of remaining sequences with Muscle (Edgar, 2004).

- **Blosum62 matrix elements** (Henikoff and Henikoff, 1992), obtained from this well-known substitution matrix.
- **van der Waals volume** (Bondi, 1964), a measure of the size of the amino acid side chain.
- **Hydrophobicity scale** (Fauchere, 1983), corresponds to the free energy of water/octanol transfer.

4.2.3. Regression models

We built five BRCA1-specific and five BRCA2-specific regression models for the prediction of functional assay values of BRCA1/2 variants. All models were built using python package scikit-learn (Pedregosa *et al.*, 2011). The regression techniques implemented are described below:

Multiple linear regression (MLR) is a statistical technique that models a linear relationship between a set of independent variables and a dependent variable (target) (Hastie *et al.*, 2009).

Ridge regression is a regularized regression technique that applies L2 regularization to shrink the size of the coefficients of the independent variables (Hastie *et al.*, 2009). The amount of shrinkage depends on the complexity parameter λ . In our Ridge models we used λ values ranging from 0.01 to 5, with a total of 20 evenly spaced values in between.

Lasso regression is also a regularized regression technique that applies L1 regularization to the coefficients, a penalty term proportional to the absolute magnitude of the coefficients (Hastie *et al.*, 2009). In this case sufficiently large λ values can shrink coefficients to exactly zero. In our Lasso models we used λ values ranging from 0.05 to 5, with a total of 15 evenly spaced values in between.

Elastic Net regression is a combined version of Ridge and Lasso. In this case, both L1 and L2 regularizations are applied to the coefficients of the input features to balance the advantages of both penalties (Hastie *et al.*, 2009). In our Lasso models we used: λ values ranging from 0.05 to 5 with a total of 15 evenly spaced values in between for L2; values ranging from 0.1 to 1 with a total of 10 evenly spaced values in between for L1 regularization.

Kernel regression is a non-parametric regression technique that estimates the target variable by weighting the contribution of nearby training examples based on their distance and a kernel function (Hastie *et al.*, 2009). In our Kernel models with radial basis function, we used λ values ranging from 0.01 to 5, with a total of 20 evenly spaced values in between; and gamma values ranging from 0.1 to 10 with a total of 7 evenly spaced values in between.

These techniques were used for building BRCA1- and BRCA2-specific regression models using the HDR and SGE assays in different combinations. All models, except MLR, were subjected to a hyperparameter tuning process with cross-validation.

4.2.4. Binarization of the regression models

As a follow-up step to the regression models, we binarized the continuous predictions using decision thresholds given in the original functional assay publications. For BRCA1 models trained using the HDR assay dataset (Starita *et al.*, 2015), predictions below 0.53 were labeled as pathogenic, and those above were labeled as benign. For the BRCA1 models trained using the SGE assay dataset (Findlay *et al.*, 2018), predictions below -1.328 were labeled as pathogenic, and those above -0.748 were labeled as benign. For BRCA2 models trained using HDR assay dataset (Richardson *et al.*, 2021), predictions below 2.25 were labeled as pathogenic, and those above were labeled as benign.

4.2.5. Binary classifiers

For comparison purposes, we also built two binary classifiers (pathogenic/benign) per protein, using the python package scikit-learn (Pedregosa *et al.*, 2011). The binary classifiers were built using two different techniques:

Random Forest (RF), which is an ensemble learning method used in classification problems (Breiman, 2001; Hastie *et al.*, 2009). It involves building several decision trees using subsets of training data and input features and aggregating the predictions to make a final decision. The default parameters of the scikit-learn package were used for building the random forest models.

Neural Network (NN), which is another commonly used classifier. It involves building a network of interconnected neurons to learn and model the relationship between the input features and the target classes (Hastie *et al.*, 2009). It is mostly suited for capturing complex non-linear relationship in the data. We built our NN models with one hidden layer and 5 hidden units.

To obtain BRCA1- and BRCA2-specific classifiers we used the variants in the CP dataset, which were labeled as either benign or pathogenic. To correct for the imbalance between these two variant types in CP, we applied the SMOTE algorithm as implemented in the python package scikit-learn (Pedregosa *et al.*, 2011). The predictive performances of the resulting models were assessed using leave-one-out cross validation (LOOCV).

Lastly, for benchmarking purposes we predicted the variants in the CP dataset with ten standard pathogenicity predictors: Align-GVGD (Tavtigian *et al.*, 2006), SIFT (Kumar, Henikoff and Ng, 2009), PolyPhen2 (Adzhubei *et al.*, 2010), PONP2 (Niroula, Urolagin and Vihinen, 2015), CADD (Kircher *et al.*, 2014), PMut (López-Ferrando *et al.*, 2017), FATHMM (Shihab *et al.*, 2013), MutationTaster2 (Schwarz *et al.*, 2014), REVEL (Ioannidis *et al.*, 2016), and VEST4 (Carter *et al.*, 2013). The predictions were downloaded using dbNSFP database (Liu, Jian and Boerwinkle, 2011; Liu *et al.*, 2020).

4.2.6. Performance assessment

The performance of the regression model predictions is measured using Spearman's rank correlation, ρ (Winter, Gosling and Potter, 2016).

The performance of the binary predictions was measured using four parameters: sensitivity, specificity, accuracy, Matthew's correlation coefficient (MCC). The formulas of these parameters are previously described in section **1.4.3. Parameters for evaluating the performance.**

4.3. Results

The predictive performances of BRCA1 and BRCA2-specific models are described below under three main sections. In **section 4.3.1**, we focus on examining the ability of the selected predictive features in the quantitative modeling of the HDR and SGE assays for the protein variants. In **section 4.3.2**, we test the success rates of the continuous estimates when used for pathogenicity prediction. Finally, in **section 4.3.3**, we compare the results of our two in-house classification predictors against a set of ten representative *in silico* tools.

4.3.1. Modelling the impact of variants on endophenotype

BRCA1- and BRCA2-specific regression models were built using different combinations of functional assays and train/test strategies. In **Table 4.1** we describe all the combinations used for developing our models.

Table 4.1: Description of the training/test set combinations used for all the models developed in this work. Train and Test columns show the abbreviation of the functional assays used and the dataset size in parenthesis. HDR: homology directed repair assay; SGE: saturation genome editing; CP: clinical phenotype; MLR: Multiple linear regression. Table is adapted from Özkan, Padilla and de la Cruz, 2021.

PROTEIN	TRAIN	TEST	VALIDATION	MODELS
BRCA1	HDR (44)	HDR (44)	Leave-one-out cross validation	MLR
	HDR (44)	SGE (1800)	Independent assays	MLR, Ridge, Lasso, Elastic, Kernel
	SGE (1800)	HDR (44)	Independent assays	MLR, Ridge, Lasso, Elastic, Kernel
	SGE (1377)	SGE (460)	Train-test split assays	MLR, Ridge, Lasso, Elastic, Kernel
	HDR (44)	CP (321)	Independent assays	MLR, Elastic, Kernel
	SGE (1686)	CP (350)	Independent assays	MLR, Elastic, Kernel
	CP (350)	CP (350)	Leave-one-out cross validation	Random Forest, Neural Network
BRCA2	HDR (252)	HDR (252)	Leave-one-out cross validation	MLR
	HDR (189)	HDR (63)	Train-test split assays	MLR, Ridge, Lasso, Elastic, Kernel
	HDR (252)	CP (252)	Independent assays	MLR, Elastic, Kernel
	CP (308)	CP (308)	Leave-one-out cross validation	Random Forest, Neural Network

The overall workflow of the LOOCV scheme is shown in **Figure 4.2, a**. The models that were evaluated with different train/test combinations were subjected to hyperparameter tuning with cross-validation during their training, as shown in **Figure 4.3**.

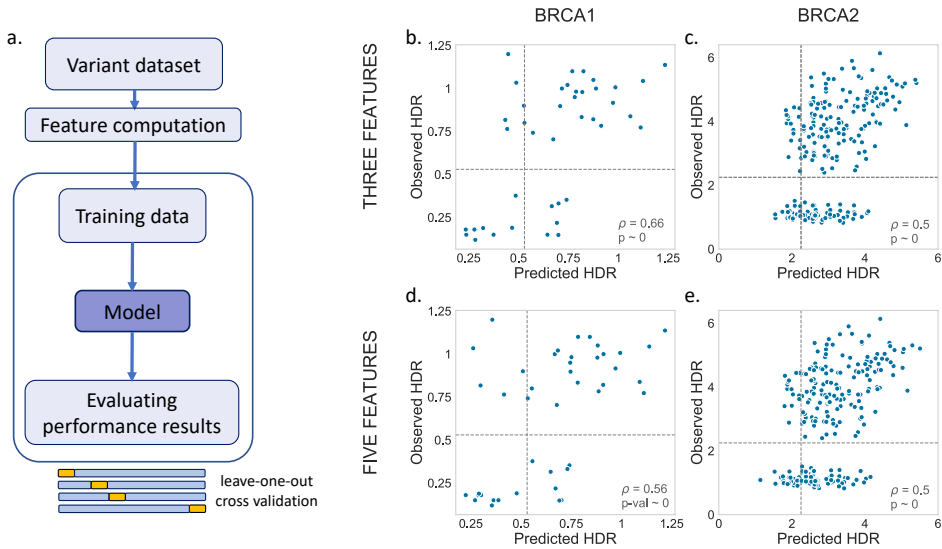


Figure 4.2: Cross-validation of the Multiple linear regression (MLR) model and observed vs predicted performances of BRCA1 and BRCA2. (a) Workflow for the leave-one-out cross validated Multiple Linear Regression model. The pipeline is applied separately for BRCA1 ((b) and (d)) and BRCA2 ((c) and (e)) using the corresponding HDR datasets. For each protein, two regression models were built: one with three and another with five input features. The scatter plots (b-d) show the cross-validated performances of the four models, where the horizontal and the vertical axes correspond to the predicted and the observed HDR. The Spearman rank correlation (ρ) values and the corresponding p -values are shown at the bottom-right of each scatter plots. The dashed lines indicate the functionality thresholds defined in the literature for these assays. Figure taken from Özkan, Padilla and de la Cruz, 2021.

4.3.1.1. Predicting the outcome of homology-directed repair (HDR) assays using multiple linear regression (MLR) models

For BRCA1, the variants from HDR dataset were too few (44 variants) to split them into train and test set shown in **Figure 4.3**. Therefore, we

started our analysis with the evaluation of MLR models following the LOOCV scheme. For BRCA2, although the HDR dataset had sufficient variants, we applied the workflow of building MLR models with LOOCV, for comparison purposes.

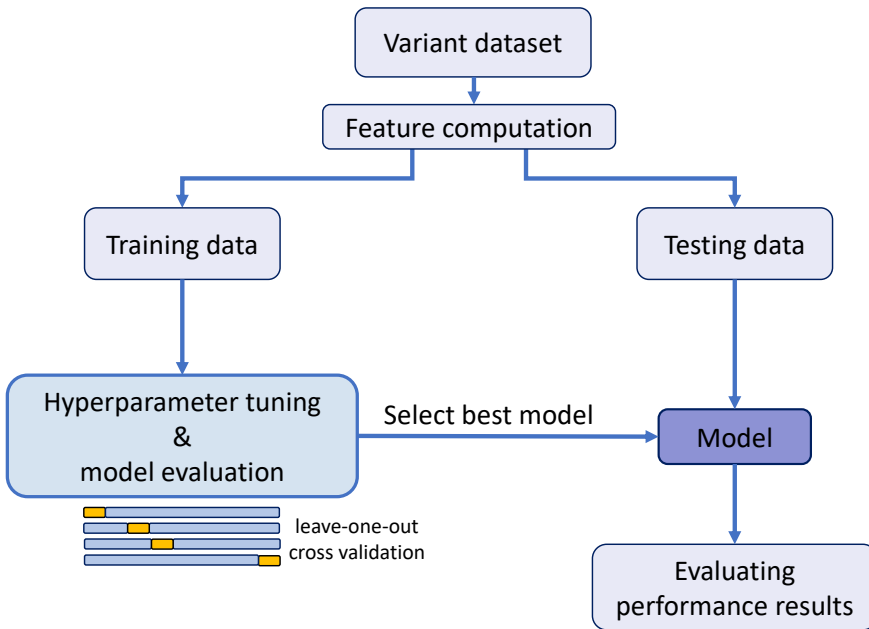


Figure 4.3: Workflow used for developing the regression models of functional assay estimates. The boxes represent the main steps of the process, which include obtention of the variant dataset, the characterization of the variants with predictive features, training and hyperparameter tuning, and performance estimation. Due to the size of the training set, hyperparameter tuning was applied using leave-one-out cross validation. Figure taken from Özkan, Padilla and de la Cruz, 2021.

For the MLR models, we built two versions; one with three features (Blosom62, Shannon's entropy and position-specific scoring matrix) and another with five features (Blosom62, Shannon's entropy, position-specific scoring matrix, van der Waals and hydrophobicity). The results are shown in **Figures 4.2.b-e**. We noticed a moderate relationship between observed and predicted values for both BRCA1 (three features: $\rho = 0.66$, $p\text{-value} \sim 0$; five features: $\rho = 0.56$, $p\text{-value} \sim 0$) and BRCA2 (three features: $\rho = 0.50$, $p\text{-value} \sim 0$; five features: $\rho = 0.50$, $p\text{-value} \sim 0$). These results show that for BRCA1/2, and within the limits imposed by our datasets, we can use the chosen features to estimate the change of an endophenotype upon a mutation with moderate accuracy.

4.3.1.2. Predicting the outcome of BRCA1/2 variants with other regression models

Next, we tested the MLR, Ridge, Lasso, Elastic and Kernel regressions using different combinations of training/test data (**Table 4.1**) following the scheme in **Figure 4.3**. The performance of each model is shown in **Figure 4.4**.

BRCA1 models are shown in **Figure 4.4 (upper box, a-o.)**, for the MLR, Elastic, Kernel, Ridge, and Lasso models. As we can see, the p values confirm a moderate predictive ability for the different regressions. Most of the results for the *train SGE/test SGE* and *train HDR/test SGE* combinations were comparable, except for the Kernel regression applied to the *train HDR/test SGE* (**Figure 4.4,h**). In this case, the resulting p value = 0.19, was lower than the other cases of the same train/test combination ($p = 0.43$).

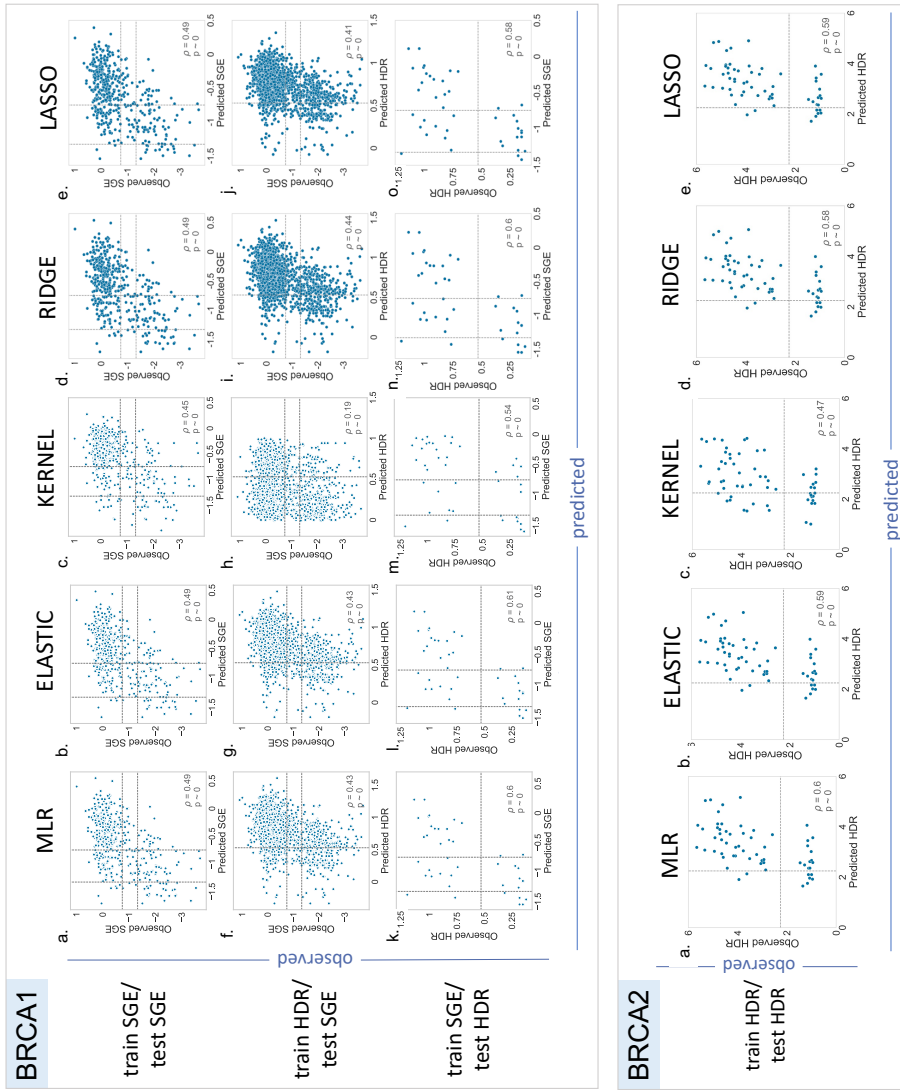


Figure 4.4: Observed vs predicted performances of five different regression models (Multiple linear regression (MLR), Elastic, Kernel, Ridge, Lasso) using Omsa-based features on three train/test set combinations for BRCA1 and one for BRCA2. BRCA1 models are shown in the upper block (a-o) and BRCA2 models in the lower block (a-e). In the scatter plots, the horizontal and vertical axes correspond to the predicted and observed values, respectively. The Spearman rank correlation (ρ) values and the

corresponding p -values are shown at the bottom-right of each scatter plot. Each row corresponds to the results obtained with a specific combination of training/test sets (see Table 4.1). In the cases where the training and test sets are the same, 75% and 25% of the variants are devoted for training and testing, respectively. The dashed lines indicate the functionality thresholds defined in the literature for these assays. Figure is adapted from Özkan, Padilla and de la Cruz, 2021.

The *train SGE/test HDR* cases deserve special mention because their p values were the highest (in the vicinity of 0.6). This improvement may be due to fact that the training set was the largest (1800 variants), and the parameter estimates were, consequently, more accurate. However, we cannot discard the possibility of a small-sample effect, resulting from the reduced number of variants, 44, in the test set.

For BRCA2 (**Figure 4.4, lower box**), we had a smaller number of variants with functional assays available thus limiting the number of train–test combinations (**Table 4.1**). The results obtained for MLR, Elastic, Kernel, Ridge and Lasso were not substantially different from those found for BRCA1: p values in the moderate range (0.47–0.60), with Kernel regression showing the poorest results ($p = 0.47$).

An aspect worth noting for both proteins is that all the parametric models (MLR, Ridge, Lasso and Elastic) displayed low or very low coefficients for the van der Waals volume and hydrophobicity features, the two new descriptors of the amino acid changes. In fact, the Lasso and Elastic methods favored an explicit elimination of hydrophobicity (**Table S4.1**). In summary, we found that for BRCA1/2, within the limits imposed by our datasets, we could estimate the quantitative trait change upon mutation with moderate accuracy, and the refinements of the coefficients with the

regularizations did not improve the predictive power significantly compared to more simplistic MLR models.

4.3.1.3. Effect of MSA on endophenotype estimates

Next, we reproduced all the results in the previous sections using the second version of the MSAs. The results obtained with Hmsa alignments are shown in **Figure 4.5**. We observed some minor differences in the ρ values relative to the Omsa results. For example, the Elastic model for the pair train SGE/test SGE gave ρ 's of 0.49 and 0.47 for Omsa and Hmsa, respectively; the Kernel approach for the pair train SGE/test HDR gave 0.54 and 0.60, respectively, etc.

Overall, the results obtained support our previous conclusion: for BRCA1 and BRCA2, we can estimate, with moderate accuracy, the endophenotype change upon mutation.

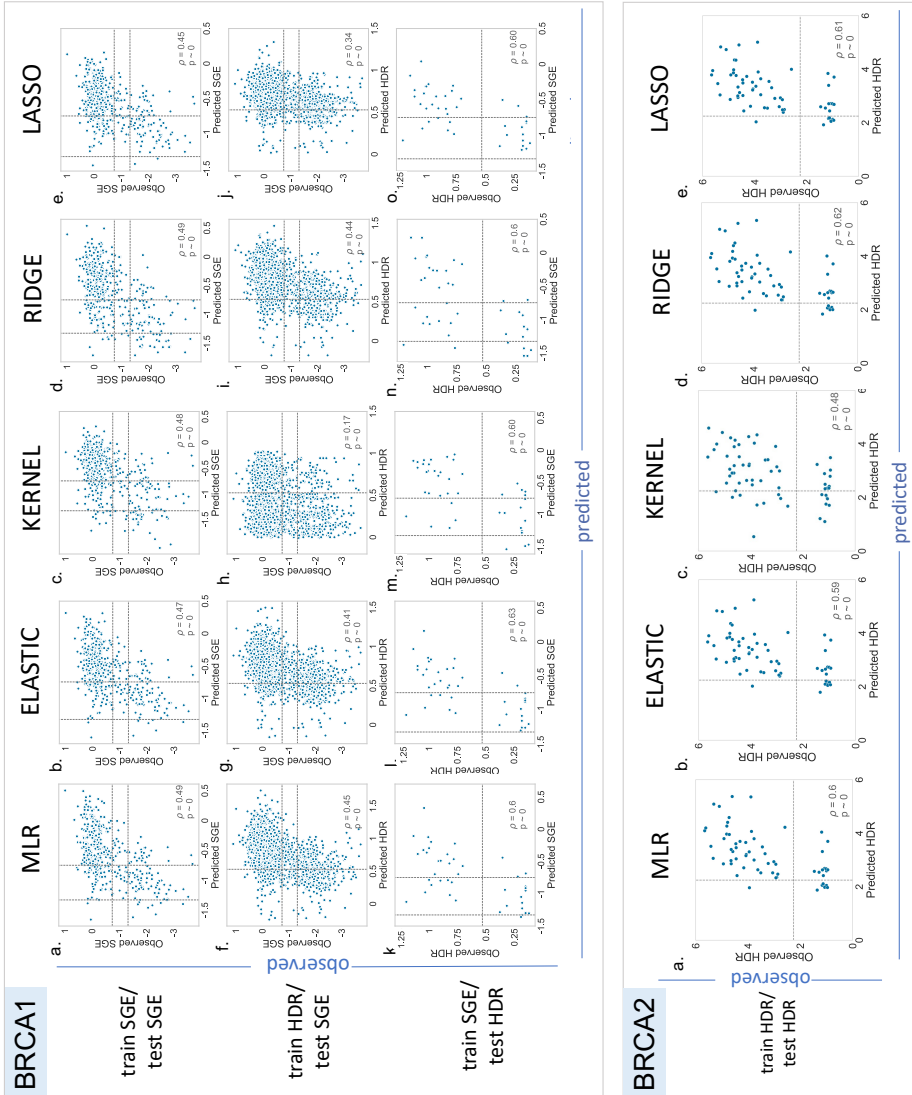


Figure 4.5: Observed vs predicted performances of five different regression models (Multiple linear regression (MLR), Elastic, Kernel, Ridge, Lasso) using Hmsa-based features. BRCA1 models are shown in the upper block (a-o) and BRCA2 models in the lower block (a-e). In the scatter plots, the horizontal and vertical axes correspond to the predicted and observed values, respectively. The Spearman rank correlation (ρ) values and the corresponding p -values are shown at the bottom-right of each scatter plot. Each

row corresponds to the results obtained with a specific combination of training/test sets (see **Table 4.1**). In the cases where the training and test sets are the same, 75% and 25% of the variants are devoted for training and testing, respectively. The dashed lines indicate the functionality thresholds defined in the literature for these assays. Figure is adapted from Özkan, Padilla and de la Cruz, 2021.

4.3.2. Performance of the regression models for pathogenicity prediction

In this section, we assess how good the quantitative predictions are for pathogenicity prediction, i.e., to which extent we can predict if a variant is pathogenic or benign using the results of the regression models. This evaluation is crucial for understanding how well these models can be used in clinical applications. To test this idea, we followed a four-step procedure. First, we obtained the CP dataset, a collection of variants of known pathogenic/benign nature (see section **4.2.1. Dataset retrieval**). Second, we trained the regression models using the HDR functional assay sets minus the variants in the CP dataset. Third, we used these models to estimate the functional assay values of the variants in the CP dataset. And fourth, we transformed the resulting outputs into binary predictions (pathogenic/benign), using the functional assay thresholds provided in the original publications (see section **4.2.4. Binarization of the regression models**).

The results obtained for BRCA1/2 are shown in **Figure 4.6** (BRCA1 in upper box, BRCA2 lower box). The accuracies and MCCs in **Figure 4.6 a** and **c** for BRCA1 and BRCA2, respectively; the sensitivities and specificities are shown in **Figure 4.6 b** and **d** for BRCA1 and BRCA2, respectively (**Table S4.2**).

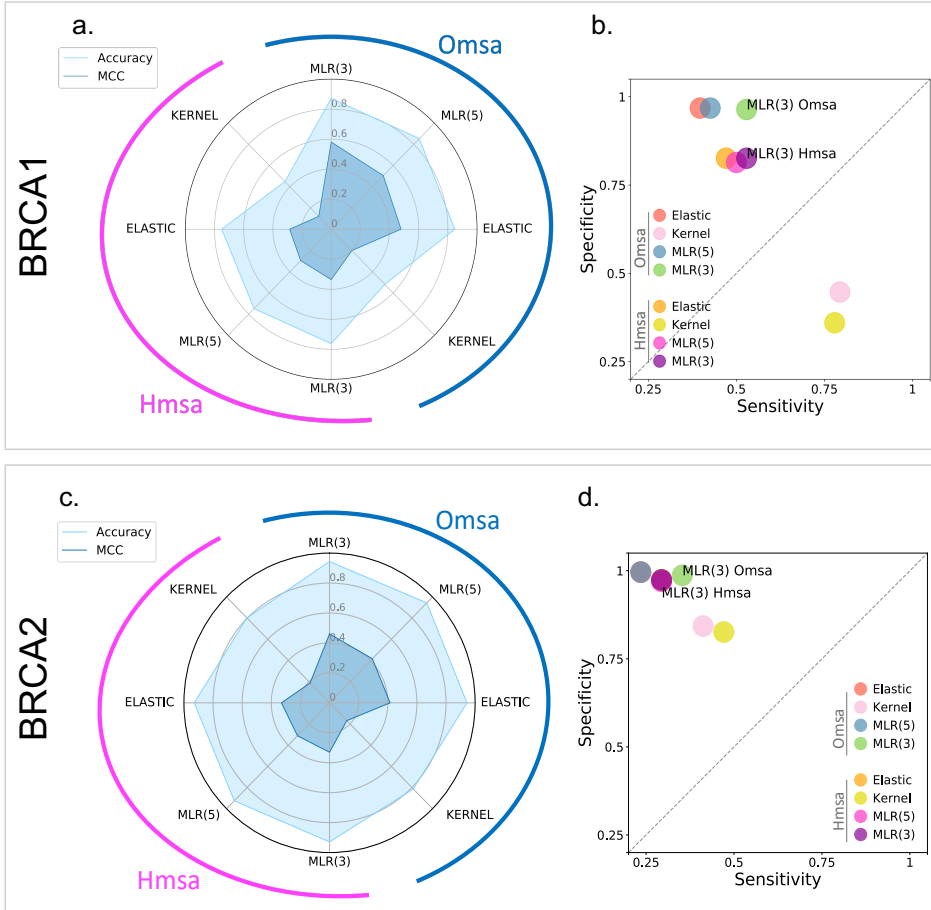


Figure 4.6: Performances of regression-based pathogenicity predictions. Using the functionality thresholds from the literature (see Materials and Methods), continuous endophenotype predictions are binarized to pathogenic/benign. The performances of the models are described using four parameters: Accuracy and MCC in (a) and (c) for BRCA1 and BRCA2, respectively. The half circles external to the radar plots identify the results obtained with different multiple sequence alignments: Omsa (blue) and Hmsa (pink). The other two parameters (sensitivity and specificity) are shown in (b) and (d) for BRCA1 and BRCA2, respectively. Each model is represented with a different color and MLR (3) model is labeled. Figure is adapted from Özkan, Padilla and de la Cruz, 2021.

Overall, we observed some common trends for both proteins. For instance, Kernel-based predictions had the poorest performances of all, with MCCs between 0.12 to 0.2 for BRCA1, and 0.17 to 0.19 for BRCA2. Furthermore, there is an apparent contradiction between the moderately good values for accuracy (75–87% for BRCA1; 80–94% for BRCA2) and the more modest MCC values (0.34–0.58 for BRCA1; 0.2–0.4, for BRCA2). This contradiction is explained by the considerable imbalance between specificity and sensitivity for all the methods (**Figures 4.6.b** and **d** for BRCA1 and BRCA2, respectively): while specificities are over 80% for both proteins, sensitivities are closer to, or below, 50% for BRCA1 and well below 50% for BRCA2. The imbalance between specificity and sensitivity is higher for BRCA2 (**Figure 4.6.d**) than for BRCA1 (**Figure 4.6.b**) and is consistent with the lower MCCs displayed for BRCA2. Lastly, the predictive performance of the Omsa-based methods was better than that of the Hmsa-based methods, particularly if we focused on the MCC. However, in both cases, the MCCs are relatively good, supporting a discriminant ability different from random.

The results indicated that all the models had a certain predictive power and that, among them, MLR was the one with the highest MCC values. We can conclude from these analyses that the computational regression estimates can be transformed into pathogenicity predictions, with moderate success rates and with a bias towards higher specificities. In general, the performances were better for BRCA1 than BRCA2 models.

4.3.3. Benchmarking the regression-based pathogenicity predictions against standard *in silico* predictions

Lastly, we built two in-house predictors, a Random Forest (RF) and a Neural Network (NN) classifier, for comparison against standard *in silico* tools to ensure a fair comparison between the regression-based predictors and the standard tools, given the difference between these tools. (see section **4.2.5 Binary classifiers**).

The results are shown in **Figure 4.7 (BRCA1 in upper box, BRCA2 lower box)**. If we first consider the in-house predictors, we find that the performances of RF and NN are similar, although from the point of view of MCC, particularly when using Omsa, RF has a somewhat better performance. For this reason, and to favor presentation clarity, we will restrict our analysis to RF predictors in the remainder of this work.

RF predictors were shown to be competitive, with MCCs (BRCA1:0.71 - BRCA2:0.62) comparable or better than those of the standard tools like PMut (BRCA1:0.65 - BRCA2:0.59), VEST4 (BRCA1:0.55 - BRCA2:0.51), and REVEL (BRCA1:0.37 - BRCA2:0.52) (**Table S4.3**). In fact, only Align-GVGD (BRCA1:0.79 - BRCA2:0.71) had better MCCs than the RFs. For BRCA1, the balance between the sensitivity and specificity values (**Figure 4.7.b**) for RF-Omsa was better than that of any other tool. For BRCA2, the situation was somewhat different, some of the best-performing standard tools (FATHMM (Shihab *et al.*, 2013), REVEL (Ioannidis *et al.*, 2016) and VEST4 (Carter *et al.*, 2013)) showed better balances than RF-Omsa. The competitiveness of the RFs dropped a little when using Hmsa instead of Omsa (**Figures 4.7.c and d**); however, they still had significant success rate. Overall, the results confirmed that RFs trained with the features used in this work have good success rates in the

pathogenicity prediction of BRCA1/2 missense variants, comparable to that of the standard tools.

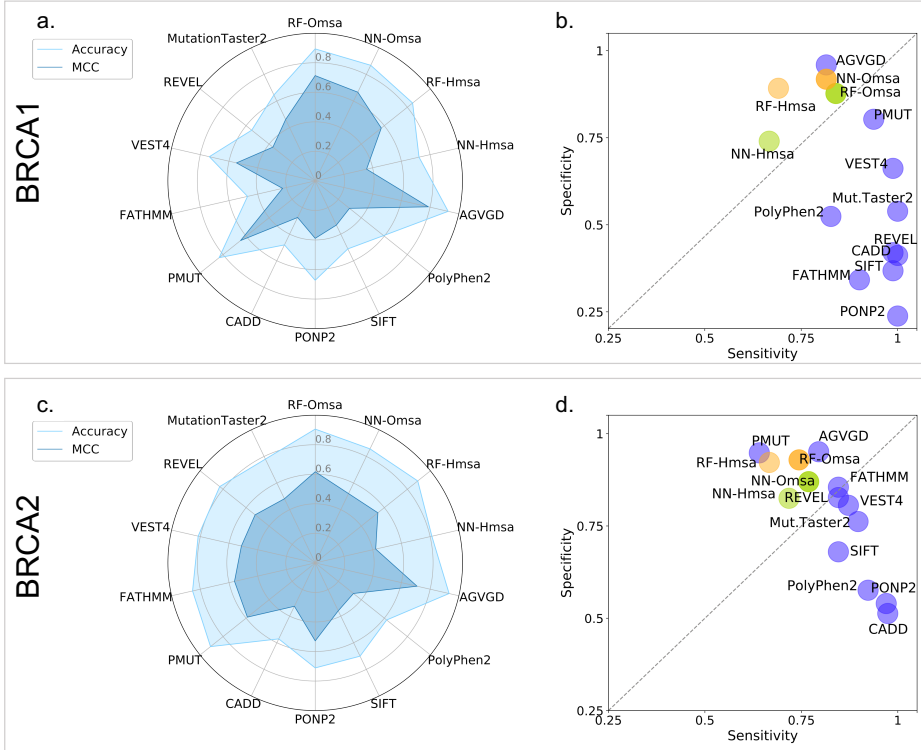


Figure 4.7: Comparing the performances of the pathogenicity predictors of BRCA1 and BRCA2 against ten standard tools. The performances of ten well-known pathogenicity predictors are compared against a family of RF and NN predictors specific for BRCA1 and BRCA2, separately. For each RF and NN predictors two versions are generated, based on the origin of the multiple-sequence alignment utilized to compute the conservation-based properties (Omsa and Hmsa). The performances of the models are described using four parameters: Accuracy and MCC in (a) and (c) for BRCA1 and BRCA2, respectively. The other two parameters, sensitivity, and specificity are shown in (b) and (d) for BRCA1 and BRCA2, respectively. Figure is adapted from Özkan, Padilla and de la Cruz, 2021.

In comparing regression-based pathogenicity predictions to those generated by the RFs, we observe that the RFs outperform the former in terms of MCC (**Figure 4.8.a**) and sensitivity (**Figure 4.8.b**). The trend is reversed in specificity (**Figure 4.8.c**) although the difference was minor. When the prediction results of the MLR-based predictors were compared against the standard predictors used in **Figure 4.7**, we found that the MLR models had better MCCs for BRCA1 (except for Pmut and Align-GVGD), a trend reversed for BRCA2. For both proteins, regression-based pathogenicity predictions have specificities clearly above sensitivities, while the opposite is the case for the standard pathogenicity predictors, except for Align-GVGD and Pmut for BRCA1.

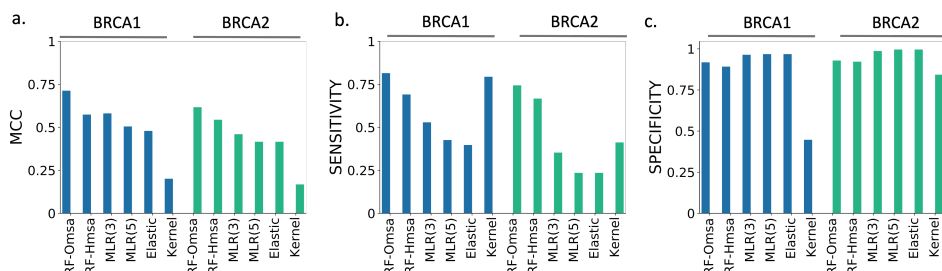


Figure 4.8: Performance comparisons between regression-based and binary predictions. The performance of RF and regression models (MLR, Elastic and Kernel) are shown using bar plots for three parameters: **(a)** MCC, **(b)** sensitivity and **(c)** specificity. BRCA1 and BRCA2 results are colored to blue and green, respectively. Figure is taken from Özkan, Padilla and de la Cruz, 2021.

Finally, we compared RF and MLR at the variant level, looking for discrepancies between them. The idea was to understand how RF outperforms MLR: by correctly predicting the same variants as MLR, plus some new cases, or by increasing the overall number of correct

predictions, regardless of whether they coincide or not with the correct MLR predictions. As shown in **Figure 4.9**, RF failed the predictions for some variants that were correctly predicted with MLR. Interestingly, we also saw that part of the RF failures were borderline cases from the endophenotype point of view.

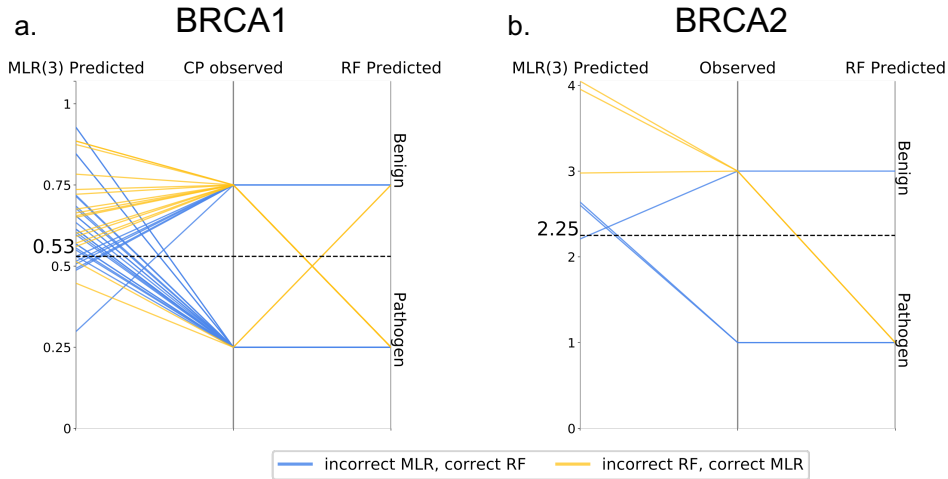


Figure 4.9: Variant-level description of the discrepancies between endophenotype-based pathogenicity predictions (Multiple Linear Regression (MLR)) and binary (Random Forest (RF)) models for BRCA1 and BRCA2. The plots are shown separately for each protein; **(a)** BRCA1 and **(b)** BRCA2. Three vertical lines in each plot represent, from left to right, the following: the values of the endophenotype estimates (obtained with MLR), the actual pathogenic classes of the variants (benign-top, pathogenic-bottom), and the RF predictions for the variants. In this representation system, each variant is characterized by a triplet, e.g., 0.75, Benign, Pathogen. The triplets are united by a line, which, in our example, would go from 0.75 (first column) to Benign (second column) and to Pathogen (third column). We display only the variants for which MLR was successful and RF failed (yellow line); and vice versa (blue line). The dashed line indicates the functionality threshold for the homology-directed repair assay. Figure is taken from Özkan, Padilla and de la Cruz, 2021.

This comparison at the variant level of the two approaches showed that, for some cases, the regression-based predictions were correct while the corresponding RF predictions failed. This indicated the existence of a certain degree of complementarity between both approaches that could support their combined use. In summary, the performances of the regression-based estimates, when used for pathogenicity prediction, were comparable to that of the standard tools in the field in the case of BRCA1 and somewhat lower for BRCA2. However, they showed an unwanted bias towards benign predictions.

4.4. Discussion

Endophenotypes are quantitative measurements of low-level phenotypes whose changes upon mutation may be easier to predict than binary pathogenic/benign status (Masica and Karchin, 2016). This proximity may simplify the prediction problem, by reducing the need for terms representing a genetic background. This is an important advantage given the high complexity of these terms (De Visser and Krug, 2014) and lack of data for their appropriate representation for several proteins.

Overall, our results indicate that the use of endophenotypes, in this case the functional assay values of BRCA1/2, for pathogenicity prediction yields reasonable outcomes. These results are nearly as competitive as those achieved with pathogenicity predictors specifically trained for predicting the clinical outcome. In accordance with previous work on our group on regression-based pathogenicity predictors (Padilla *et al.*, 2019), we find that the success rate of our predictions was different between BRCA1 and BRCA2, being lower for the latter (**Figure 4.8**). This difference, also noted by Hart *et al.* (2018) in their BRCA1/2 predictors,

indicated that genetic background may not be the only source of variability in the performances of the *in silico* tools. The properties related to the molecular functions of the protein, like 3D structure or pattern of protein–protein interactions, etc., are different between BRCA1 and BRCA2 and may not be equally captured by the MSA-based properties included in our models. This suggests that the estimates of the BRCA2-related endophenotypes may require additional properties.

While our findings show promise, it is important to acknowledge the limitations inherent in our models. The number of experimentally validated deleterious variants in BRCA1 and BRCA2, key for model development, is limited, imposing a challenge for the development of machine learning models. Indeed, data limitations lead us to use only very few parameters, mostly conservation-related features, for building our models. Although these features provide an indirect measure of a variant's impact, complementary features to improve the model performances should be added, such as structure-based properties or features related to protein stability. This need for new features is explored more in depth in our next chapter, where quantitative predictors of variant effects are generalized to all proteins.

Finally, it is important to mention that, independently from the predictive ability, an advantage that supports the use of regression-based predictions is related to the interpretability of the results. In regression models, particularly when regularization is used, the output is directly interpretable in terms of the independent variables. This degree of interpretability is not always straightforward in machine learning classifiers, given their black box nature (Rudin, 2019).

4.5. Conclusion

Although, with their present performances, *in silico* tools are not yet recommended as a stand-alone source of evidence, the recent surge of new ideas and techniques in the prediction field may change this situation sooner than expected. In our case, our results confirm that it is possible to computationally estimate the endophenotype change upon mutation, i.e., the output of functional assays, for variants in BRCA1/2 proteins. These estimations may reach an accuracy level compatible with their subsequent use for pathogenicity predictions. We found that the training of the regression models had some technical subtleties. We hope that future efforts will increase their success rates, bringing them closer to clinical applications.

4.6. Supporting information

List of material in the supporting information:

- **Table S4.1:** Regression coefficients for the different regression models in this work shown in Figure 4.4.
- **Table S4.2:** Performances of the endophenotype-based pathogenicity predictions shown in Figure 4.6
- **Table S4.3:** Performances of the pathogenicity predictors shown in Figure 4.7.

Table S4.1: Regression coefficients for the different regression models in this work shown in Figure 4.4. Column MODEL gives the type of model trained (see Materials and Methods). Column 'TRAINING/TEST SETS' contains the abbreviation of the training/test datasets; in parentheses we show the values of the hyperparameters for the different regression models. Column INTERCEPT has the values of the intercept in each regression model. The remaining columns provide the coefficients of the different properties used in the model.

PROTEIN	MODEL	TRAIN/TEST SETS	INTERCEPT	BLOSUM62	PSSM	ENTROPY	VAN DER WAALS	HYDROPHOBICITY
BRCA1	MLR	HDR/HDR_loocv REF (3)	1,73	0,04	-0,23	0,1	-	-
BRCA1		HDR/HDR_loocv	1,66	0,05	-0,22	0,1	0,002	-0,002
BRCA1		SGE/SGE	-0,48	0,14	-0,12	0,41	-0,0005	0,02
BRCA1		HDR/SGE	1,66	0,05	-0,22	0,1	0,002	-0,002
BRCA1		SGE/HDR	-0,48	0,13	-0,11	0,42	-0,001	0,02
BRCA1	RIDGE	SGE/SGE ($\alpha = 13,6$)	-0,46	0,14	-0,12	0,4	-0,0005	0,02
BRCA1		HDR/SGE ($\alpha = 7$)	1,44	0,06	-0,18	0,1	0,002	-0,003
BRCA1		SGE/HDR ($\alpha = 12,6$)	-0,47	0,13	-0,11	0,41	-0,001	0,02
BRCA1	LASSO	SGE/SGE ($\alpha = 0,05$)	-0,54	0,11	-0,08	0,36	-0,002	0
BRCA1		HDR/SGE ($\alpha = 0,05$)	1,47	0,05	-0,17	0,04	0,0006	0
BRCA1		SGE/HDR ($\alpha = 0,05$)	-0,54	0,1	-0,07	0,36	-0,002	0
BRCA1	ELASTIC	SGE/SGE ($\alpha = 0,05$ / $L1 = 0,2$)	-0,43	0,12	-0,11	0,38	-0,0009	0
BRCA1		HDR/SGE ($\alpha = 0,05$ / $L1 = 0,4$)	1,53	0,05	-0,2	0,08	0,001	0
BRCA1		SGE/HDR ($\alpha = 0,05$ / $L1 = 0,2$)	-0,42	0,11	-0,1	0,38	-0,001	0
BRCA2	MLR	HDR/HDR_loocv REF (3)	3,56	0,34	-0,13	0,82	-	-
BRCA2		HDR/HDR_loocv	3,6	0,33	-0,15	0,82	0,004	-0,09
BRCA2		HDR/HDR	3,93	0,32	-0,23	0,72	0,004	-0,07
BRCA2	RIDGE	HDR/HDR ($\alpha = 9,6$)	4	0,31	-0,25	0,63	0,004	-0,08
BRCA2	LASSO	HDR/HDR ($\alpha = 0,05$)	4	0,31	-0,22	0,6	0,003	-0,03
BRCA2	ELASTIC	HDR/HDR ($\alpha = 0,05$ / $L1 = 0,1$)	4,04	0,31	-0,24	0,62	0,004	-0,07

Table S4.2: Performances of the endophenotype-based pathogenicity predictions shown in Figure 4.6. Column ‘PREDICTOR’ provides the name of the regression model originating the endophenotype estimates, the number of features employed (F3: three, F5: five), the msa used to compute the conservation-based properties. We only give the results for the HDR-trained regression models. MCC-Matthews Correlation Coefficient; SN-Sensitivity; SP-Specificity; ACC-Accuracy; PPV-Positive Predictive Value; NPV-Negative Predictive Value; TP-True Positives (correct pathogenic predictions); TN-True Negatives (correct benign predictions); FP-False Positives (incorrect pathogenicity predictions); FN-False Negatives (incorrect benign predictions); P-Number of pathogenic variants; N-Number of benign variants; TOTAL-Total number of variants (pathogenic+benign); COVER: fraction of variants with a prediction produced.

PROTEIN	PREDICTOR	MCC	SN	SP	ACC	PPV	NPV	TP	TN	FP	FN	P	N	TOTAL	COVER
BRCA1	MLR_F3_Omsa_HDR/CP	0.581	0.529	0.964	0.872	0.8	0.884	36	244	9	32	68	253	321	1.0
BRCA1	MLR_F5_Omsa	0.505	0.426	0.968	0.854	0.784	0.863	29	245	8	39	68	253	321	1.0
BRCA1	ELASTIC_F5_Omsa	0.479	0.397	0.968	0.847	0.771	0.857	27	245	8	41	68	253	321	1.0
BRCA1	KERNEL_F5_Omsa	0.201	0.794	0.447	0.52	0.278	0.89	54	113	140	14	68	253	321	1.0
BRCA1	MLR_F3_Hmsa	0.336	0.529	0.826	0.763	0.45	0.867	36	209	44	32	68	253	321	1.0
BRCA1	MLR_F5_Hmsa	0.296	0.5	0.814	0.748	0.42	0.858	34	206	47	34	68	253	321	1.0
BRCA1	ELASTIC_F5_Hmsa	0.285	0.471	0.826	0.751	0.421	0.853	32	209	44	36	68	253	321	1.0
BRCA1	KERNEL_F5_Hmsa	0.121	0.779	0.36	0.449	0.247	0.858	53	91	162	15	68	253	321	1.0
BRCA2	MLR_F3_Omsa	0.46	0.353	0.987	0.944	0.667	0.955	6	232	3	11	17	235	252	1.0
BRCA2	MLR_F5_Omsa	0.416	0.235	0.996	0.944	0.8	0.947	4	234	1	13	17	235	252	1.0
BRCA2	ELASTIC_F5_Omsa	0.416	0.235	0.996	0.944	0.8	0.947	4	234	1	13	17	235	252	1.0
BRCA2	KERNEL_F5_Omsa	0.168	0.412	0.843	0.813	0.159	0.952	7	198	37	10	17	235	252	1.0
BRCA2	MLR_F3_Hmsa	0.33	0.294	0.974	0.929	0.455	0.95	5	229	6	12	17	235	252	1.0
BRCA2	MLR_F5_Hmsa	0.311	0.294	0.97	0.925	0.417	0.95	5	228	7	12	17	235	252	1.0
BRCA2	ELASTIC_F5_Hmsa	0.33	0.294	0.974	0.929	0.455	0.95	5	229	6	12	17	235	252	1.0
BRCA2	KERNEL_F5_Hmsa	0.188	0.471	0.826	0.802	0.163	0.956	8	194	41	9	17	235	252	1.0

Table S4.3: Performances of the pathogenicity predictors shown in Figure 4.7. Column *PREDICTOR* has the name of the tools used. All of them are mentioned (standard tools) or described (our RF and NN predictors). The names of the remaining columns correspond to the following: MCC-Matthews Correlation Coefficient; SN-Sensitivity; SP-Specificity; ACC-Accuracy; PPV-Positive Predictive Value; NPV-Negative Predictive Value; TP-True Positives (correct pathogenic predictions); TN-True Negatives (correct benign predictions); FP-False Positives (incorrect pathogenicity predictions); FN-False Negatives (incorrect benign predictions); P-Number of pathogenic variants; N-Number of benign variants; TOTAL-Total number of variants (pathogenic+benign); COVER: fraction of variants with a prediction produced.

PROTEIN	PREDICTOR	MCC	SN	SP	ACC	PPV	NPV	TP	TN	FP	FN	P	N	TOTAL	COVER
BRCA1	P38398_RF_Omsa	0.713	0.815	0.918	0.894	0.75	0.943	66	247	22	15	81	269	350	1.0
	P38398_NN_Omsa	0.667	0.84	0.877	0.869	0.673	0.948	68	236	33	13	81	269	350	1.0
	P38398_RF_Hmsa	0.574	0.691	0.892	0.846	0.659	0.906	56	240	29	25	81	269	350	1.0
	P38398_NN_Hmsa	0.358	0.667	0.74	0.723	0.435	0.881	54	199	70	27	81	269	350	1.0
	AGVGD	0.788	0.815	0.959	0.926	0.857	0.945	66	258	11	15	81	269	350	1.0
	POLYPHEN2	0.298	0.827	0.524	0.594	0.344	0.91	67	141	128	14	81	269	350	1.0
	SIFT	0.332	0.988	0.368	0.511	0.32	0.99	80	99	170	1	81	269	350	1.0
	PONP2	0.389	1.0	0.238	0.673	0.636	1.0	56	10	32	0	56	42	98	1.0
	CADD	0.275	1.0	0.412	0.481	0.184	1.0	27	84	120	0	27	204	231	1.0
	PMUT	0.648	0.938	0.803	0.834	0.589	0.977	76	216	53	5	81	269	350	1.0
	FATHMM	0.227	0.901	0.342	0.471	0.292	0.92	73	92	177	8	81	269	350	0.997
	VEST4	0.548	0.988	0.662	0.737	0.468	0.994	80	178	91	1	81	269	350	0.997
	REVEL	0.367	0.988	0.42	0.551	0.339	0.991	80	113	156	1	81	269	350	0.997
	MUTATIONASTER2	0.462	1.0	0.539	0.646	0.395	1.0	81	145	124	0	81	269	350	0.997
BRCA2	P51587_RF_Omsa	0.617	0.744	0.929	0.906	0.604	0.962	29	250	19	10	39	269	308	1.0
	P51587_NN_Omsa	0.521	0.769	0.87	0.857	0.462	0.963	30	234	35	9	39	269	308	1.0
	P51587_RF_Hmsa	0.544	0.667	0.922	0.89	0.553	0.95	26	248	21	13	39	269	308	1.0
	P51587_NN_Hmsa	0.421	0.718	0.825	0.812	0.373	0.953	28	222	47	11	39	269	308	1.0
	AGVGD	0.709	0.795	0.951	0.932	0.705	0.97	31	255	13	8	39	268	307	1.0
	POLYPHEN2	0.332	0.923	0.576	0.62	0.24	0.981	36	155	114	3	39	269	308	1.0
	SIFT	0.36	0.846	0.68	0.701	0.277	0.968	33	183	86	6	39	269	308	1.0
	PONP2	0.528	0.97	0.54	0.711	0.582	0.964	32	27	23	1	33	50	83	1.0
	CADD	0.326	0.974	0.513	0.571	0.225	0.993	38	138	131	1	39	269	308	1.0
	PMUT	0.589	0.641	0.948	0.909	0.641	0.948	25	255	14	14	39	269	308	1.0
	FATHMM	0.562	0.846	0.855	0.854	0.478	0.972	33	212	36	6	39	248	287	0.564
	VEST4	0.514	0.872	0.806	0.815	0.415	0.976	34	200	48	5	39	248	287	0.564
	REVEL	0.522	0.846	0.827	0.829	0.434	0.972	33	205	43	6	39	248	287	0.564
	MUTATIONASTER2	0.482	0.897	0.762	0.78	0.372	0.979	35	189	59	4	39	248	287	0.564

5.

Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors

5.1. Introduction

Quantitative prediction of the functional consequences of missense variants has emerged as a vital aspect across research domains (Masica and Karchin, 2016), with potential applications in protein evolution modeling, rational protein design, etc. Particularly in precision medicine, accurate estimation of the functional effects caused by missense variants has the potential to revolutionize genomic information utilization (Zhou *et al.*, 2019). Currently, the computational tools that are used for aiding in interpretation of the variants are typically binary pathogenicity predictors (Grimm *et al.*, 2015; Niroula and Vihinen, 2016; Özkan *et al.*, 2021). Despite not being perfectly accurate, these methods have proven as sufficiently effective to warrant their inclusion in the ACMG/AMP guidelines for variant interpretation (Richards *et al.*, 2015). However, they have inherent limitations that restrict their usefulness (Sun and Yu, 2019; Liu *et al.*, 2022). The simplified binary classifications fail to capture the natural gradient of variant effects, resulting in potentially inaccurate intermediate scores (Starita *et al.*, 2017; Cubuk *et al.*, 2021). Because pathogenicity predictors assume that a variant's impact is constant regardless of the genetic context, they cannot capture this influence in the disease phenotype (Vihinen, 2020; Tabet *et al.*, 2022), and therefore oversimplify that some variants may be harmful depending on the individual's overall genetic makeup. Additionally, such predictors lack interpretability due to the black box nature of the predictors (Rudin, 2019; Azodi, Tang and Shiu, 2020; Musolf *et al.*, 2022), e.g., difficulties in understanding the working mechanism or how they end up with certain decisions.

Despite recent advances in protein science, developing *de novo* theoretical models for the quantitative estimation of missense variant

effects in clinical applications is challenging. In this scenario, data-driven approaches based on deep mutational scanning (DMS) experiments are starting to provide a promising alternative (Fowler and Fields, 2014; Livesey and Marsh, 2020). DMS data integrate the molecular impact of variants on the protein's function, structure, and stability (Bhasin and Varadarajan, 2021). With the increasing amount of DMS experiments published (Livesey and Marsh, 2020), the variant datasets with functional assays could be used for the development of predictors (Masica and Karchin, 2016). This idea has recently crystallized in the general method ENVISION (Gray *et al.*, 2018), a reference predictor devised to estimate the quantitative impact of variants. More methods based on Deep Learning approaches, like ECNet (Luo *et al.*, 2021), are also making progress in this area. Collectively, these studies address the predictive challenge from various perspectives; however, their modest progress underscores the inherent complexity of the problem. Additionally, they have not been fully tested in prediction of the binary impact of variants. Therefore, development of quantitative tools for predicting the molecular impact of missense variants holds promise for handling the challenges faced with the variant interpretation problem. Given these circumstances, it is natural to question whether the quantitative approach to predicting variant effects will meet the requirements of clinical variant interpretation before its competitors: DMS experiments and binary classifiers. This consideration arises when we focus on two significant factors: Firstly, DMS experiments, the experimental counterpart to the quantitative approach, are rapidly advancing and gaining recognition for their valuable contributions to clinical diagnosis (Starita *et al.*, 2017; Gelman *et al.*, 2019; Tabet *et al.*, 2022). Secondly, binary classifiers have exhibited remarkable performance levels, prompting a reevaluation of their role in clinical variant interpretation, and highlighting their potential for variant annotation (Garcia, Andrade and Palmero, 2022; Pejaver *et al.*, 2022).

In this study, we tackle the challenge of predicting the molecular impact of missense variants by developing binary classifiers based on quantitative tools. To this end, we divide the prediction problem into two parts (**Figure 5.1**). First, we construct a set of protein-specific prediction (PSP) models of the variant's impact, using data from a selected set of DMS experiments. The results of these models are extended to any variant from any protein in the clinical genome using a simple Ensemble learning-based formalism. Second, we apply a simple approach to binarize the results of these predictors and test their predictive performance in the interpretation of pathogenic/benign variants. Through rigorous testing on an independent validation dataset, we show that this methodology can reach a predictive level comparable to that of established methods, keeping a high potential for interpretability.

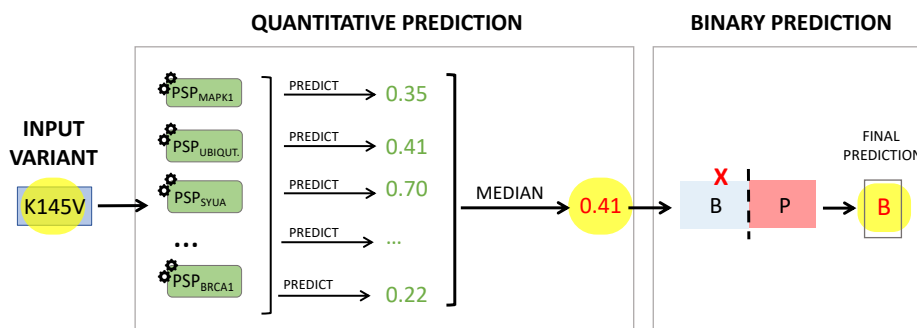


Figure 5.1: The pathogenicity prediction pipeline presented in this thesis. For a given variant in an unknown protein, we first generate a quantitative prediction using a set of PSPs trained on DMS assays. Next, we apply an Ensemble approach, using the median of the resulting values. In a second step, we apply a binarization threshold (dashed, vertical line) to transform the median value obtained (the continuous estimate) into a two-state outcome (B-Benign/P-Pathogenic for the variant).

5.2. Materials and Methods

The Material and Methods of this study are divided into seven sections: (i) dataset retrieval; (ii) discriminant features; (iii) construction of protein-specific predictors; (iv) Ensemble learning-based predictor; (v) obtaining binary predictions using the Ensemble-based predictor, (vi) benchmarking, and (vii) performance metrics.

5.2.1. Dataset retrieval

For the development of quantitative protein-specific predictors, we collected DMS experiment values. To facilitate the validation of our ensemble-model with binary prediction, we obtained clinically labeled variant datasets. The next sections delineate the procedures involved in preparing these datasets.

5.2.1.1. Collection of variant datasets from DMS experiments

We collected DMS experiments from the literature. Five publications were selected for dataset curation (Gray *et al.*, 2018; Riesselman, Ingraham and Marks, 2018; Reeb, Wirth and Rost, 2020; Dunham and Beltrao, 2021; Frazer *et al.*, 2021) where the authors have compiled and carefully curated DMS experiments for various proteins. To ensure the relevance to our study objectives, we applied a filtering process with the following criteria: availability of a minimum of 1000 missense variants per experiment (to ensure adequate amount of data for training our models) and availability of protein structure data in the AlphaFold database (Jumper *et al.*, 2021). When multiple functional assays were available for a protein, either the most recent one or the assay with more coverage was chosen, resulting in a final dataset containing one assay per protein.

All DMS experiment assays we collected, and their selection criteria are summarized in **Table S5.1**.

After applying these filters, our final dataset consisted of 30 functional assays (**Table 5.1**). The selected assays showed high heterogeneity due to variations in experimental designs and scoring ranges, leading to differences in score distributions. Moreover, some assays contained outliers that could adversely impact the learning process of algorithms. To address these issues, all scores were normalized using the normalization scheme described in the study of Gray et al. (2018). This normalization scheme can be summarized in three main steps:

- step 0: if scores higher than wild-type indicate loss-of-function, flip scores by multiplying with -1
- step 1: normalize scores to the wild-type and do log-transform:

$$S_{reported\ i} = \log_2\left(\frac{S}{S_{wild\ type}}\right)$$

- step 2: apply the normalization scheme using the formula below:

$$S_{normalized} = \frac{(S_{reported\ i} - S_{median\ synonymous})}{(-S_{median\ bottom\ 1\%})} + 1$$

where S is the wild-type normalized, log transformed functional assay score.

Lastly, after normalization, the datasets are filtered to contain only missense variants. The selected assay set, dataset source tables, wild-type scores and how normalization formula was applied are shown in **Table 5.1**.

Table 5.1: The final list of deep mutational scanning experiment datasets selected for the protein-specific predictors.

REFERENCE STUDY	UNIPROT PROTEIN IDENTIFIER	PROTEIN	ORGANISM	# OF FILTERED NOISY SYNONYMOUS VARIANTS	WILD TYPE SCORE	NORMALIZATION	ASSAY DATA REFERENCE
McLaughlin et al. 2012	P31016	PSD95	<i>R. norvegicus</i>	1577	1	Gray et al. normalized data used	Supp. Table 1 from Gray et al. (scaled_effect1)
Mejamed et al. 2015	P01447	PAB1	<i>S. cerevisiae</i>	1188	1	Gray et al. normalized data used	Supp. Table 1 from Gray et al. (scaled_effect1)
Meinlov et al. 2014	P00552	Aminoglycoside kinase	<i>K. pneumoniae</i>	4234	1	Gray et al. normalized data used	Supp. Table 1 from Gray et al. (scaled_effect1)
Mehra et al. 2016	P0829	HSP90	<i>S. cerevisiae</i>	4021	1	Gray et al. normalized data used	Supp. Table 1 from Gray et al. (scaled_effect1)
Olson et al. 2014	P06554	Protein G	<i>Streptococcus sp. group G</i>	1026	1	Gray et al. normalized data used	Supp. Table 1 from Gray et al. (scaled_effect1)
Roscoe et al. 2014	P0CG63	Ubiquitin	<i>S. cerevisiae</i>	1085	1	Gray et al. normalized data used	Supp. Table 1 from Gray et al. (scaled_effect1)
Brennan et al. 2016	P28482	MMPK1	<i>H. sapiens</i>	6810	0	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Brennan2016_MMPK1/scores_norm.npz
Fridly et al. 2018	P38398	BRCA1	<i>H. sapiens</i>	1837	0	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Fridly2018_BRCA1/scores_norm.npz
Heredia et al. 2018	P51681	CZR5	<i>H. sapiens</i>	4668	0	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Heredia2018_CZR5/scores_norm.npz
Heredia et al. 2018	P61073	CXCR4	<i>H. sapiens</i>	3877	0	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Heredia2018_CXCR4/scores_norm.npz
Kitzman et al. 2015	P04386	Regulatory protein GAL4	<i>S. cerevisiae</i>	1196	0	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Kitzman2015_GAL4/scores_norm.npz
Mateyke et al. 2018	P51580	TPMT	<i>H. sapiens</i>	3689	1	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Mateyke2018_TPMT/scores_norm.npz
Rockaf-Shmuel et al. 2015	P20589	MTH3 (haeIII)	<i>H. egyptius</i>	1350	1	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/RockafShmuel2015_MTH3/scores_norm.npz
Sarkisyan et al. 2016	P42212	GFP	<i>Ae. victoria</i>	11081	3.7	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Sarkisyan2016_GFP/scores_norm.npz
Stiffner et al. 2015	P62593	TEM1 β-lactamase	<i>E. coli</i>	4997	0	Gray et al. normalized assay data from Reeb et al.	(https://doi.org/10.17632/2zwksp7m1k.1) Parsed from data/Stiffner2015_TEM1/scores_norm.npz
Bandaru et al. 2017	P01112	Ras	<i>H. sapiens</i>	3134	0	step 2	Supp. Table 2 From Resselman et al. (RASH_HUMAN_Kurlyan)
Kesic et al. 2016	P9222	IF1	<i>E. coli</i>	1367	1	step 2	Supp. Table 2 From Resselman et al. (IF1_ECOLI_Kishony)
Welle et al. 2017	P0P23	CALM1	<i>H. sapiens</i>	1813	1	step 2	Supp. Table 2 From Resselman et al. (CALM1_HUMAN_Roth2017)
Welle et al. 2017	P63165	TPK1	<i>H. sapiens</i>	3181	1	step 2	Supp. Table 2 From Resselman et al. (TPK1_HUMAN_Roth2017)
Welle et al. 2017	Q9H354	Small ubiquitin-related modifier 1	<i>H. sapiens</i>	1700	1	step 2	Supp. Table 2 From Resselman et al. (UBC3_HUMAN_Roth2017)
Welle et al. 2017	P63279	SUMO1	<i>H. sapiens</i>	2563	1	step 2	Supp. Table 2 From Resselman et al. (SUMO1_HUMAN_Roth2017)
Welle et al. 2017	P11436	Aliphatic amidase	<i>Paeniglossa</i>	6227	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Winnbeck et al. 2017	P12391	Protein kinase Src	<i>H. sapiens</i>	3714	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Ahler et al. 2019	P04637	Cellular tumor antigen p53	<i>H. sapiens</i>	7444	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Giacomelli et al. 2018	P35520	Cysathionine beta-synthase	<i>H. sapiens</i>	10450	1	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Sun et al. 2020	Q9BQ86	Vinculin	<i>H. sapiens</i>	2695	0.9	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Chasson et al. 2020	P43246	DNA mismatch repair protein MSH2	<i>H. sapiens</i>	16749	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Jia et al. 2021	P07550	Beta-2 adrenergic receptor	<i>H. sapiens</i>	7800	1	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Jones et al. 2020	P04844	Phosphatase and tensin homolog	<i>H. sapiens</i>	6564	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Mignelli et al. 2018	P37840	Alpha-synuclein	<i>H. sapiens</i>	2599	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)
Newberry et al. 2020	P37840	Alpha-synuclein	<i>H. sapiens</i>	2599	0	step 2	Supp. Table 2 From Resselman et al. (AMIE_PSEAE_Whitehead)

This normalization process scaled variant scores. By doing so, a score equal to 1 indicated the wild-type score, while a score below 1 indicated decreased activity compared to the wild-type, and a score above 1 indicated increased activity compared to the wild-type. Distribution of the final set of 30 DMS experiments after normalization is shown in **Figure 5.2**. The distribution of each protein's DMS experiment values, before and after normalization, is shown in **Figure S5.1**.

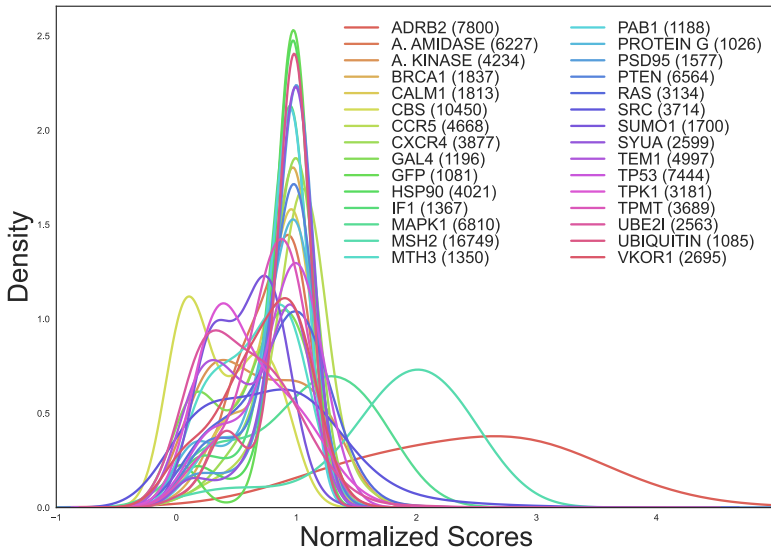


Figure 5.2: Normalized DMS experiments for the 30 proteins used in this study. The names of the proteins, colors used to represent their distribution, and the number of variants in each protein's set is shown in the legend.

5.2.1.2. Collection of clinically labeled variants

We obtained variants from HumSavar (UniProt) (Bateman *et al.*, 2017) and ClinVar (Landrum *et al.*, 2016) databases to test the predictive ability of our binary (pathogenic/benign) classifier. During the retrieval process from ClinVar, we applied several filters. Firstly, we only included

nonsynonymous variants in our analysis. Secondly, we selected variants labeled as pathogenic, likely pathogenic, pathogenic/likely pathogenic, benign, likely benign, or benign/likely benign to ensure their relevance to our study objectives. Variants with the review status “no assertion criteria provided” were excluded from our dataset. In order to have two-class labels we merged benign, likely benign, or benign/likely benign labels and named them as Likely Benign/Benign (LB/B). The same criteria were applied to pathogenic variants to name them as Likely Pathogenic/Pathogenic (LP/P). For HumSavar variants, we selected variants with labels LP/P or LB/B. Proteins present in the DMS experiment dataset were removed from both datasets. The two datasets were then merged to create a unified "clinical dataset", with priority given to the variant labels from ClinVar in case of discrepancies.

We further applied the following filters to the dataset: availability of AlphaFold structure prediction, availability of predictions from *in silico* tools selected for benchmarking, selection of proteins with more than 20 Pathogenic and 20 Benign variants. The final clinical dataset comprised 19,883 variants (12,768 pathogen and 7205 benign) from 126 unique proteins (**Table S5.2**).

5.2.2. Discriminant features

We used a total of 14 features designed to capture complementary aspects of the impact of missense variants: 5 sequence-based and 9 structure-based (**Figure 5.3**). They are described in detail below:

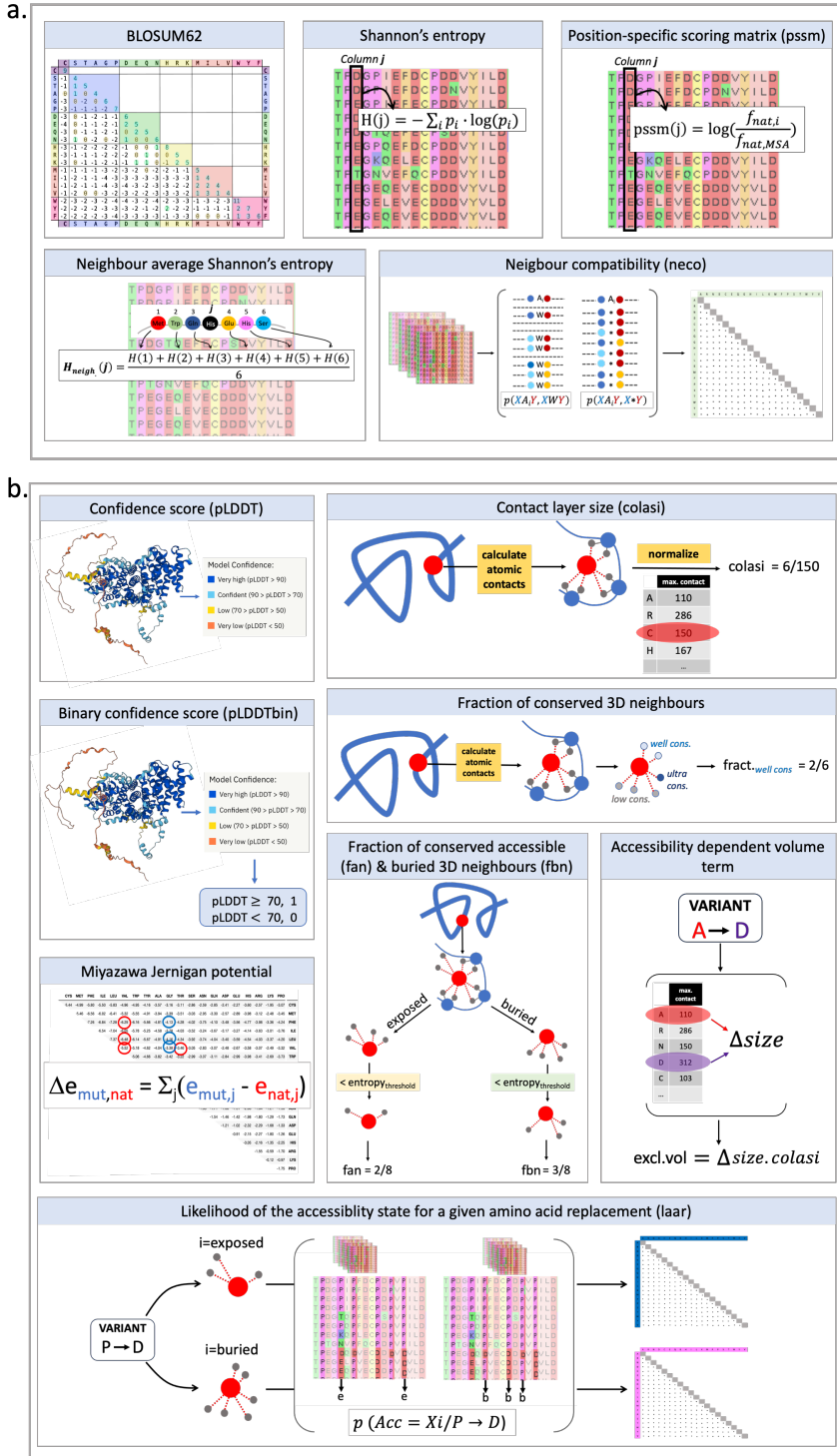


Figure 5.3: Features used to build the predictors. (a) Sequence-based features: BLOSUM62 (matrix image taken from Wikipedia, 2023, Shannon's entropy, position-specific scoring matrix (pssm), neighbor average Shannon's entropy and neighbor compatibility (neco). MSA images taken from (Bioinformatics at COMAV, no date). (b) Structure-based features: Confidence score (pLDDT) (structure image taken from AlphaFold database (Jumper, 2022)), binary confidence score (pLDDTbin), contact layer size (colasi), fraction of conserved 3D neighbors, fraction of conserved accessible (fan) and buried 3D neighbors (fbn), accessibility dependent volume term, Miyazawa Jernigan potential (matrix image taken from Miyazawa and Jernigan, 1996 and likelihood of the accessibility state for a given amino acid replacement (laar).

5.2.2.1. Sequence-based features

Sequence-based measures provide easy-to-compute proxies of the molecular impact of missense variants. We used 5 sequence-based features for our models (**Figure 5.3, a**):

1. Blosum62 matrix elements (Henikoff and Henikoff, 1992). They summarize key differences in amino acid properties, such as size and hydrophobicity, between native and mutant amino acids.
2. Shannon's entropy (Cover and Thomas, 2005), derived from the multiple sequence alignment (MSA) of the protein family. It evaluates the overall conservation at the mutation locus. It is equal to:

$$-\sum_i p_i \cdot \log(p_i),$$

where p_i is the frequency of an amino acid with index i ; i goes over all natural amino acids at the variant's location in the MSA.

3. Position-specific scoring matrix (pssm) (Ng and Henikoff, 2001; Ferrer-Costa, Orozco and de la Cruz, 2004), derived from the MSA of the protein family. It evaluates the conservation of the native residue at the mutation locus. It is equal to:

$$\log \left(\frac{f_{nat,i}}{f_{nat,MSA}} \right)$$

where $f_{nat,i}$ and $f_{nat,MSA}$ are the frequencies of the native amino acid at the variant's location, i , and in the whole MSA, respectively.

4. Neighbor average Shannon's entropy, derived from the MSA of the protein family. It is the average of the Shannon's entropy values of the sequence neighbors (-3/+3) of the mutated residue. The mutated residue is excluded from the computation.

5. Neighbor compatibility (neco). This sequence-based parameter aims to estimate the impact of neighboring residues on the native amino acid's likelihood of undergoing a specific substitution. The parameter's values constitute a 20x20 substitution matrix utilized for scoring variants. To derive these matrix elements, we utilize the differences between the human sequence and its homologs from other species, specifically when the neighbors of the human amino acid are conserved. This serves as an approximation for the mutation process in the human protein. To implement this concept, we utilize a simple probabilistic model. In this model, the element corresponding to the substitution of a native residue A_i (where subindex i represents the native residue's location in the human sequence) by a mutant residue B is computed using the formula:

$$\ln \left(\frac{p(XBY/human=XA_iY)}{p(X*Y/human=XA_iY)} \right),$$

where $p(XBY/human = XA_iY)$ is the probability that a B amino acid in an MSA column is part of an XBY pattern in a non-human protein, given that the human sequence is represented by XA_iY . It is obtained dividing the number of XBY cases in MSA columns by that of B amino acids in MSA columns, when the human sequence in all these columns has the triplet XA_iY .

In these computations, the human sequence carries the triplet XA_iY and will be excluded from the counts. Additionally, we will only consider those MSA columns having at least one B. The term $p(X * Y/h.s. = XA_iY)$ is a normalization factor. It is calculated dividing the total of $X * Y$ cases in the MSA columns by that of “ *** ” cases in the MSA columns (where XA_iY is present in the human sequence). The actual calculations are made using the MSAs of the proteins from a well-curated dataset of 593 PDB structures (Wang and Dunbrack, 2003).

The previous properties rely on the use of MSA. These MSA were obtained using a procedure described elsewhere (Riera, Padilla and de la Cruz, 2016) that we summarize here for the benefit of the reader. For a given target protein, the following steps were taken using the UniProt version of its sequence:

- a. Identify homologs of the protein, searching UniRef100 (Suzek *et al.*, 2015) using PsiBlast (Altschul *et al.*, 1997). The parameters of this search were: E-value of 0.001 and two iterations. From the PsiBlast output, we discarded those homolog sequences with less than 40% identity to the target protein. (Target-homolog sequence identity computation: The number of matching residues obtained from the global alignment of the two proteins was divided by their average length.)

- b. The remaining sequences, including the query sequence, were aligned using Muscle (Edgar, 2004) to generate the final MSA.

5.2.2.2. Structure-based features

These features are derived from the state-of-the-art structure models produced by AlphaFold (Jumper et al., 2021). They are intended to capture complementary aspects of the impact of missense variants on protein structure and function. We used the 9 structure-based features (**Figure 5.3, b**) described below.

6. AlphaFold's per residue confidence score (pLDDT) (Jumper et al., 2021). The values of this residue-level parameter are related to the reliability of the structure predictions. It has also been shown that they are related to the disordered state of the protein region around the residue (Binder *et al.*, 2022).

7. Binary confidence score (pLDDTbin). This index, derived from pLDDT, only has two values: 1 (pLDDT > 70) and 0 (pLDDT ≤ 70). The structural features described below are all dependent on pLDDTbin: if pLDDTbin = 0, the features are all set to 0.

8. Contact layer size (colasi). This parameter, which varies between 0 and 1 (except in a few cases of residues buried in highly dense locations) is a normalized measure of the native residue's contact layer size. It is a raw measure of the network of atomic interactions around the native residue. As a residue becomes more buried its colasi value will tend to be 1; colasi will be 0 when the residue has no 3D neighbors. This feature is obtained from the AlphaFold model following a two-step procedure. Firstly, compute the number of close interatomic contacts (distance ≤ 5 Å) between the native residue's atoms and those of the remainder of the

protein, using the software Arpeggio (Jubb *et al.*, 2017). Secondly, divide this number by the maximum number of interatomic contacts observed for the native residue's type in a well-curated dataset of 593 PDB structures (Wang and Dunbrack, 2003).

9. Fraction of conserved 3D neighbors. To calculate this parameter, our starting point was the set of atomic contacts of the native residue obtained for colasi (see above). Then, we focused on those cases involving highly conserved neighboring residues (Shannon's entropy < 1.37) and divided their total by the total number of contacts. The Shannon's entropy threshold chosen, 1.37 (one third of 4.12, Shannon's entropy maximum value), is defined to ensure that most residues engaged in packing networks of functional significance, including protein-protein interactions and the protein core, are included.

10. Fractions of conserved accessible 3D neighbors (fan). To compute this feature, we first obtain, for the protein of interest, the medians m_{acc} and m_{bur} of the Shannon's Entropy for the accessible ($colasi \leq 0.5$) and buried ($colasi > 0.5$) residues, respectively. To compute fan and fbn (next feature) we apply the same procedure. We divide the list of the native's residue contacts into two categories: accessible and buried. Next, we identify the accessible neighbors of the native residue and count how many of them have a Shannon's Entropy value below m_{acc} . Fan is then obtained dividing the number of atomic contacts contributed by these residues by the total number of native contacts.

11. Fractions of conserved buried 3D neighbors (fbn). Fbn is computed analogous to fan, using m_{bur} instead of m_{acc} .

12. Miyazawa-Jernigan potential. This potential, of which we use the 1996 version (Miyazawa and Jernigan, 1996), is based on a probabilistic

model that captures the predisposition of the 20 natural amino acids to contact each other. It is embodied in a 20x20 matrix and we use it as follows. Firstly, we identify the neighboring residues whose average side chain atom lies within a distance of $< 6.5 \text{ \AA}$ from that of the native residue. For glycines we employ the C-alpha atom for the distance computations. Subsequently, we calculate the difference in contact energies between the sums of native-neighbour and mutant-neighbour interactions, using the 20x20 matrix provided by Miyazawa and Jernigan (Table 3, upper diagonal) (Miyazawa and Jernigan, 1996) applying the following formula:

$$\Delta e_{mut,nat} = \sum_j (e_{mut,j} - e_{nat,j}),$$

where j goes through the list of neighbors; *mut* and *nat* refer to the mutant and native residues, and $e_{mut,j}$ and $e_{nat,j}$ are the values of Miyazawa and Jernigan matrix (upper triangle) associated to the corresponding pairs of contacting residues (e.g., $e_{mut,j}$ is the interaction energy between the *mut* residue and the neighboring residue j -th).

13. Accessibility dependent volume term. This term is designed to capture packing disruptions related to size differences between amino acids, weighting them by colasi, to consider environment differences. The formula is equal to:

$$\Delta Acc. \text{ dependent volume} = \Delta size.colasi,$$

where $\Delta size = (max\ ng_{mut} - max\ ng_{nat})$ and $max\ ng_{mut}$ is the maximum number of contacts the mutant can have (in our curated PDB dataset) and $max\ ng_{nat}$ is the same for the native. These two terms are related to the amino acid volumes of each residue.

14. Likelihood of the accessibility state for a given amino acid replacement (laar). Only two accessibility states are considered in this computation: buried (colasi >0.5) and exposed (colasi ≤0.5). The parameter's values form two 20x20 substitution matrices, one for buried and one for exposed residues. To derive the matrix elements, we approximate the mutation process in the human protein following the philosophy used for Neco (see above). For example, for a given *A to W* substitution and an accessibility state X_i (buried or exposed) of the native residue, we apply the following formula:

$$\ln \left(\frac{p(Acc=X_i/A \rightarrow W)}{p(Acc=X_i/A)} \right),$$

where $p(Acc = X_i/A \rightarrow W)$ is the probability that, given an *A to W* substitution in an MSA column, the accessibility of the human sequence is X_i . This probability is calculated by dividing the number of *A to W* substitutions in MSA columns for which the accessibility of the human sequence is X_i by the total number of *A to W* substitutions in MSA columns. $p(Acc = X_i/A)$ is a normalization factor, which is the probability that the accessibility of a sequence is X_i in all the MSA columns where the human sequence has an *A* amino acid. These computations result in the 20x20 matrices for the exposed and buried residues. The calculations are made considering all the MSAs of the proteins from a well-curated dataset of 593 PDB structures (Wang and Dunbrack, 2003).

5.2.2.3. Normalization of the features

All features were normalized during training step to transform their range to $[0,1]$ using the MinMaxScaler function in scikit-learn package (Pedregosa *et al.*, 2011). The following formula was used for scaling the features:

$$x_{TRAIN\ scaled} = \frac{x_{TRAIN} - x_{min\ TRAIN}}{x_{max\ TRAIN} - x_{min\ TRAIN}},$$

where x_{TRAIN} is the original value of the feature, $x_{min\ TRAIN}$ and $x_{max\ TRAIN}$ are the minimum and maximum values of the feature in the training dataset. To avoid any kind of bias between training and testing sets, the normalization of the features in the testing set variants was done using the $x_{min\ TRAIN}$ and $x_{max\ TRAIN}$ of the related feature column. The workflow of this process is shown in **Figure 5.4**.

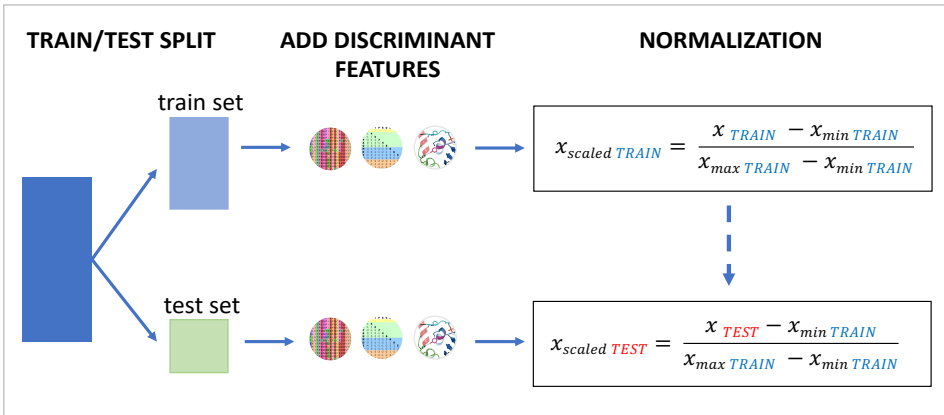


Figure 5.4: Normalization of the training and testing set features. Once the dataset is split into train and test sets, discriminant features are added to the variants in both groups, then normalization is applied: first to the training set features, and then to the test set features.

5.2.3. Construction of protein-specific predictors

A PSP model was built for each protein separately using the protein's associated DMS experimental assay. For each variant, the corresponding DMS value was used as target variable, and the 14 features corresponded to the independent variables in the regression model. Using these features and target variables, we built 30 PSP, and evaluated the performances following an individual cross-validation scheme. This workflow is illustrated in **Figure 5.5**.

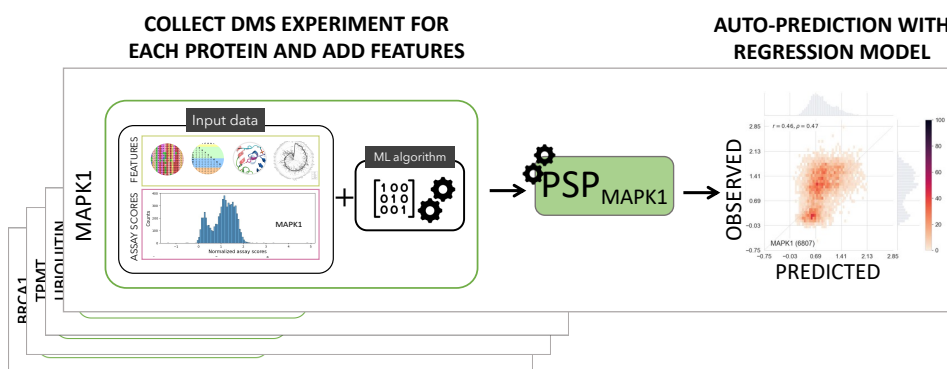


Figure 5.5: Building the 30 protein-specific predictors. For each protein in our dataset, we built a PSP using the DMS experimental values as target variable. The 14 features were attached to each variant, and a regression model was built. The performance of each predictor was then calculated using cross-validation.

In the subsequent sections, we describe the technical details of this pipeline, including the training techniques and the performance testing/validation methods.

5.2.3.1. Training technique

For our PSP we use a standard **Multiple Linear Regression (MLR)** model (Hastie *et al.*, 2009), which assumes a linear relationship between the input variable(s) and the dependent variable. The formulation is the following: $y = w_0 + w_1x_1 + w_2x_2 \dots + b_nx_n$, where y is the dependent variable, w_0 is the intercept, x_i ($i=1,n$) are the independent variables and b_i their weights. We built our MLR models for the protein variants, separately for each protein, using the values of the corresponding DMS experiments as the dependent variable, and the 14 features as the independent variables. All MLR models were obtained using python's scikit-learn package (Pedregosa *et al.*, 2011).

5.2.3.2. Undersampling

We observed an imbalanced bimodal-like distribution for most of the functional assays in our dataset. This may pose a problem when training regression models, potentially biasing the weights of various features. To address this concern, we employed a two-step **undersampling** procedure for each assay prior to model training. First, we fitted a Gaussian Mixture Model (GMM) (Pedregosa *et al.*, 2011) to each assay, to find the best threshold separating the two peaks. Second, we randomly reduced the number of variants contributed by the larger group to meet the contribution of the smaller group. This process was applied to the training set during the cross-validation procedure (**Figure 5.6**).

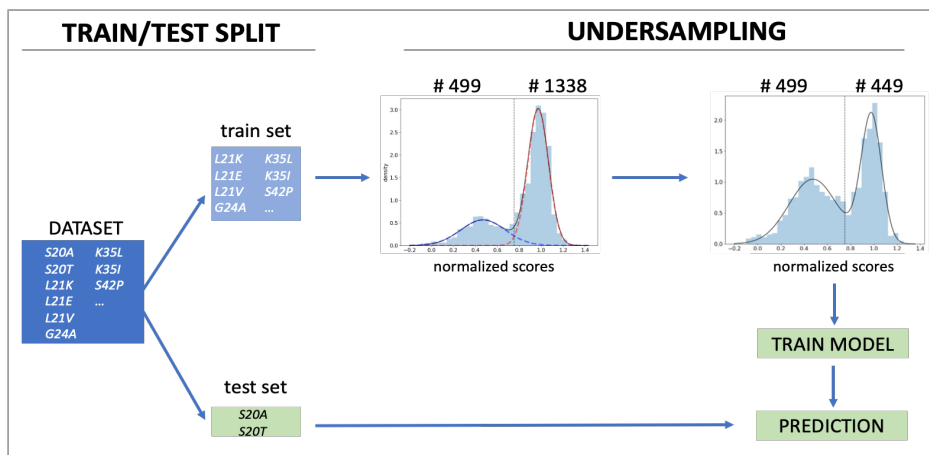


Figure 5.6: Workflow of the train/test split method using leave-one-position-out and application of undersampling for the training set. At each iteration, variants at the same position are left out while rest is used for training. Training set is fitted to GMM to identify the boundary between two peaks, and randomly selected variants are eliminated from the majority group. This data is then used for training the model to predict the variant(s) in testing set.

5.2.3.3. Cross validation

To evaluate the performance of the models, we employed a rigorous cross-validation procedure to prevent information leaks between the training and the test dataset. We will refer to the cross-validation version as leave-one-position-out (LOPO) cross validation. While it resembles the standard leave-one-variant-out cross validation, it is applied to protein positions instead of variants (**Figure 5.6**). In LOPO, all the variants at one position are collectively held out as the testing set, while the variants at the remaining positions are used as the training set. The process is iterated until all positions are left out once.

5.2.4. Generating PSP predictions for the whole clinical genome: An Ensemble learning-based predictor

The regression-based procedure explained in the previous sections is only valid for those proteins for which DMS are available. That is, for the 30 proteins in our dataset. To produce quantitative predictions for other proteins we need a different approach. Inspired in the Ensemble Learning theory we applied the following strategy (**Figure 5.1**): (i) use the PSP obtained to generate predictions for variants from other proteins; (ii) combine these predictions using the median. This approach is based on the conjecture that PSPs capture different components of the quantitative impact of variants, which follow the same physicochemical rules in all proteins. We refer to the resulting methodology as Ensemble or Ensemble-based method. In this work, we chose the subset of 10 PSP that gave the best performance when tested in other proteins. For estimating the generalization power of this approach, we followed the scheme described in **Figure 5.7**.

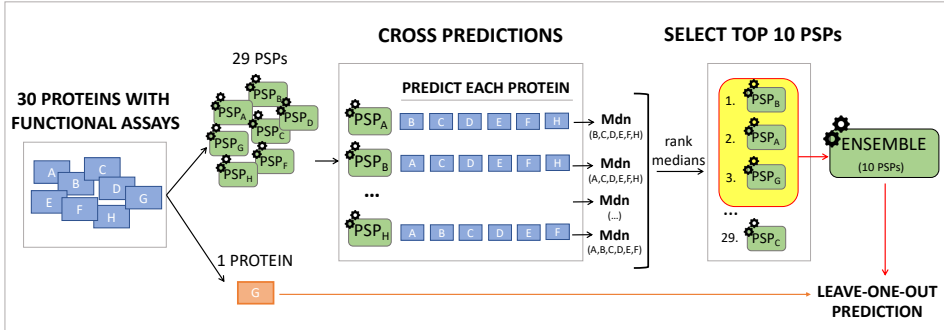


Figure 5.7: Estimating the generalization power of the Ensemble-based model. PSP: Protein-specific predictor; Mdn: Median. In this leave-one-protein-out cross validation scheme, one protein was set aside for testing while remaining 29 were used for training. Performance of each PSP in the training set was obtained by taking the median of their performances on the remaining 28 proteins and ranked to choose the top 10 PSPs. These PSPs were then combined into an Ensemble predictor for its generalization to other proteins.

5.2.5. Obtaining binary (pathogenic/benign) predictions for any protein using the Ensemble-based predictor

The second part of this project is focused on obtaining binary (pathogenic/benign) predictions for the variants, using our Ensemble-based predictor. While this part is conceptually very relevant, from a technical perspective, it presents a notably simpler task than the obtention of the PSPs. It only requires defining a threshold to binarize the results of the Ensemble-based predictor, so that any missense variant can be classified as pathogenic or benign. To obtain this threshold, or cut-off, we followed a four-step workflow, explained in **Figure 5.8**, based on the use of variants with benign/pathogenic clinical annotations (see our previous section **5.2.1.2. Collection of clinically labeled variants**). From our clinical dataset, we selected those proteins with at least 50 pathogenic and 50 benign variants. This gave a set of 26 proteins (**Table**

S5.2) to which we applied the Ensemble model to produce quantitative predictions for their clinically annotated variants, trying different threshold values, as explained in **Figure 5.8** to obtain the desired threshold value.

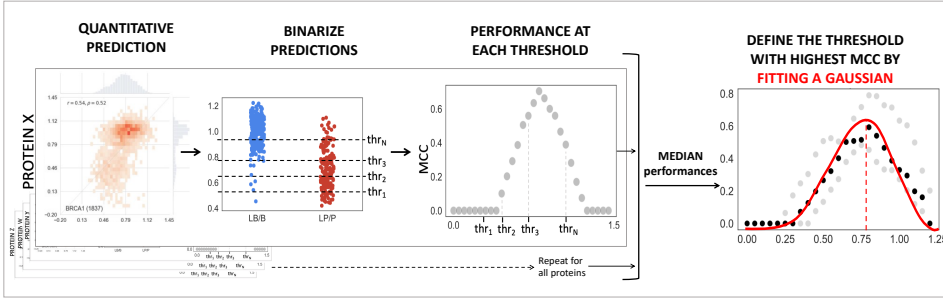


Figure 5.8: Workflow of the selection of a binarization threshold using our Ensemble predictor. The process consists of four main steps: (i) Application of the Ensemble model to the clinically annotated variants in the set of 26 proteins. (ii) and (iii) Exploring different values of the binarization threshold: we systematically explored the values of the binarization threshold (horizontal dashed lines in ‘BINARIZE PREDICTION’ figure) and for each value we computed the corresponding MCC (grey circles in PERFORMANCE OF EACH THRESHOLD figure). (iv) Obtain the optimal threshold: we pooled together the curves for all proteins, took the median of the MCCs at each threshold value (black circles), and identified, using a gaussian mixture fit, the threshold that gives the maximum MCC.

This binarized model was then validated on the remainder of the clinical dataset variants, which consisted of 100 proteins with 12,691 variants.

5.2.6. Benchmarking

We employed two benchmarking analyses to compare the performance of the two prediction types used in this work against that of similar methods from other authors. One for the quantitative predictions (PSPs for the 30 proteins, and general quantitative Ensemble model); and the second, for the binarized Ensemble predictions.

To benchmark our quantitative models, we selected five *in silico* tools that predict the magnitudes of functional variant effects:

- ENVISION (Gray *et al.*, 2018). A supervised method that also uses DMS experiments for training purposes and predicts the quantitative effect of missense variants. ENVISION predictions for available proteins were downloaded directly from their website https://envision.gs.washington.edu/shiny/envision_new/
- ECNeT (Luo *et al.*, 2021). A deep-learning algorithm based on the use of homologous sequences that predicts functional fitness scores to aid protein engineering. In their study, they explore a split-wise split train/test strategy that is analogous to our LOPO approach. The ECNeT predictions for the whole proteome were not available, therefore we extracted their Spearman's rank correlations from their supplementary figure 10.c of Luo et al. (2021).
- DeepSequence (Riesselman, Ingraham and Marks, 2018). An unsupervised probabilistic method that uses a variational autoencoder model to identify patterns from multiple sequence alignments of homologous sequences for given protein. DeepSequence scores for the available proteins were downloaded from Supplementary Table 2 of Riesselman et al. (2018). No generalization to the whole proteome is provided.

- EVE (Frazer *et al.*, 2021). Developed by the same group of DeepSequence, is an unsupervised method that uses a variational autoencoder model to learn from multiple sequence alignments of a given protein. EVE scores for an exhaustive mutagenesis of the whole proteome were downloaded directly from their website <https://evemodel.org>
- EVMutation (Hopf *et al.*, 2017). An unsupervised statistical method that models co-variations between pairs of protein residues using the multiple sequence alignment of a given protein and predicts variant effects magnitudes using the log-ratio of sequence probabilities between mutant and wild-type. EVmutation (epistatis) predictions for available proteins were downloaded directly from their website <https://marks.hms.harvard.edu/evmutation/>

To evaluate our binary predictions, we selected four *in silico* tools, based on their clinical value in the recent work by Pejaver *et al.* (2022):

- REVEL (Ioannidis *et al.*, 2016). A metapredictor that predicts the pathogenicity of missense variants. It is a supervised method that is trained with RF using clinically labeled variants and predictions of several individual tools. The scores range from 0 to 1, with higher scores indicating greater likelihood of disease-causing.
- VEST4 (Carter *et al.*, 2013). Supervised method for pathogenicity prediction. It is based on a RF trained using clinically labelled variants with 86 quantitative features. The scores range between 0 and 1, where 1 indicates confident prediction of pathogenicity.
- BayesDel (Feng, 2017). A metapredictor that combines the predictions of several individual tools. It is a supervised method trained with naïve Bayesian approach to produce a weighed product

of likelihood ratio as a pathogenicity score. The scores range from -1.293 to 0.757, with higher scores indicating pathogenicity.

- CADD (Kircher *et al.*, 2014). A metapredictor that scores the deleteriousness of variants with supervised learning. It is based on SVM algorithm whose scores range from 1 to 99, with higher scores (> 30) indicating deleteriousness.

Predictions from these *in silico* tools were downloaded using Ensembl Variant Effect Predictor (VEP) tool from their website <https://www.ensembl.org/info/docs/tools/vep/index.html> (McLaren *et al.*, 2016).

5.2.7. Performance metrics

The predictive performances of the regression models were measured using two parameters: the Pearson correlation coefficient and the Spearman rank correlation coefficient (Winter, Gosling and Potter, 2016). Both Pearson and Spearman correlations were calculated using python scipy stats package (Virtanen *et al.*, 2020).

The performance of the binary predictions was measured using five parameters: sensitivity, specificity, accuracy, MCC, and AUC. These parameters are described in **Chapter 1**, section **1.4.3. Parameters for evaluating the performance.**

5.3. Results

Below, we describe the results of this study. They are divided into four main sections: (i) feature importance and their correlations; (ii) performance of the PSPs; (iii) generalization of the PSPs to the clinical genome with Ensemble model, and (iv) binarization of the Ensemble model.

5.3.1. Feature importance and correlations

We explored the cross-correlations between 14 features in our model (**Figure 5.9,a**). In general, the Pearson correlations were weak, between 0 to 0.55, indicating low linear relationships. However, a subset of features exhibited a stronger correlation between them. This was observed mostly between some structure- and sequence-related features: e.g. the fraction conservation 3D neighbor is highly correlated with neighbor average entropy (-0.78) and Shannon's entropy (-0.67). Overall, the low to moderate correlations between features highlighted the diversity of the feature landscape of the model.

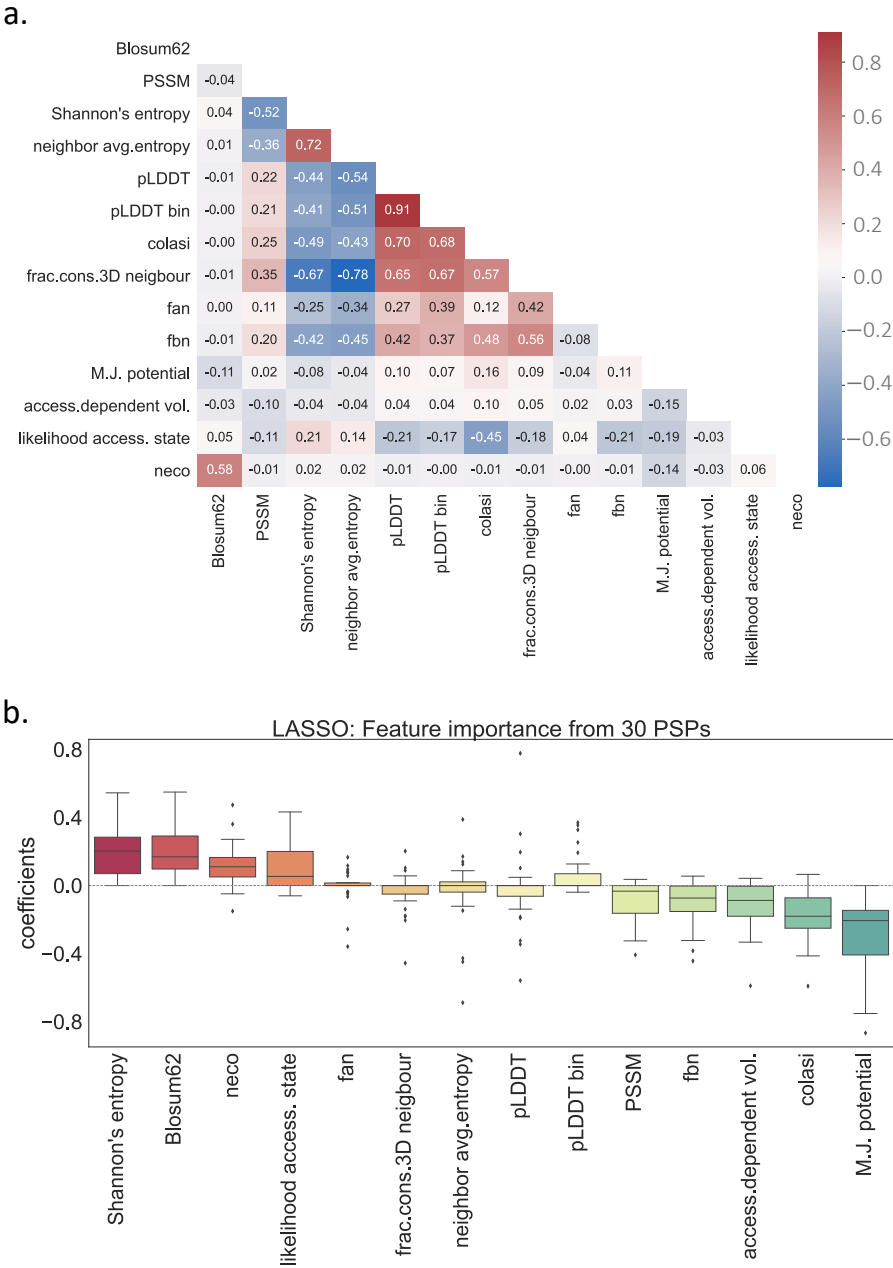


Figure 5.9: Feature importance analysis. (a) Pearson's correlation between features, from -1 (blue) to 1 (red). (b) Distributions of the coefficient values obtained from 30 protein-specific Lasso models, shown using boxplots.

In addition, we built Lasso regression models for each protein, to understand the relative importance of each feature's contribution to the models' predictive capacities. For each feature separately, we pooled the weights of the feature in the protein specific models and plotted their distributions using boxplots (**Figure 5.9,b**). We found that the key contributors (features with the highest coefficient values) were two sequence-related (Shannon's entropy, Blosum62) and two structure-related (Miyazawa and Jernigan potential and colasi) features. This shows that the predictive power of the models depends on both evolutionary conservation and structure-based features, highlighting the importance of integrating multiple aspects of variant information for accurate predictions.

5.3.2. Performance of the protein-specific predictors

We developed PSPs for 30 proteins separately. Their performances were validated using a stringent leave-one-position-out cross validation method. In **Figure 5.10** we show the Pearson correlation for all models, and heatmaps for the observed-predicted values for a selection of proteins. Heatmaps for all PSPs are shown in **Figure S5.2**.

5. Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors

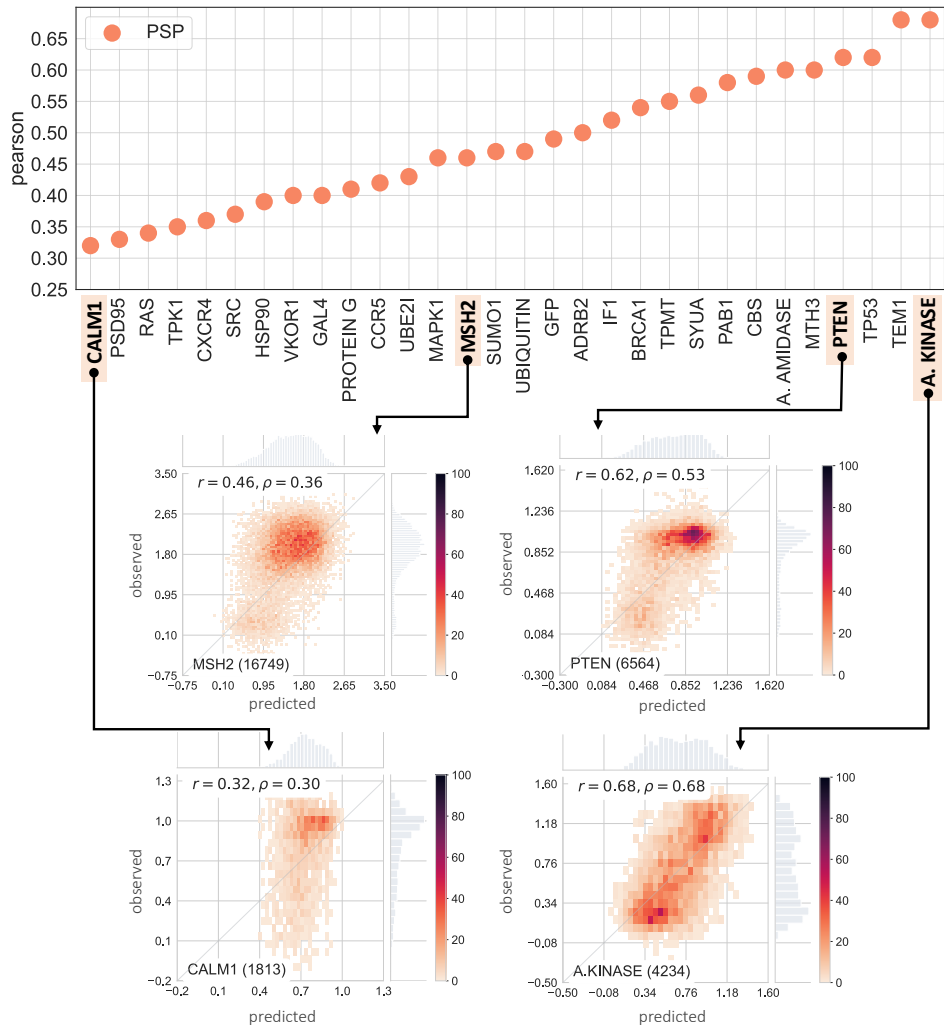


Figure 5 .10: Performance of the Protein-Specific-Predictors (PSP) for 30 proteins. The scatter plot shows the Pearson correlation coefficient values of each model. For four proteins (CALM1, MSH2, PTEN and AMINOGLYCOSIDE KINASE) the Observed vs. Predicted heatmap plots are shown with arrows.

The Pearson coefficients range from 0.32 (CALM1) to 0.68 (TEM1 and A. KINASE) confirming the predictive capabilities of our models. Obtaining Pearson's > 0.5 for 13 proteins in our dataset indicate that our features

can capture an important component of the molecular impact of missense variants. Additionally, the heatmap distributions show that, for those PSPs with higher Pearson's values, the majority of the data points lie close to the diagonal line, whereas in models with poorer performances we observe curvature or non-linear distribution. Exploration of the proteins with the poorest performances indicates deviations from the average behavior of globular proteins.

Next, we compared PSPs against ENVISION using the 20 proteins in our dataset for which ENVISION provided predictions (**Figure 5.11**).

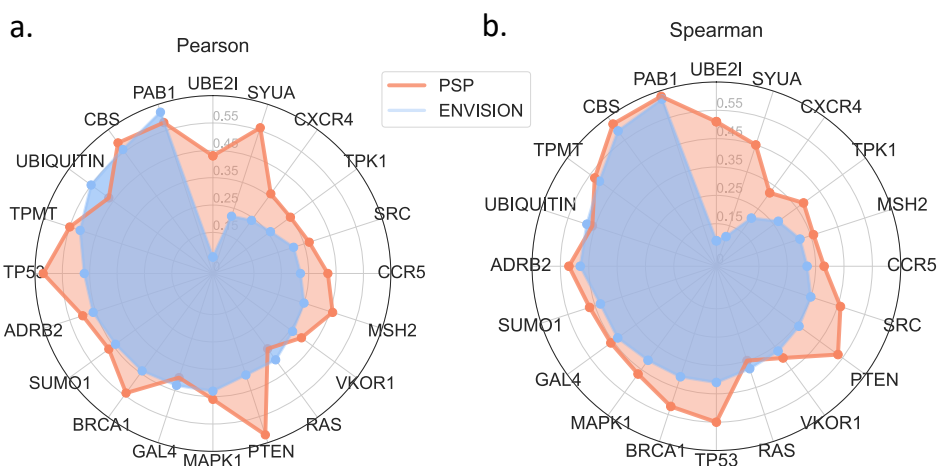


Figure 5.11: Comparison of PSPs (orange) against ENVISION (blue). Radar plots for (a) Pearson and (b) Spearman correlation coefficients.

Figure 5.11 shows that PSPs outperform ENVISION in 17 out of 20 cases for both Pearson and Spearman correlations. The trend for the Pearson correlations is particularly important because higher values of this correlation indicate more accurate predictions of the molecular impact predictions. Given that both predictors are built with similar goals, this comparison ensures a fair, comprehensive assessment of our PSPs.

5.3.3. Ensemble learning model: Extending protein-specific predictions to clinical genome

The Ensemble-based model is constructed for extending the use of our PSP models to proteins for which there are no DMS training data available. To explore the feasibility of this approach, we first tested the performance of each PSP on the remaining 29 proteins' variants; these are what we called cross-prediction tests. Overall, we observed (**Figure 5.12,a**) cross-prediction Pearsons up to 0.7. When compared these values to those obtained on the PSP own protein (**Figure 5.12,b**), we saw that they were comparable, except for the protein SYUA. The behavior of this protein suggests that the sequence/structure properties of this protein deviate from those of standard globular proteins, something confirmed when we look at its structure.

Collectively, these results showed that PSP models capture some of the main components of missense variant impacts, even in the presence of protein-specific biases, suggesting their potential utility in constructing a general predictor. After considering several options to put this information together from a relatively complex, algebra-based approach to the training of an additional layer of regressors, we opted for a simple, non-parametric approach: use 10 top-ranking PSPs to predict the impact of a variant in a new protein and take the median of these predictions as the outcome. To identify the 10 top-ranking PSPs and implement this Ensemble model, we used the correlations shown in **Figure 5.12,a**, within a cross-validated scheme (**Figure 5.7**).

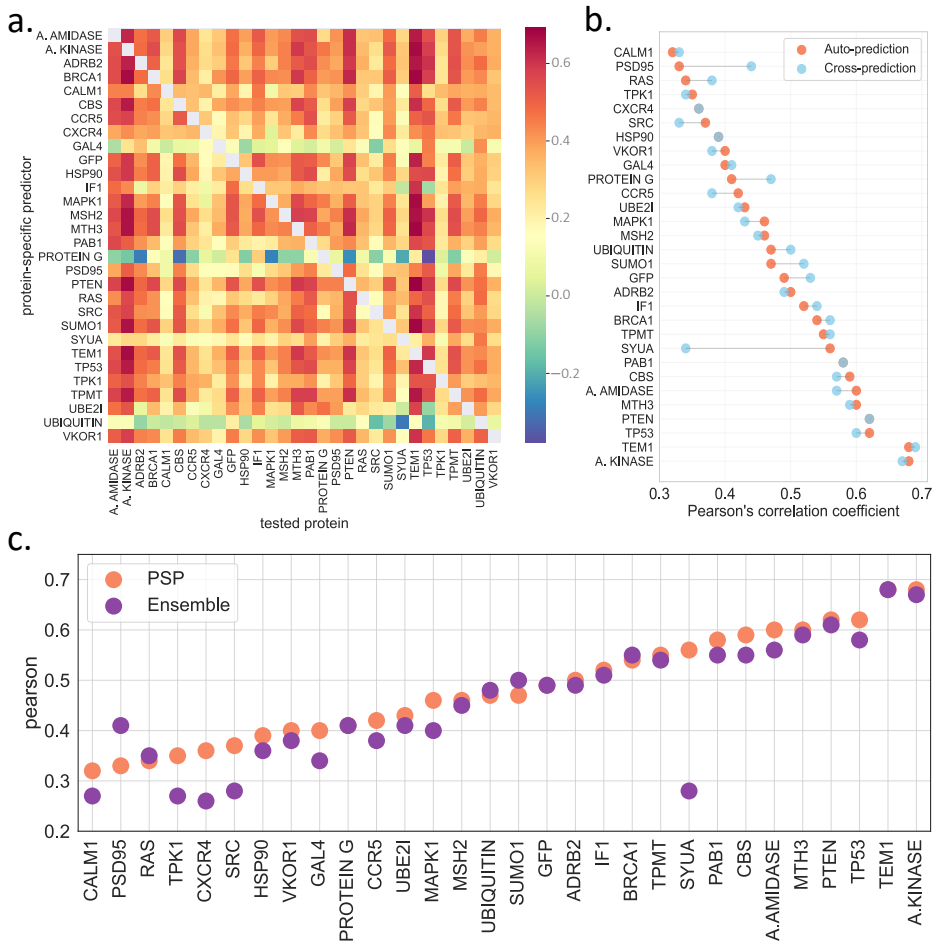


Figure 5.12: Cross predictions using the protein-specific-predictors. (a) Performance of the cross-predictions is shown representing Pearson's correlation coefficient with heatmaps, where x-axis is the tested protein and y axis is the trained protein. (b) Comparison of the PSP cross-validated performance (orange) and the performance of the highest prediction obtained by another PSP (cross-prediction) (c) Comparison of the PSP cross-validated performance (orange) and the Ensemble model built using the median of 10 PSPs (purple).

To explore the predictive value of this approach, we compared the performances obtained for the 30 proteins in our dataset against of those of the auto-predictions (or cross-validated PSP performances) (**Figure 5.12,c**). As expected, there was a performance drop in the Ensemble model when compared to the auto-predictive performances; however, the differences observed were not substantial, particularly when the PSP correlation was above 0.5, i.e., when the protein followed a globular protein model. The only clear exception was the case of SYUA.

We then benchmarked the quantitative Ensemble-based predictions against five predictors: two supervised methods (ENVISION and ECNet) also trained using DMS experiments, and three unsupervised models (DeepSequence, EVMutation and EVE). These predictors were selected for comparison given that none of them were trained using clinically labeled variants.

Comparisons of the Pearson correlations between Ensemble and four predictors (**Figure 5.13**) show that, Ensemble model outperforms ENVISION in 13 out of 20 cases (**Figure 5.13,a**); DeepSequence in 4 out of 19 cases (**Figure 5.13,b**); EVMutation in 8 out of 14 cases (**Figure 5.13,c**) and EVE in 6 out of 11 cases (**Figure 5.13,d**). Overall, these results show that our model, has a promising performance, except for its comparison to DeepSequence, given the compared top-ranking predictors.

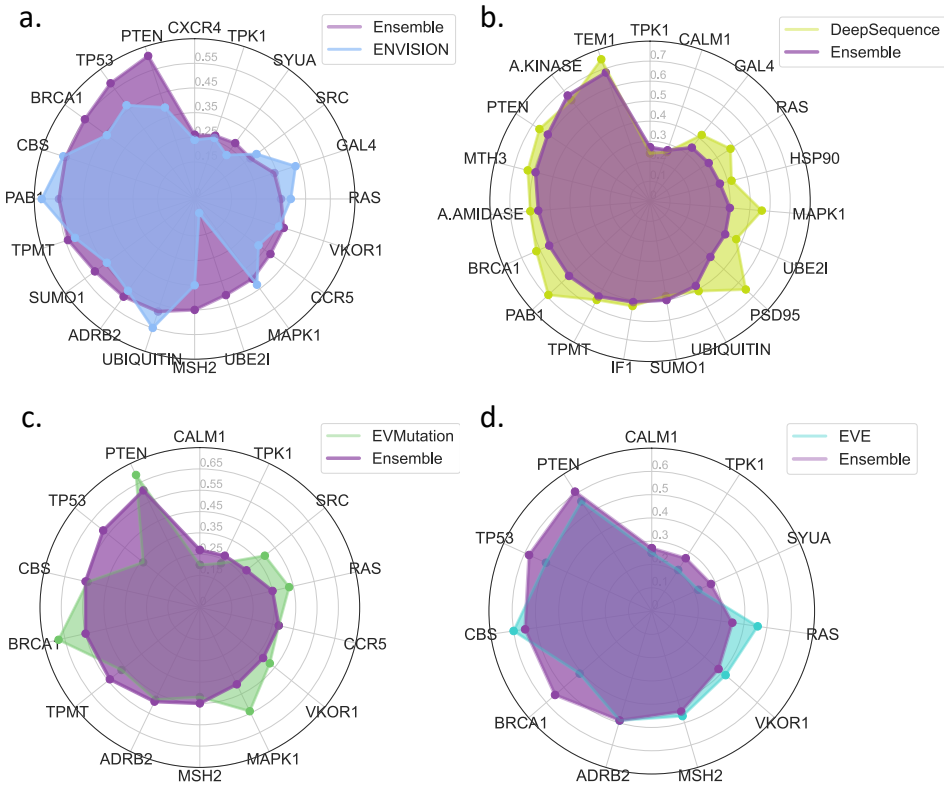


Figure 5.13: Comparison of our Ensemble-based model against four predictors using Pearson's correlations. (a) Comparison with ENVISION for 20 proteins; (b) comparison with DeepSequence for 19 proteins; (c) comparison with EVMutation for 14 proteins; and (d) comparison with EVE for 11 proteins.

A similar situation was observed when we compared the Spearman's rank correlations (**Figure 5.14**), where the differences between the predictor pairs compared were smaller, except for DeepSequence. Ensemble model outperforms ECNet in 7 out of 14 cases.

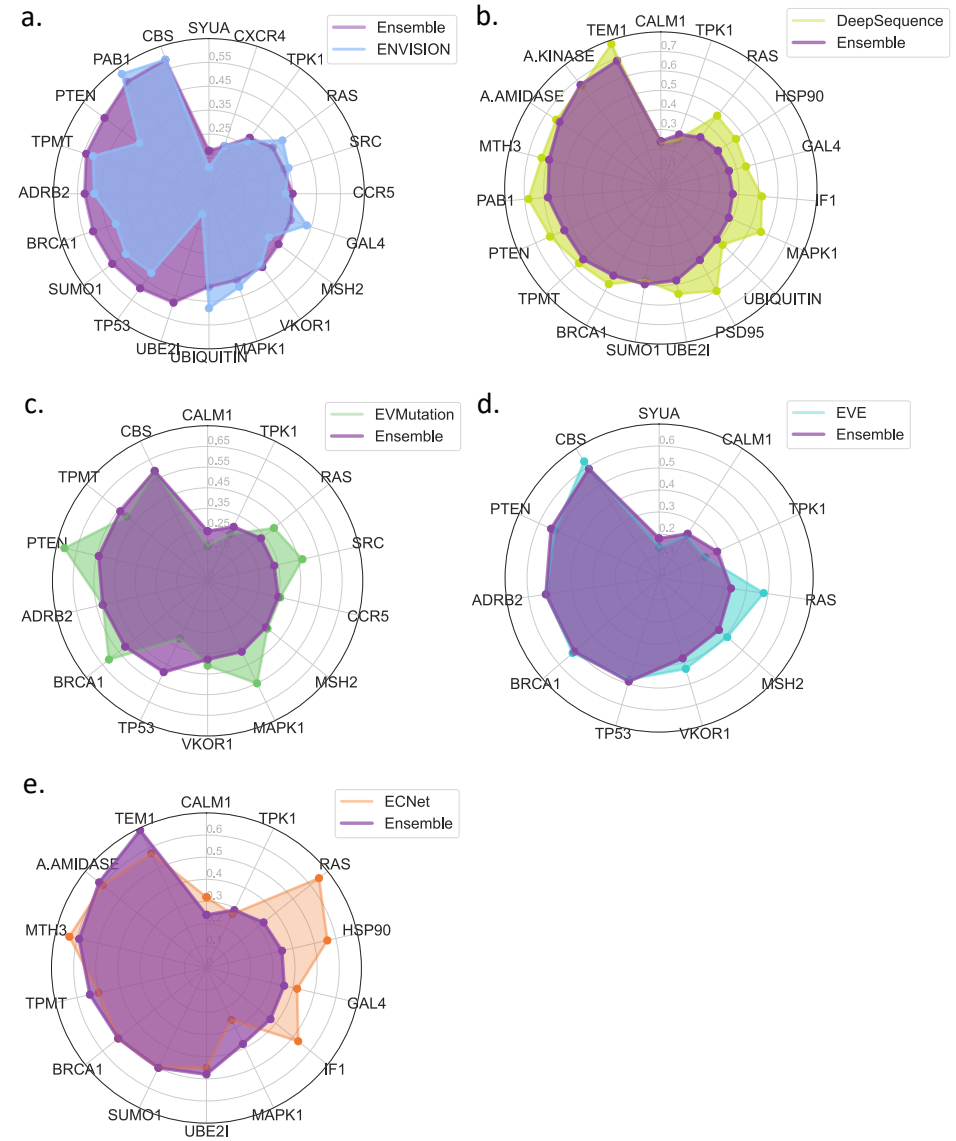
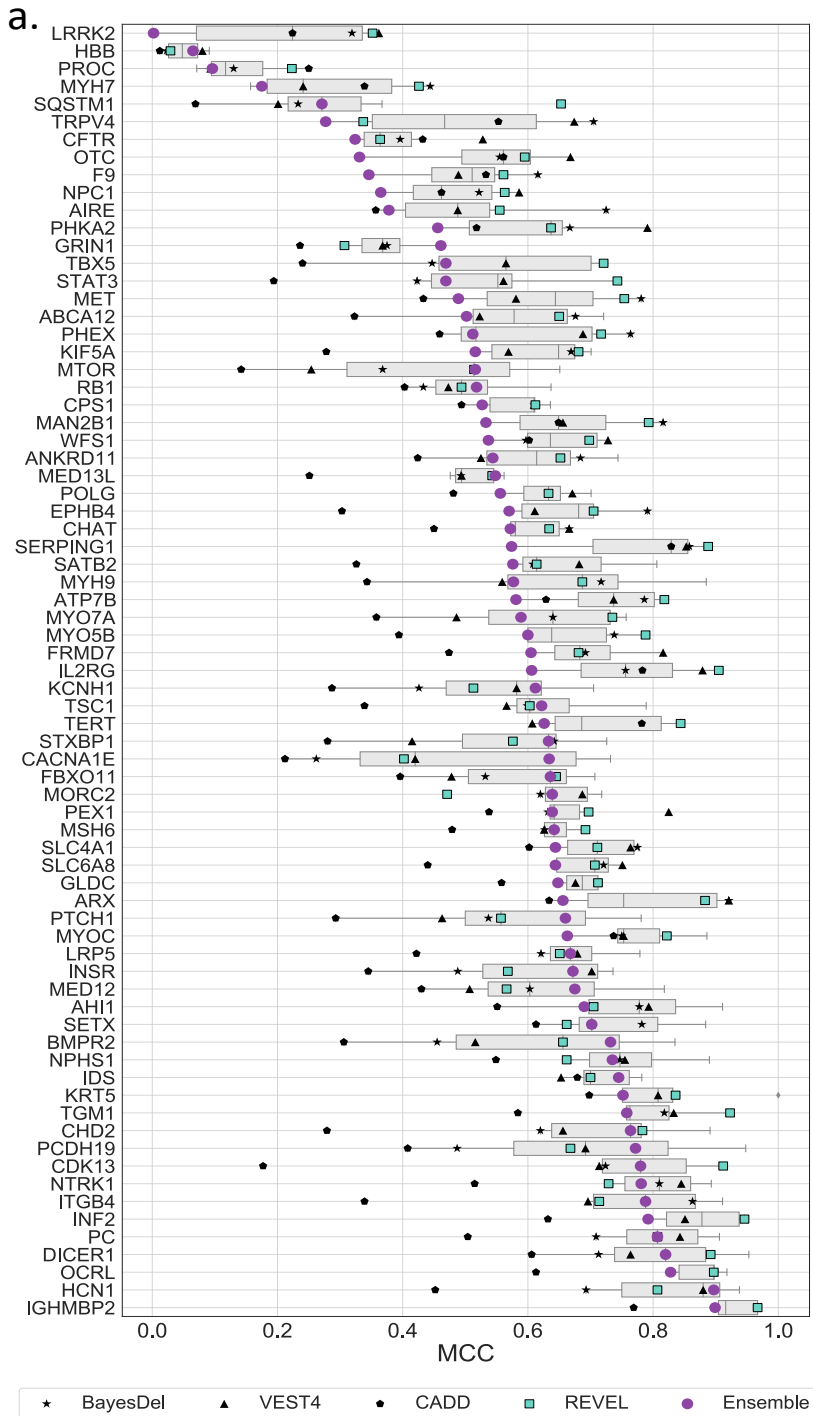


Figure 5.14: Comparison of the Ensemble-based model against four predictors using Spearman's rank correlations. (a) Comparison with ENVISION for 20 proteins; (b) comparison with DeepSequence for 19 proteins; (c) comparison with EVMutation for 14 proteins; (d) comparison with EVE for 11 proteins; and (e) comparison with ECNet for 8 proteins.

5.3.4. Binarization of the Ensemble model

The final phase of this project was binarization of the Ensemble-based model estimates for predicting the pathogenicity nature (pathogenic/benign) of protein variants. Following the scheme described in Material and Methods **5.2.4.2. Obtaining a threshold for binarizing quantitative predictions** and **Figure 5.8**, by using 26 proteins with more than 30 pathogenic and benign variants, we obtained a binarization threshold value of 0.84. Variants with an Ensemble estimate below and above 0.84 were labeled as pathogenic and benign, respectively.

To validate the predictive power of the resulting binary predictions, we compared their performance to those of a set of *in silico* tools: REVEL, VEST4, BayesDel, and CADD. These tools were chosen because they have been shown to provide strong level of evidence for pathogenicity (PP3) and moderate level benignity (BP4) criteria in the application of ACMG/AMP criteria (Pejaver *et al.*, 2022). For this comparison, we used the proteins from the clinically labeled dataset not used in the derivation of the binarization threshold: 82 proteins with 9330 variants (5229 pathogenic and 4101 benign). The performances are compared using four parameters (MCC, accuracy, sensitivity, and specificity), as shown in **Figure 15.5**.



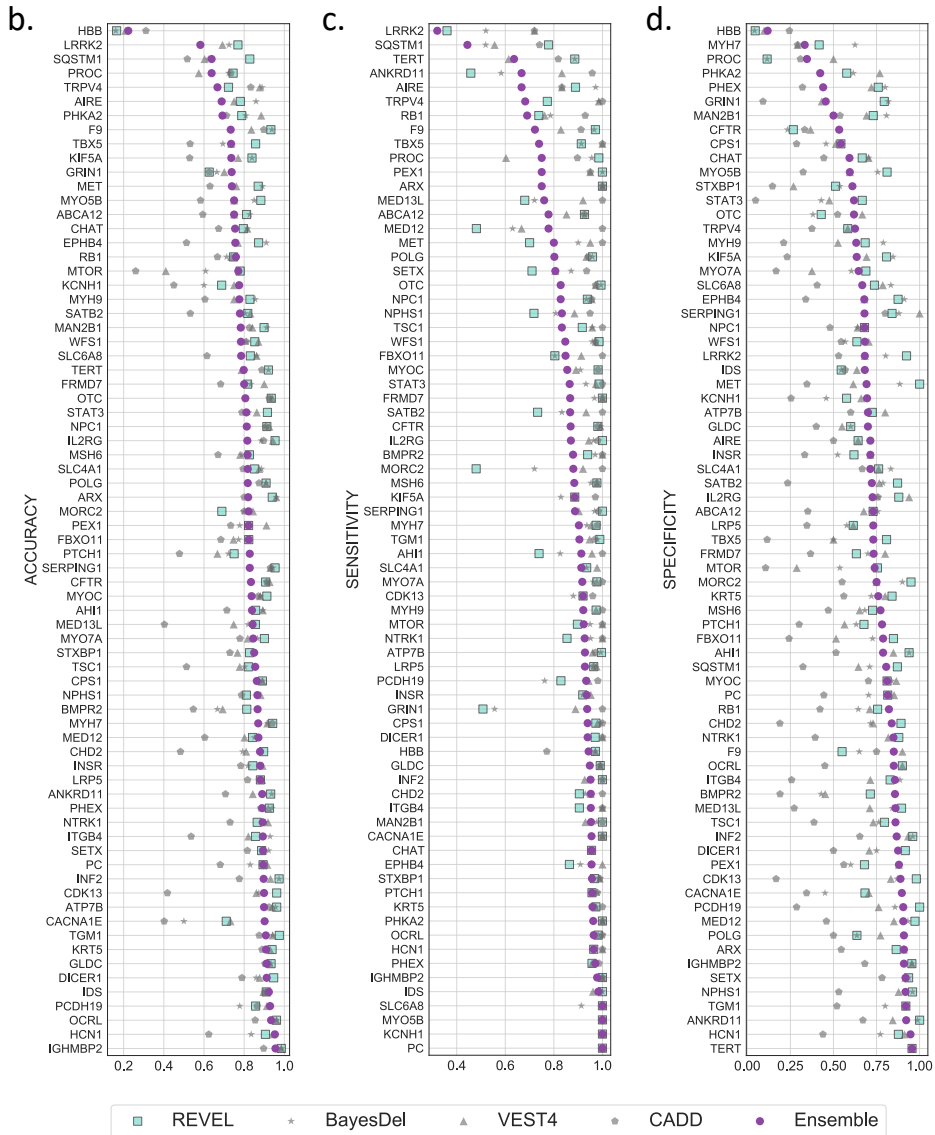


Figure 5.15: Comparison of the binary Ensemble model against REVEL, BayesDel, VEST4 and CADD. Comparison of the predictors on the missense variant dataset of 82 proteins. (a) MCC distributions (shown also with boxplots), (b) Accuracy, (c) Sensitivity and (d) Specificity.

Overall, we find that the MCC of our Ensemble-based model cover a broad range going from around 0.1 to 0.9. Among the four *in silico* tools considered, our model outperformed CADD for the high majority of cases (63 out of 82 cases) when we specifically focus on MCC rankings (**Figure 5.15,a**). On the other side, our model was outperformed by REVEL (57 out of 82 cases), VEST4 and BayesDel (54 out of 82 cases) for more than half of the protein predictions. In the case of accuracy comparisons (**Figure 15.5,b**), we observe that our model had accuracies higher than 0.6 for all proteins, except for HBB and LRRK2, and the differences between our model and compared predictors are very minor when the performance is above 0.8. In the case of sensitivity (**Figure 15.5,c**), we observe all models showing high predictive power, where our model appears poorer. Lastly, for specificity (**Figure 15.5,d**) results, we observe a slightly higher performance of our model compared to the other tools.

An overview of these results is also summarized in **Figure 5.16** using boxplots. Globally, we see that our Ensemble predictor has lower sensitivities than the other methods and comparable or somewhat higher specificities. This explains the fact that our MCCs tend to be below those of other methods. It also indicates that the procedure for finding the threshold may be at the origin of the problem.

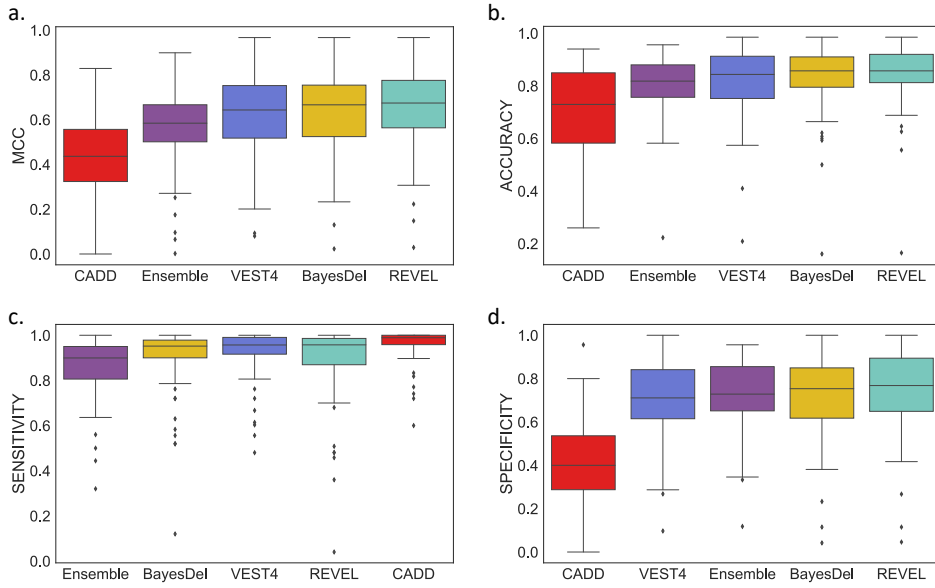


Figure 5.16: Per predictor comparison of the binary Ensemble model against REVEL, BayesDel, VEST4 and CADD. Comparison of the predictors on the missense variant dataset of 82 proteins. (a) MCC, (b) Accuracy, (c) Sensitivity and (d) Specificity.

Next, we evaluated the performance of our Ensemble-based model by constructing ROC curves by pooling all 9330 variants of the 82 proteins (**Figure 5.17**). The ROC curves showed that Ensemble model's predictions were better than random with the AUC equal to 0.87, indicating good overall performance. Like for other parameters, our model outperforms CADD (AUC equals to 0.85) but is outperformed by REVEL, VEST4 and BayesDel predictors (AUC equals to 0.92, 0.91 and 0.91, respectively).

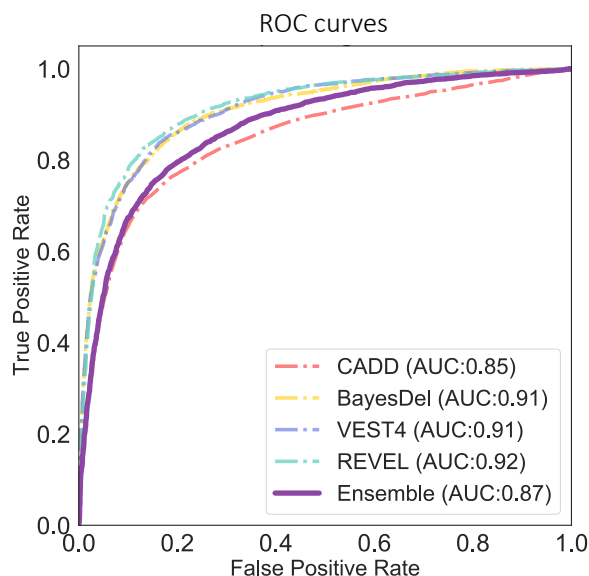


Figure 5.17: Receiver operating characteristic (ROC) curves comparison between the Ensemble-based model and four in silico tools.

Given that our Ensemble model approaches the prediction problem through training with DMS assay values, and no information of clinically labeled data was involved, the closeness of the ROC curves with the compared pathogenicity tools show that our models can capture the binary nature of the problem.

Lastly, we compared the continuous statistical scores of each pathogenicity predictor with the Ensemble-model predictions (**Figure 5.18**). The goal here was to have a quantitative evaluation to see the consistency between our Ensemble model's molecular impact predictions and the pathogenicity scores.

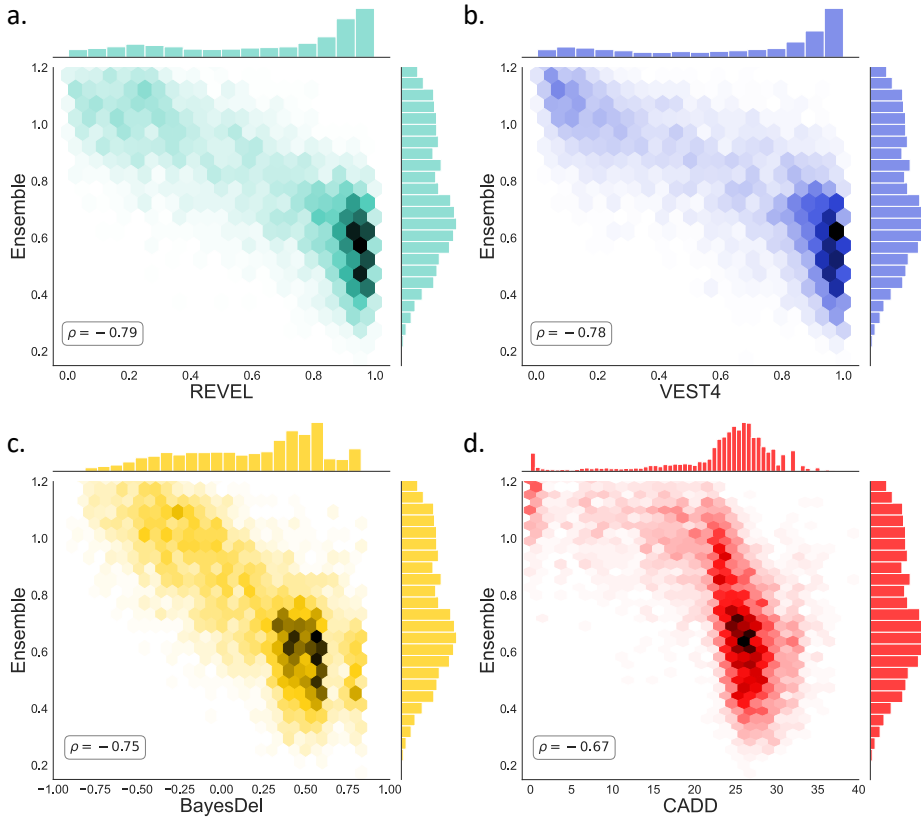


Figure 5.18: Correlation between Ensemble model and four in silico tools. Predictions of the Ensemble-based (y-axis) and the compared in silico tools (x-axis) are shown for (a) REVEL, (b) VEST4, (c) BayesDel and (d) CADD.

We observed that our model, despite not being trained on clinically labeled variant dataset, showed high alignment with the state-of-the-art tools' predictions. We observed the highest agreement with REVEL ($\rho = -0.79$) (Figure 5.18,a), followed by VEST4 ($\rho = -0.78$) (Figure 5.18,b), BayesDel ($\rho = -0.75$) (Figure 5.18,c) and lowest correlation with CADD ($\rho = -0.67$) (Figure 5.18,d).

5.4. Discussion

Currently, available pathogenicity predictors cannot be used as standalone tools in the clinical setting (Bean and Hegde, 2017; Sun and Yu, 2019). In this context, the availability of an increasing number of massive DMS experiments suggests that ML models could use this information to produce an alternative solution to the variant interpretation problem. For instance, they could be used in filling the missing information in these experiments (Weile *et al.*, 2017; Wu *et al.*, 2019). Or, they can be used in aiding the selection of representative variants for their experimental validation, while predicting the outcomes for the remaining variants with computational tools (Sruthi and Prakash, 2020).

However, the quantitative estimation of the impact of missense variants on protein function is a problem far from trivial. For example, a recent study by Livesey *et al.* (2020) indicates that binary classifiers display better Spearman correlations with the values of DMS experiments than methods directly trained on these experiments. An immediate consequence of this paradoxical result is that the standard approach to the modeling of continuous data, based on the use of regression models and biochemical/biophysical/conservation features, does not seem to be competitive. And second, it suggests that metapredictor models, widely used for binary classifiers, may work better at fitting functional assays, a challenging result given the low interpretability of these models.

In our work, we have explored these issues in depth, with a clear overall goal: to build a method that starts from the quantitative estimation of a variant's functional impact and then turns it into a binary prediction (pathogenic/benign) for that variant. To this end, we have decomposed the prediction problem into two parts (**Figure 5.1**). In the first part, we

have focused on (i) improving the quantitative prediction of missense variants' impact introducing novel features, particularly by considering state-of-the-art 3D information; and (ii) extending the resulting methodology to the clinical genome, using Ensemble Learning theory. In the second part, we have focused on the binarization of the quantitative predictions, testing the results on an independent set of variants with clinical annotations, retrieved from UniProt and ClinVar databases.

For the first part, the quantitative prediction problem, our starting point was the use of PSP (**Figure 5.5**) because they have gained popularity over the years. In fact, they show competitive performances against all-genes pathogenicity predictors, outperforming them in many cases (Masica *et al.*, 2012; Adebali *et al.*, 2016; Riera, Padilla and de la Cruz, 2016; Padilla *et al.*, 2019; Aljarf *et al.*, 2022; Fang *et al.*, 2022), due to their ability to account for gene-specific properties.

Our PSPs were developed combining sequence- and structure-based features (**Figure 5.3**). We observed that the combination of sequence and structure-based features in our models enabled them to capture complementary aspects of the variant's impact. Sequence-based features reflect different aspects of evolutionary conservation, which are widely used in many ML models (Niroula and Vihinen, 2016; Tang and Thomas, 2016). Resulting from a complex combination of organismal population dynamics and molecular biophysics, conservation works as a proxy for functional and structural properties. Indeed, it has been observed that variants at evolutionarily conserved sites tend to disrupt protein's function, and similarly, those substitutions observed between humans and closely related species tend to be benign (Sunyaev, Ramensky and Bork, 2000; Miller and Kumar, 2001; Wang and Moul, 2001). Therefore, it is not surprising to see that our strongest predictive

features include entropy and position-specific scoring matrix (**Figure 5.9,b**). Regarding structure-based properties, the feature importance analysis showed that two of these properties, accessibility (colasi) and the Miyazawa and Jernigan potential, have high influence on the model's learning (**Figure 5.9,b**). These two properties, which are computed using 3D information from AlphaFold models (Jumper et al., 2021), provide a coarse-grained view of the protein energetics.

The performance of the PSPs showed that, we were able to predict the DMS data with high accuracies for many proteins in our dataset, with Pearsons higher than 0.5 for more than half of the cases (**Figure 5.10**). Comparisons with ENVISION, a tool with a similar technical goal, showed our models superior performance for predicting the DMS values more efficiently. A factor impacting the proteins with poor performances could be explained by their differences than average globular proteins, with structural transitions and protein disorder playing a substantial role. For instance, CALM1, the protein with the poorest performance, is involved in calcium signaling and has a dynamic structure (Jensen *et al.*, 2018). PSD95, the protein with the second lowest ranking performance, has an intricate network of protein-protein interactions (McLaughlin *et al.*, 2012). In this case, missense variants occurring in crucial binding interfaces could have a greater impact on its function, making prediction more challenging. On the other hand, the poor performances we obtained for some of the proteins in our study have also shown similar behavior for other predictors. For example, in the recent benchmarking study of Livesey et al. (2023), DMS experiments of many proteins including CALM1, TPK1, CXCR4 showed low correlations with 55 *in silico* tools, with Spearman's rank correlations approximately around 0.3. This is consistent with our PSP performances for these proteins, pointing to directions for improving our models, e.g., proteins deviating from a strict

globular protein paradigm may require the development of specific predictors with more unique characterization.

In summary, we found that our PSPs provided a promising starting point to the variant interpretation problem. However, two problems remained open at this stage: (i) how to extend the results to proteins for which no DMS experiments are available; and (ii) how to use quantitative estimates for the binary classification of variants.

Therefore, as our next step, we extended our models to the clinical genome by following an Ensemble Learning approach. Ensemble methods have emerged as promising solutions to address the challenges and limitations associated with individual machine learning models, offering a state-of-the-art performance (Sagi and Rokach, 2018; Cao *et al.*, 2020). The underlying principle of ensemble methods lies in the recognition that individual models capture distinct aspects of the problem at hand and, by combining their predictions, we can produce a more informed decision. In our study, we chose the median approach to integrate PSP predictions due to its technical simplicity and its ability to mitigate the influence of noisy predictions from individual PSP models.

Comparison of the quantitative predictions obtained from Ensemble model against four predictors (ENVISION, DeepSequence, EVMutation and EVE) (**Figure 5.13**) showed that our model has a promising performance against that of top-ranking predictors in the field. In the study of Livesey *et al.* (2023), EVE and DeepSequence were ranked first overall, out of 55 *in silico* tools used for benchmarking with various DMS studies. In our case, the Ensemble predictor outperformed both tools for various proteins and was very close for the cases it was outperformed. However, given the complexity of DeepSequence method, it is unclear to

which extent the model captures some correlation relating residue location and protein function. We prioritized comparisons of Pearson correlations because our model was trained to predict the DMS experimental values, not the ranking of mutations. When we focus on the comparison between Ensemble and ENVISION, our model clearly outperforms ENVISION for most of the cases.

It is important to mention that both the quality and the type of DMS experiment might present potential pitfalls to our models, affecting the training process. Given that fitness is measured differently depending on the experimental set up of the DMS study (e.g. growth, abundance, binding affinity, fluorescence etc.), this could potentially affect how the molecular impact is represented in terms of numerical values. Although our median strategy aims to minimize or handle this effect partially by combining multiple predictions, our future work also includes a rigorous selection of DMS experiments to improve the effectiveness and representativeness of our models.

Finally, we compared the binarized predictions of the quantitative estimates against a set of *in silico* tools that are identified by Pejaver et al. (2022) as highly promising for clinical applications. The results show that, our model was outperformed by REVEL, VEST and BayesDel for several cases, although when we pool the individual predictions together, the median performances were highly close to each other (**Figure 5.16**, **Figure 5.17**). However, it is worth noting that the variant dataset we used contains variants utilized for training of the four reference tools, therefore that the performance of these tools may be somewhat overoptimistic. High sensitivities and poorer specificities observed in compared *in silico* tools were consistent with findings in other studies (de la Campa, Padilla and de la Cruz, 2017; Cubuk *et al.*, 2021), showing that the author

provided thresholds tend to maintain high sensitivities. In our case, the Ensemble model's showed higher ranking in specificities and lower in sensitivities, which also points out our model's limitation regarding the binarization criteria. Consequently, we are exploring an alternative approach that combines sensitivity and specificity values as a reference, rather than solely relying on the MCC, to address this challenge more effectively. Lastly, exploring the relationship between our Ensemble model's predictions and those of the established methods (**Figure 5.18**) showed high concordance between the two approaches, indicating that our model's predictions are in high agreement regarding the variant pathogenicity prediction.

The predictive performance of our Ensemble model is significant given that it approaches to the prediction problem from a completely different angle than its competitors: our training set was the DMS experiment values rather than clinically labeled variants. Despite of the difficulties found, at its current status, our model ranks at a high position when compared against state-of-the-art tools. We expect that better versions of our approach will benefit from improvements on the binarization strategy, from the publication of new DMS assays, etc.

5.5. Conclusion

Our results show that we can produce fundamental approaches, based on the modeling of the functional impact of variants, to solve the pathogenicity prediction problem. Although the performance of our method is not yet complete it is almost as good as that of *in silico* tools specifically trained to solve this problem. This result opens the way to solve one of the main challenges of personalized medicine in the 21st century (Shendure, Findlay and Snyder, 2019): the development of back-to-basics models that solve the variant interpretation problem.

5.6. Supporting Information

List of material in the supporting information:

- **Table S5.1:** List of deep mutational scanning experiment studies collected, description of the datasets, and our selection criteria.
- **Table S5.2:** The clinically labeled variant dataset.
- **Figure S5.1:** Distribution of the DMS experiments for 30 proteins with histograms, before (PRE) and after (POST) normalization.
- **Figure S5.2:** Leave-one-position-out performance of 30 protein-specific predictors.

Table S5.1: List of deep mutational scanning experiment studies collected, description of the datasets, and our selection criteria.

SOURCE STUDY	REFERENCE	PMID	UNIPROT	PROTEIN IDENTIFIER	PROTEIN	ORGANISM	# OF UNFILTERED VARIANTS RETRIEVED FROM THE REFERENCE STUDY	SOURCE OF ASSAY INFORMATION	SELECTION	REASON FOR NO SELECTION
Gray et al. 2018	McLaughlin et al. 2012	23041932	P31016	PSD95	PSD95	<i>A. norvegicus</i>	1577	Supp. Table 1 from Gray et al.	yes	
	Melamed et al. 2015	25671604	P04147	PAB1	PAB1	<i>S. cerevisiae</i>	1188	Supp. Table 1 from Gray et al.	yes	
	Melnikov et al. 2014	24914046	P00552	Aminoglycosid e kinase	Aminoglycoside kinase	<i>K. pneumoniae</i>	5300	Supp. Table 1 from Gray et al.	yes	
	Mishra et al. 2016	27068472	P02829	HSP90	HSP90	<i>S. cerevisiae</i>	4021	Supp. Table 1 from Gray et al.	yes	
	Olson et al. 2014	25455030	P06654	Protein G	Protein G	<i>Streptococcus sp. group G</i>	1045	Supp. Table 1 from Gray et al.	yes	
	Roscoe et al. 2014	24862281	P0CG63	Ubiquitin	Ubiquitin	<i>S. cerevisiae</i>	1085	Supp. Table 1 from Gray et al.	yes	
	Frimberg et al. 2014	24567513	P62593	TEM1 β-lactamase	TEM1 β-lactamase	<i>E. coli</i>	-	Supp. Table 1 from Gray et al.	no	Study with higher coverage selected
	Fowler et al. 2010	20711194	P46937	YAP65	YAP65	<i>H. sapiens</i>	-	Supp. Table 1 from Gray et al.	no	Less than 1000 variants
	Roscoe et al. 2013	23376099	P0CG63	Ubiquitin	Ubiquitin	<i>S. cerevisiae</i>	-	Supp. Table 1 from Gray et al.	no	More recent assay selected
	Starita et al. 2013	23509263	Q8E500	UBE4B	UBE4B	<i>M. musculus</i>	-	Supp. Table 1 from Gray et al.	no	Less than 1000 variants
	Starita et al. 2015	25823446	P38398	BRCA1	BRCA1	<i>H. sapiens</i>	-	Supp. Table 1 from Gray et al.	no	More recent assay selected
	Starita et al. 2015	25823446	P38398	BRCA1	BRCA1	<i>H. sapiens</i>	-	Supp. Table 1 from Gray et al.	no	More recent assay selected
Reeb et al. 2020	Brenan et al. 2016	27760319	P28482	MAPK1	Mitogen-activated protein kinase 1	<i>H. sapiens</i>	6810	Table S1 from Reeb et al.	yes	
	Findlay et al. 2018	30209399	P38398	BRCA1	Breast cancer type 1 susceptibility protein	<i>H. sapiens</i>	1837	Table S1 from Reeb et al.	yes	
	Heredia et al. 2018	29678950	P51681	CDR5	C-C chemokine receptor type 5	<i>H. sapiens</i>	4668	Table S1 from Reeb et al.	yes	
	Heredia et al. 2018	29678950	P61073	CXCR4	C-X-C chemokine receptor type 4	<i>H. sapiens</i>	3877	Table S1 from Reeb et al.	yes	
	Kitzman et al. 2015	25559584	P04386	GAL4	Regulatory protein GAL4	<i>S. cerevisiae</i>	1196	Table S1 from Reeb et al.	yes	
	Mattreyek et al. 2018	29785012	P51580	TPMT	Thiopurine S-methyltransferase	<i>H. sapiens</i>	3689	Table S1 from Reeb et al.	yes	
	Rochah-Shmuel et al. 2015	26274323	P20589	MTIH3 (haellim)	Modification methylase HaeIII	<i>H. aegyptius</i>	1957	Table S1 from Reeb et al.	yes	
	Sarkisyan et al. 2016	27193686	P42212	GFP	Green fluorescent protein	<i>Ae. victoria</i>	1084	Table S1 from Reeb et al.	yes	
	Stiffler et al. 2015	25723163	P62593	TEM1 β-lactamase	TEM1 β-lactamase	<i>E. coli</i>	4997	Table S1 from Reeb et al.	yes	
	Adkar et al. 2012	22325784	P62554	ccdB	Toxin CcdB	<i>E. coli</i>	-	Table S1 from Reeb et al.	no	Assay values are not continuous, they are ranks
	Araya et al. 2012	23035249	P46937	YAP65	Transcriptional coactivator YAP1	<i>H. sapiens</i>	-	Table S1 from Reeb et al.	no	Less than 1000 variants
	Kiepas et al. 2011	21464309	P02829	HSP90	ATP-dependent molecular chaperone HSP82	<i>S. cerevisiae</i>	-	Table S1 from Reeb et al.	no	More recent assay selected

continued

5. Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors

SOURCE STUDY	REFERENCE	PMID	UNIPROT IDENTIFIER	PROTEIN	ORGANISM	# OF UNFILTERED VARIANTS RETRIEVED FROM THE REFERENCE STUDY	SOURCE OF ASSAY INFORMATION	SELECTION	REASON FOR NO SELECTION
Reeb et al. 2020	Hietpas et al. 2013	24239404	HSF90	ATP-dependent molecular chaperone HSP82	<i>S. cerevisiae</i>	-	Table S1 from Reeb et al.	no	More recent assay selected
	Jiang et al. 2013	23825969	HSF90	ATP-dependent molecular chaperone HSP82	<i>S. cerevisiae</i>	-	Table S1 from Reeb et al.	no	More recent assay selected
	Klesmith et al. 2015	26369947	LGK	Levodopa decarboxylase	<i>L. starkeyi</i>	-	Table S1 from Reeb et al.	no	No AlphaFold structure
	Majitha et al. 2016	27749844	PPARG	Peroxisome proliferator-activated receptor gamma	<i>H. sapiens</i>	-	Table S1 from Reeb et al.	no	The origin of assay scores are not experimental
	Matreyek et al. 2018	29785012	PTEN	Phosphatase and tensin homolog	<i>H. sapiens</i>	-	Table S1 from Reeb et al.	no	Study with higher coverage selected
	Romero et al. 2015	26040002	Beta-glucosidase	Beta-glucosidase	<i>Streptomyces</i>	-	Table S1 from Reeb et al.	no	No AlphaFold structure
	Trakimayr et al. 2012	22846908	PDGFR	Immunoglobulin gamma-1 heavy chain	<i>H. sapiens</i>	-	Table S1 from Reeb et al.	no	Less than 1000 variants
	Bandaru et al. 2017	28686159	Hras	Hras	<i>H. Sapiens</i>	3134	Suppl Table 1 from Riesselman et al.	yes	
	Kelsic et al. 2016	28009265	IF1	Translation initiation factor IF-1	<i>E. coli</i>	1440	Suppl Table 1 from Riesselman et al.	yes	
	Weile et al. 2017	29269382	CALM1	Calmodulin-1	<i>H. sapiens</i>	2831	Suppl Table 1 from Riesselman et al.	yes	
Riesselman et al. 2017	Weile et al. 2017	29269382	TPK1	Thiamin pyrophosphokinase 1	<i>H. sapiens</i>	4859	Suppl Table 1 from Riesselman et al.	yes	
	Weile et al. 2017	29269382	SUMO1	Small ubiquitin-related modifier 1	<i>H. sapiens</i>	2019	Suppl Table 1 from Riesselman et al.	yes	
	Weile et al. 2017	29269382	UBE2I	SUMO-conjugating enzyme UBC9	<i>H. sapiens</i>	3002	Suppl Table 1 from Riesselman et al.	yes	
	Wrenbeck et al. 2017	28585537	Alphatic amidase	Alphatic amidase	<i>Paeruginosa</i>	6757	Suppl Table 1 from Riesselman et al.	yes	
	Aakre et al. 2015	26478181	Toxin-antitoxin complex	Toxin-antitoxin complex	<i>M. opportunistum</i>	-	Suppl Table 1 from Riesselman et al.	no	Low coverage
	Chan et al. 2017	28262665	TIM Barrel	TIM Barrel	<i>T. thermophilus</i>	-	Suppl Table 1 from Riesselman et al.	no	No AlphaFold structure
	Chan et al. 2017	28262665	Q06121	TIM Barrel	<i>S. sapientiae</i>	-	Suppl Table 1 from Riesselman et al.	no	Low coverage
	Chan et al. 2017	28262665	Q56319	TIM Barrel	<i>T. maritima</i>	-	Suppl Table 1 from Riesselman et al.	no	Low coverage
	Deng et al. 2012	23017428	TEM1 beta-lactamase	beta-lactamase	<i>E. coli</i>	-	Suppl Table 1 from Riesselman et al.	no	Study with higher coverage selected
	Doud et al. 2016	27271655	Hemagglutinin	Influenza hemagglutinin	<i>Influenza virus</i>	-	Suppl Table 1 from Riesselman et al.	no	No AlphaFold structure
continued	Haddox et al. 2018	29590010	HIV env protein (BFS20)	HIV env protein (BFS20)	<i>Human immunodeficiency virus 1</i>	-	Suppl Table 1 from Riesselman et al.	no	No AlphaFold structure
	Haddox et al. 2018	29590010	HIV env protein (BG505)	HIV env protein (BG505)	<i>Human immunodeficiency virus 1</i>	-	Suppl Table 1 from Riesselman et al.	no	No AlphaFold structure
	Jacquier et al. 2013	23878237	TEM1 beta-lactamase	beta-lactamase	<i>E. coli</i>	-	Suppl Table 1 from Riesselman et al.	no	Study with higher coverage selected
	Li et al. 2016	27080104	Yeast tRNA (CCU, Arg)	Yeast tRNA (CCU, Arg)	<i>S. Cerevisiae</i>	-	Suppl Table 1 from Riesselman et al.	no	No AlphaFold structure

SOURCE STUDY	REFERENCE	PMID	UNIPROT	PROTEIN IDENTIFIER	PROTEIN	ORGANISM	# OF UNFILTERED VARIANTS RETRIEVED FROM THE REFERENCE STUDY	SOURCE OF ASSAY INFORMATION	SELECTION	REASON FOR NO SELECTION
Riesselman et al. 2017	Mavor et al. 2016	27111525	P0C409	Ubiquitin	Ubiquitin	<i>S. cerevisiae</i>	-	Suppl. Table 1 from Riesselman et al.	no	Roscoe et al. 2014 growth assay is selected
	Pekusueva et al. 2019	30969563	P06633	IGPD	Imidazoleglycerol-phosphate dehydratase	<i>S. cerevisiae</i>	-	Suppl. Table 1 from Riesselman et al.	no	The origin of assay scores are not experimental
	Qi et al. 2014	24722365	Q99188	N55A	Hepatitis C NS5A	<i>Hepatitis C virus</i>	-	Suppl. Table 1 from Riesselman et al.	no	No AlphaFold structure
	Wu et al. 2015	26132554	P15659	Polymerase acidic protein	Influenza polymerase PA subunit	<i>Influenza virus</i>	-	Suppl. Table 1 from Riesselman et al.	no	No AlphaFold structure
	Ahler et al. 2019	30956043	P12931	p60-Src	Proto-oncogene tyrosine-protein kinase Src	<i>H. sapiens</i>	3714	Dataset EV1 from Dunham et al.	yes	
Dunham et al. 2020	Giacomelli et al. 2018	30224644	P04637	TP53	Cellular tumor antigen p53	<i>H. sapiens</i>	7856	Dataset EV1 from Dunham et al.	yes	
	Sun et al. 2020	32000841	P35520	CBS	Cystathionine beta-synthase	<i>H. sapiens</i>	10450	Dataset EV1 from Dunham et al.	yes	
	Askenberg et al. 2017	28346537	I67AH8	NP	Nucleoprotein	<i>Influenza virus</i>	-	Dataset EV1 from Dunham et al.	no	No AlphaFold structure
	Bolognesi et al. 2019	31519910	Q13148	TDPA3	TAR DNA-binding protein 43	<i>H. sapiens</i>	-	Dataset EV1 from Dunham et al.	no	Less than 1000 variants
	Dorrity et al. 2018	30068600	P13574	STE12	Protein STE12	<i>S. cerevisiae</i>	-	Dataset EV1 from Dunham et al.	no	Less than 1000 variants
	Doud et al. 2015	26226986	I67AH8	NP	Nucleoprotein	<i>Influenza virus</i>	-	Dataset EV1 from Dunham et al.	no	No AlphaFold structure
	Findlay et al. 2014	25141179	Q9UK59	DBR1	Lariat debranching enzyme	<i>H. sapiens</i>	-	Dataset EV1 from Dunham et al.	no	Less than 1000 variants
	Hartman et al. 2018	29643335	P03612	CP	Capsid protein	<i>Bacteriophage MS2</i>	-	Dataset EV1 from Dunham et al.	no	No AlphaFold structure
	Lee et al. 2018	30104379	P03437	Hemagglutinin	Hemagglutinin	<i>Influenza virus</i>	-	Dataset EV1 from Dunham et al.	no	No AlphaFold structure
	Steinberg et al. 2016	27173379	Q65161	TEM1 β-lactamase	Beta-lactamase	<i>E. coli</i>	-	Dataset EV1 from Dunham et al.	no	Study with higher coverage selected
Frazer et al. 2021	Wagenaar et al. 2014	24112705	P15056	BRAF	Serine/threonine-protein kinase B-raf	<i>H. sapiens</i>	73	Dataset EV1 from Dunham et al.	no	Less than 1000 variants
	Chasson et al. 2020	32870157	Q9BQB6	VKOR1	Vitamin K epoxide reductase complex subunit 1	<i>H. sapiens</i>	2822	Suppl. Table 4 from Frazer et al.	yes	
	Jia et al. 2021	33357406	P43246	MSH2	DNA mismatch repair protein Msh2	<i>H. sapiens</i>	16749	Suppl. Table 4 from Frazer et al.	yes	
	Jones et al. 2020	33084570	P07550	ADRB2	Beta-2 adrenergic receptor	<i>H. sapiens</i>	7800	Suppl. Table 4 from Frazer et al.	yes	
	Migheli et al. 2018	29706350	P60484	PTEN	Phosphatase and tensin homolog	<i>H. sapiens</i>	7657	Suppl. Table 4 from Frazer et al.	yes	
	Neuberger et al. 2020	32152544	P37840	SNCA/SYUA	Alpha-synuclein	<i>H. sapiens</i>	2739	Suppl. Table 4 from Frazer et al.	yes	
	Gray et al. 2019	31558564	P05067	Anyloid beta precursor	Amyloid-beta precursor protein	<i>H. sapiens</i>	-	Suppl. Table 4 from Frazer et al.	no	Less than 1000 variants
	Kotler et al. 2018	29979965	P04637	TP53	Cellular tumor antigen p53	<i>H. sapiens</i>	-	Suppl. Table 4 from Frazer et al.	no	Study with higher coverage selected
	Penn et al. 2020	32181350	P08100	Rhodopsin	Rhodopsin	<i>H. sapiens</i>	-	Suppl. Table 4 from Frazer et al.	no	Less than 1000 variants
	Seuma et al. 2021	33522485	P05067	Anyloid beta precursor	Amyloid-beta precursor protein	<i>H. sapiens</i>	-	Suppl. Table 4 from Frazer et al.	no	Less than 1000 variants

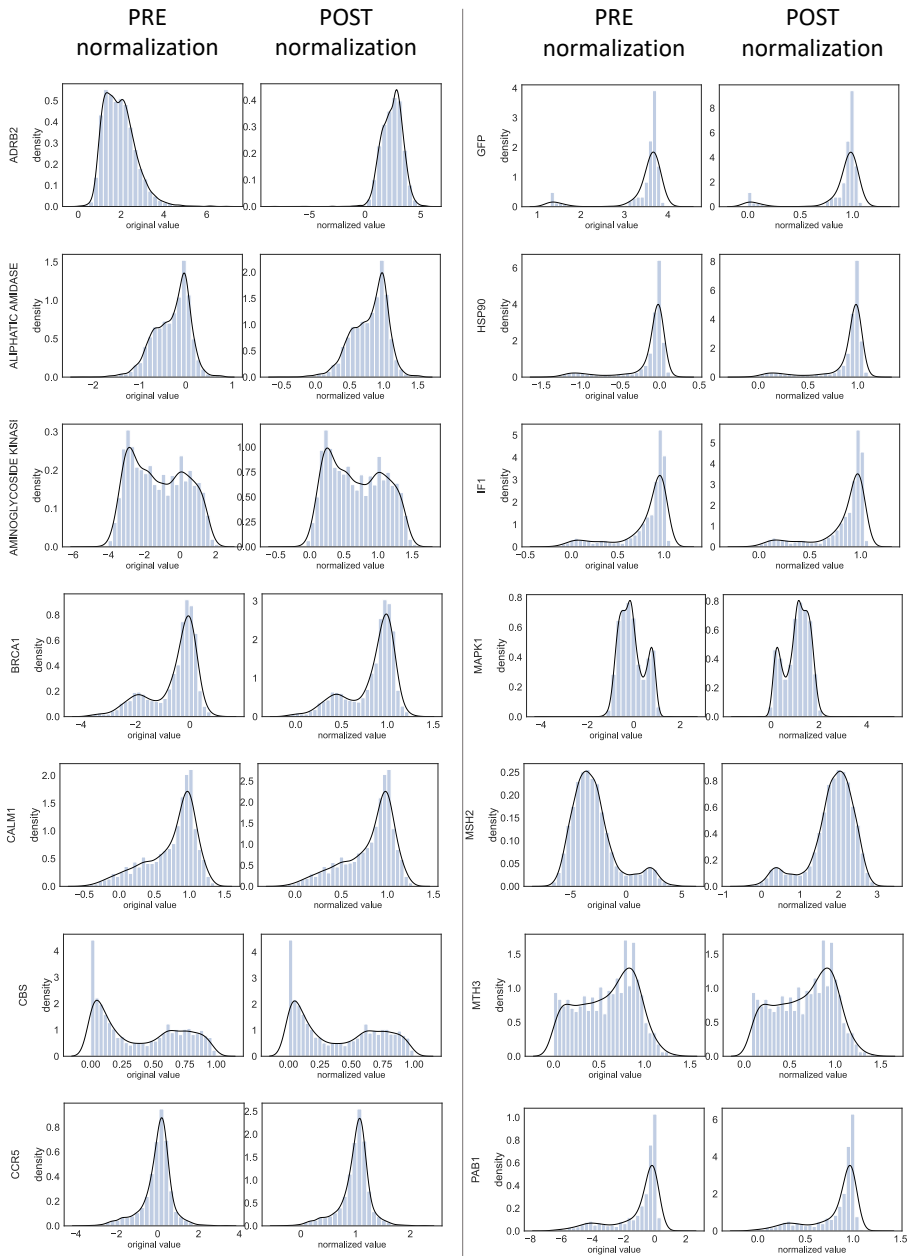
Table S5.2: The clinically labeled variant dataset. The dataset is curated from HumSavar and ClinVar variants. The table shows the UniProtID, Protein name, number of pathogenic and benign variants.

UNIPROT	PROTEIN	# OF PATHOGEN	# OF BENIGN	UNIPROT	PROTEIN	# OF PATHOGEN	# OF BENIGN
P33897	ABCD1	217	55	Q9BXM7	PINK1	24	43
P49815	TSC2	154	280	P40763	STAT3	74	21
Q96L73	NSD1	125	142	P04070	PROC	68	26
P32004	L1CAM	73	64	Q9BZS1	FOXP3	22	27
Q01955	COL4A3	107	58	P68871	HBB	35	238
P02458	COL2A1	331	113	Q05193	DNM1	36	45
P40692	MLH1	180	57	P48029	SLC6A8	23	42
P53420	COL4A4	81	67	P42345	MTOR	39	188
P35498	SCN1A	895	62	O95622	ADCY5	20	42
O95255	ABCC6	101	51	P20908	COL5A1	29	116
P46100	ATRX	60	249	Q9Y6X9	MORC2	25	20
P43246	MSH2	134	164	P38935	IGHMBP2	46	22
Q13224	GRIN2B	113	101	O15118	NPC1	227	25
P17813	ENG	84	79	O43933	PEX1	20	25
P78504	JAG1	66	63	O60741	HCN1	28	57
P12110	COL6A2	60	76	P61764	STXBP1	92	41
Q92793	CREBBP	100	142	P22607	FGFR3	36	26
P01130	LDLR	610	70	Q86UK0	ABCA12	27	37
P30531	SLC6A1	57	54	Q9UPW6	SATB2	30	47
P51531	SMARCA2	101	62	Q13402	MYO7A	133	48
Q9UM47	NOTCH3	235	93	P11498	PC	20	27
P29400	COL4A5	461	105	P54098	POLG	121	22
O15360	FANCA	61	67	P00740	F9	205	20
P55316	FOXP1	71	54	P28329	CHAT	22	27
Q14721	KCNB1	64	93	P08123	COL1A2	378	32
Q14524	SCN5A	175	55	Q05586	GRIN1	63	44
P78562	PHEX	138	25	O43918	AIRE	36	28
P04629	NTRK1	41	33	P12883	MYH7	390	24
P41229	KDM5C	37	35	P35499	SCN4A	93	22
O14746	TERT	44	45	P06213	INSR	62	21
Q9UBP0	SPAST	175	21	P13647	KRT5	74	25

continued

UNIPROT	PROTEIN	# OF PATHOGEN	# OF BENIGN	UNIPROT	PROTEIN	# OF PATHOGEN	# OF BENIGN
P46019	PHKA2	26	26	P07949	RET	121	23
Q9HBA0	TRPV4	66	24	P23378	GLDC	116	20
P22304	ID5	178	44	Q13496	MTM1	90	22
P31785	IL2RG	54	33	P35670	ATP7B	276	40
Q92574	TSC1	24	93	Q01968	OCRL	56	20
Q71F56	MED13L	25	114	P06400	RB1	42	45
Q13501	SQSTM1	27	31	Q14004	CDK13	25	54
Q12840	KIF5A	35	52	Q07889	SOS1	60	46
Q8N157	AHI1	23	33	O95259	KCNH1	21	59
P12109	COL6A1	48	93	O75443	TECTA	36	25
Q9HC10	OTOF	24	39	Q14896	MYBPC3	78	32
Q99593	TBX5	23	26	P52701	MSH6	43	66
P22735	TGM1	95	25	P13569	CFTR	259	30
Q27J81	INF2	41	75	Q14839	CHD4	35	20
O00754	MAN2B1	43	26	Q13873	BMPR2	33	42
P78363	ABCA4	513	38	P08581	MET	20	26
Q15878	CACNA1E	22	224	O14647	CHD2	42	74
Q99972	MYOC	55	37	O60500	NPHS1	78	49
P54760	EPHB4	22	56	Q9UPY3	DICER1	33	24
Q7Z333	SETX	31	110	P31327	CPS1	148	35
Q96QS3	ARX	28	22	P35579	MYH9	38	38
P05155	SERPING1	62	25	P10253	GAA	197	29
P21333	FLNA	47	179	P04839	CYBB	89	31
P00480	OTC	180	21	P00451	F8	569	27
Q6UB99	ANKRD11	24	167	Q5S007	LRRK2	25	66
P82279	CRB1	165	23	Q6ZUT3	FRMD7	30	30
Q93074	MED12	27	74	Q5K651	SAMD9	25	29
P16144	ITGB4	21	35	Q9ULV0	MYO5B	23	37
O76024	WFS1	72	44	O75197	LRP5	83	26
Q04671	OCA2	71	37	Q86XK2	FBXO11	46	33
Q8TAB3	PCDH19	105	21	P12755	SKI	22	22
Q13635	PTCH1	47	133	P02730	SLC4A1	46	42

Figure S5.1: Distribution of the DMS experiments for 30 proteins with histograms, before (PRE) and after (POST) normalization. The protein names are stated on the y-axis.



continued

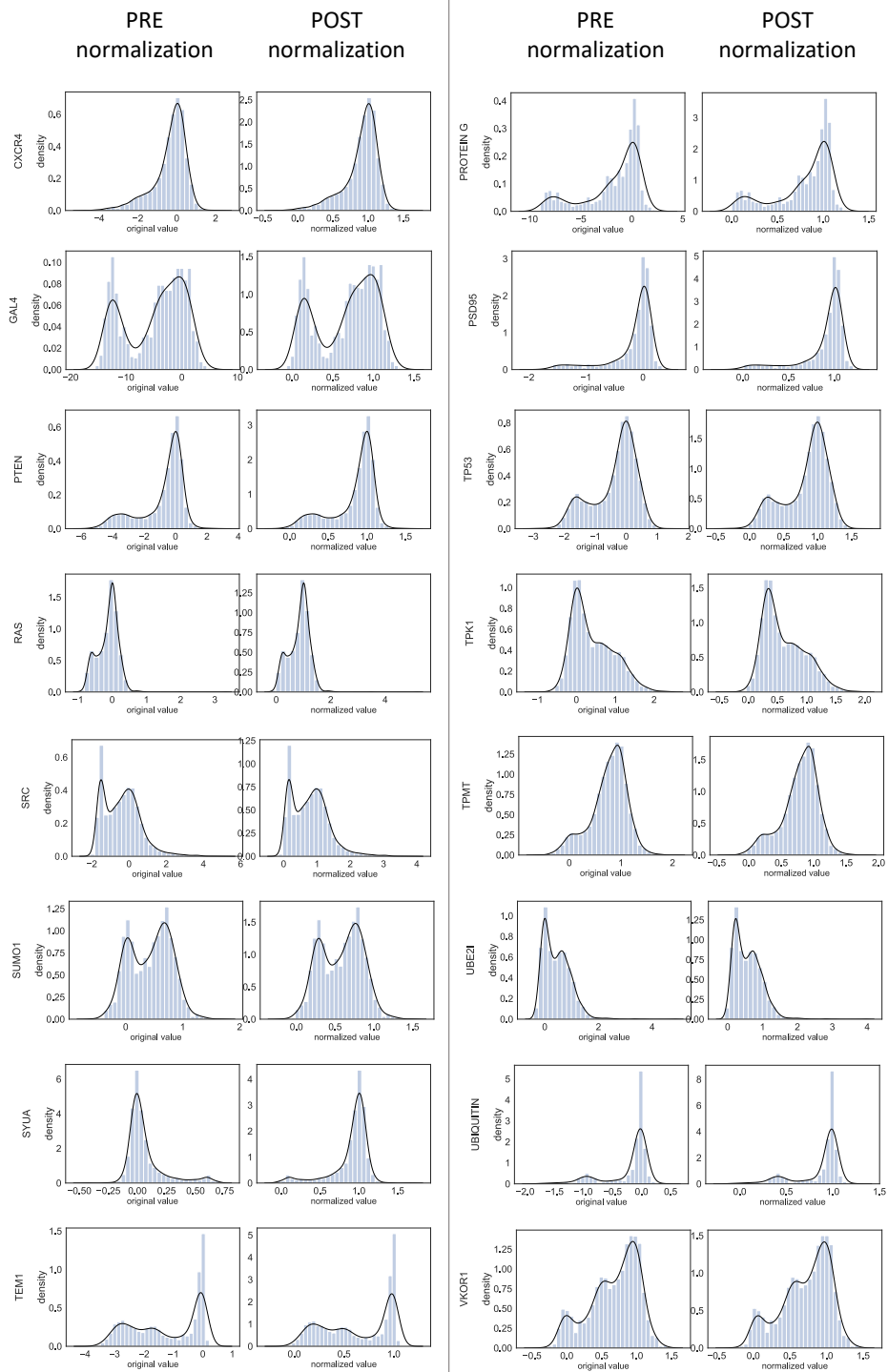
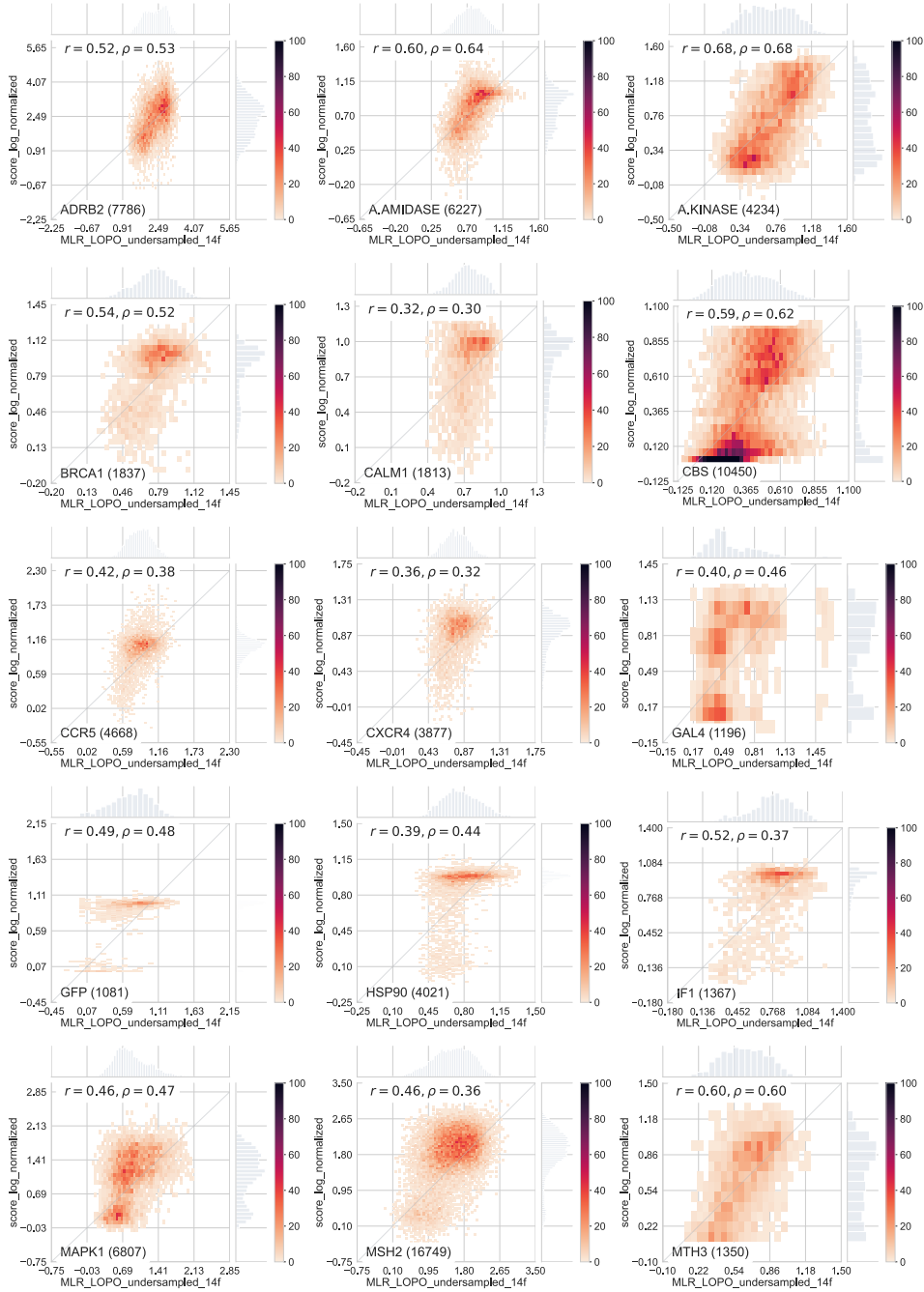


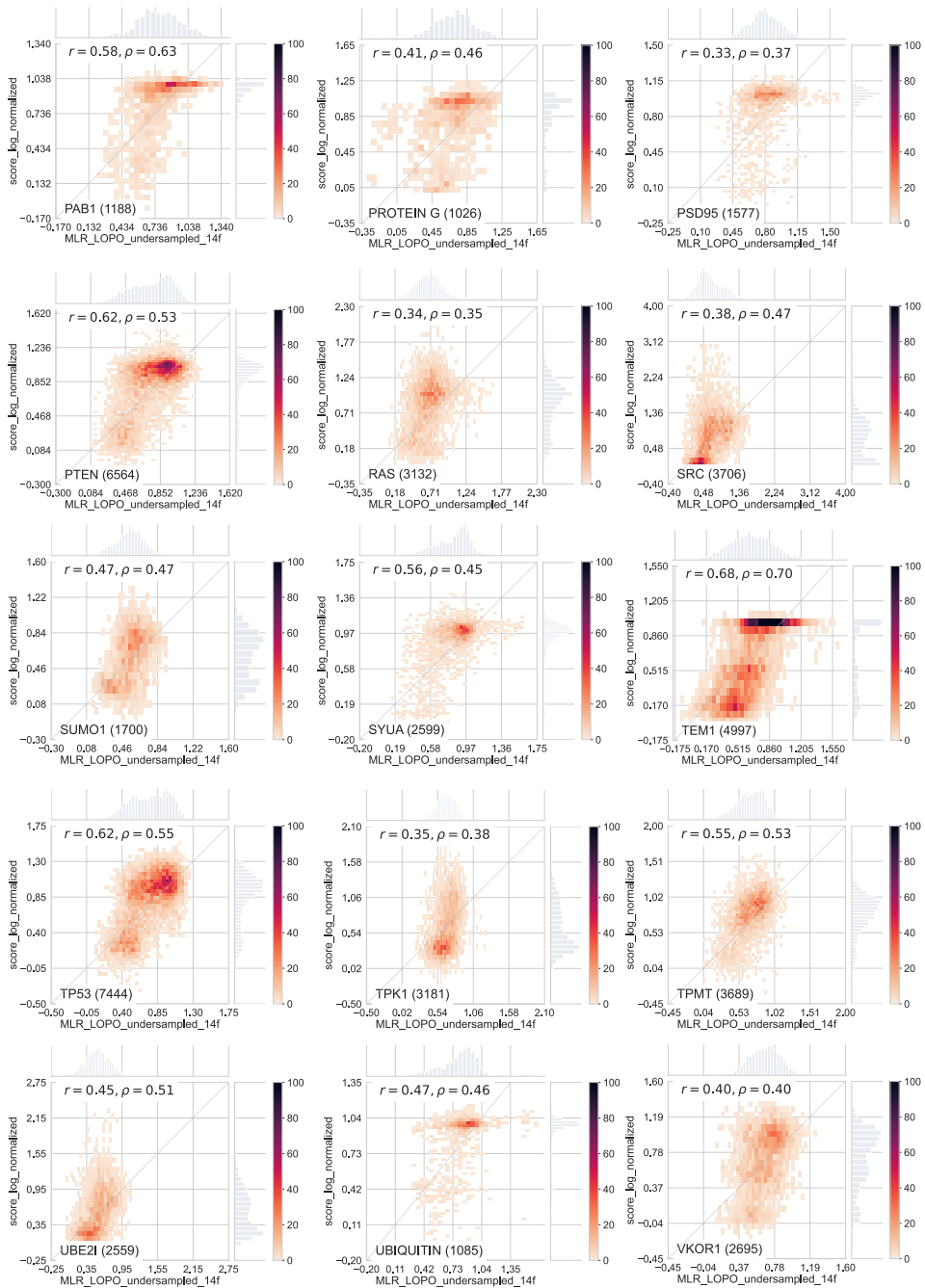
Figure S5.2: Leave-one-position-out performance of 30 protein-specific predictors. Observed vs. Predicted heatmap plots for 30 PSPs are shown. The names of the proteins, and the number of nonsynonymous variants of each dataset is stated at the lower left side of each heatmap. *

* The outlier values of four proteins' normalized assay scores were filtered out for visualization purposes (we used ADBR2 normalized assay scores ≥ -1 , MAPK1 normalized assay scores < 3 , RAS normalized assay scores < 2.5 , UBE2I normalized assay scores < 2.5 and SRC normalized assay scores < 3.12).



continued

5. Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors



6.

Discussion

Interpreting the clinical impact of variants on disease phenotype is a fundamental step for advancing precision medicine and individually tailored treatments (Wong *et al.*, 2021). This requires a comprehensive understanding of the molecular impact of variants, a conceptually and technically difficult task. Presently, the computational interpretation of variants is primarily conducted through binary pathogenicity predictors that are built using machine learning methods (Vihinen, 2012), which can contribute significantly to interpretation of sequence variants when used cautiously. These methods have the capacity to integrate large amounts of data from protein sequences, structures, and experimental assay values (Katsonis *et al.*, 2022) in predictive models. Numerous approaches have been developed for the pathogenicity prediction problem using different combinations of features, algorithms, refinements, etc. (summarized in Özkan *et al.*, 2021; Katsonis *et al.*, 2022; Liu *et al.*, 2022). While they show promising performances, they cannot be currently used in clinical settings as stand-alone devices due to their limitations (Bean and Hegde, 2017; Sun and Yu, 2019). The aim of this thesis was to develop a novel pathogenicity predictor for the binary prediction of missense variants, based on a fundamental formulation of the problem involving quantitative predictions of their molecular impacts. We described how this key challenge (Shendure, Findlay and Snyder, 2019) has been addressed in three main chapters.

In the chapter entitled **“The computational approach to the variant interpretation: principles, results, and applicability”**, we initiated our investigation by conducting a thorough examination of the overall performance of pathogenicity predictors commonly utilized in the bioinformatics field. This work has been published as a book chapter (Özkan *et al.*, 2021) in Elsevier’s ‘Clinical DNA Variant Interpretation’

book (Patrinos, 2021). In this review, we described the main characteristics of pathogenicity predictors, including how the training sets are collected, what kind of features are commonly used, the classification methods, how the performances are validated, etc. We explored the similarities and differences between general predictors (applied to any protein), protein-specific predictors, and metapredictors. We observed that, generally, metapredictors are the best performers, at the price of losing interpretability. Further, we discussed the varying performances of these tools in different benchmarking sets, how they differ from their performances in the original publication vs. independent sets. Looking ahead, we recognized that a promising advance in the field lies in the development of continuous predictors, which define the prediction problem as a quantitative one. The pioneer in this approach is ENVISION, which predicts the molecular effect of variants quantitatively.

Following this, in our study in the chapter “**Endophenotype-based strategy for pathogenicity prediction in BRCA1 & BRCA2 missense variants**”, we turned our attention to investigating the power of regression tools as main prediction engines in quantitative prediction (Özkan, Padilla and de la Cruz, 2021). This is relevant considering that there is a technical difference between the tools used for predicting the values of categorical and quantitative variables (Hastie *et al.*, 2009). We have recently started a research line in this direction testing the performance of simplistic multiple-linear-regression models in the CAGI ENIGMA challenge (Cline *et al.*, 2019). The model presented was trained to estimate the molecular impact of variants on BRCA1/2 function, using only three features (Shannon’s entropy, position-specific scoring matrix and BLOSUM62). Interestingly, this predictor ranked at top positions when compared to other predictors in the contest, including binary classifiers (Padilla *et al.*, 2019). In this part of the thesis, we used this

predictor as a starting point and explored various regression models with different feature combinations. We used the homology-directed repair (Starita *et al.*, 2017; Richardson *et al.*, 2021) and saturation genome editing assays (Findlay *et al.*, 2018) as target (dependent) variables and five features as independent variables. We tested the predictive performance of five regression techniques: multiple linear regression, ridge, lasso, elastic net, and kernel. The results of this work confirmed the competitiveness of regression-based models for the prediction of the clinical impact of variants. We showed that our sequence-based features captured the characteristics of the functional assays, showing good correlations mostly in the case of BRCA1. We further compared the performance of regression models against a simple, in-house Random Forest classifier, developed using the same set of features. We found that the binarized regression model showed similar predictive performance when compared to the Random Forest classifier, thus confirming the viability of binary predictions based on quantitative models of the variants' impact. Our data also confirmed the validity of the endophenotype conjecture (Masica and Karchin, 2016) stating that using these intermediate phenotypes may be a good alternative to the use of variant clinical annotations (pathogenic/benign).

The previous studies created the basis for the central part of our work. In our study in the chapter **“Binary pathogenicity classification of missense variants through development of quantitative protein-specific predictors”**, we explain the process of producing a general formulation of the prediction problem grounded on quantitative predictions, in line with the back-to-basics proposal of Shendure *et al.* (2019). To this end we broke down the prediction challenge into two parts: the quantitative and the binary prediction problems.

To address the quantitative prediction problem, we first built 30 protein-specific regression models based on the use of DMS experiments. This part allowed us to try different predictive features, in particular to incorporate the use of 3D information from the AlphaFold database, the recently published breakthrough (Jumper *et al.*, 2021). Then, to extend our approach to the proteins in the clinical genome, we developed an Ensemble learning-based model, using a set of PSPs.

We found that sequence and conservation-related features have an important predictive power, in accordance with what has been found in the development of other prediction tools (Kumar *et al.*, 2009; Niroula and Vihinen, 2016; Gyulkhanyan *et al.*, 2020). On the other hand, our feature importance analyses showed that structure-based features (such as colasi and the Miyazawa and Jernigan potential) were important contributors. This is relevant in the sense that these two features are coarse-grained approximations of the protein energetics. Additionally, this is consistent with recent studies which showed that integration of features obtained from AlphaFold models increased the predictive ability of pathogenicity predictors (Schmidt *et al.*, 2023).

As a result, integrating sequence- and structure-based features with the simple regression technique resulted in prediction models that accurately captured variants' molecular impacts. We compared the leave-one-position-out cross-validated performance of our protein-specific models against their reference method, ENVISION (Gray *et al.*, 2018), and our results (**Figure 5.11**) showed that our models outperformed ENVISION in 17 out of 20 proteins missense variant predictions, according to Pearson's correlation coefficients.

Subsequently, based on the conjecture that the PSPs capture overlapping aspects of a variant's molecular impact, we developed an Ensemble-based approach to extend quantitative predictions to missense variants from any protein. The median method employed has several advantages for the variant prediction problem (Cao et al., 2020; Sagi & Rokach, 2018). First, the possibility of any DMS experiment-specific bias during training is minimized; second, each protein-specific model can capture a different aspect of the molecular effect. Lastly, taking the median reduces the influence of noisy predictions from individual protein-specific models. Our model's molecular effect predictions hold the potential to aid in the variant interpretation problem by providing valuable information on the phenotypic effects of variants (Fowler et al., 2010; Gasperini et al., 2016; Patwardhan et al., 2009; Starita et al., 2017). It's predictions can contribute to DMS experiments when it fail to provide variant effect scores for all variants they assayed (Dorrity et al., 2018; Findlay et al., 2014; Firnberg et al., 2014; Matreyek et al., 2018).

Lastly, we addressed the binary prediction problem, discretizing the outcome of our Ensemble model. The results we obtained showed (**Figures 5.15**) a good performance when compared with a selection of top ranking *in silico* tools (Pejaver et al., 2022). Indeed, in our benchmarking experiment, our success rate was close to that of these tools. This confirmed that our work clears the way for the development of fundamental approaches to the prediction problem and confirms the feasibility of the fundamental approach proposed by Shendure et al. (2019).

7.

Conclusion

The conclusions of this thesis are the following:

- The revision of the literature shows that predicting the clinical impact of variants remains a major bottleneck in variant interpretation.
- The development of quantitative predictors for estimating the impact of variants holds promise for solving the variant interpretation problem.
- Computational estimation of the endophenotype change upon mutation for BRCA1/2 proteins shows promise in achieving accuracy levels suitable for pathogenicity prediction problem.
- We have developed 30 protein-specific tools for the quantitative estimation of the functional impact of variants. Direct comparison of these models against their immediate competitor, ENVISION, showed a better performance in 17 out of 20 cases.
- The ensemble-based pathogenicity predictor, built upon 10 top PSPs using the median approach, can predict the molecular impact of variants with only a small drop in performance. This shows that the median approach is a good method for extending our quantitative model to the whole clinical genome.
- Departures are observed from proteins deviating from the average globular protein or having multiple interaction partners.
- When benchmarked against other predictors, our Ensemble-based model showed good predictive performance. However, in most of the cases it was outperformed by DeepSequence, an unsupervised method.

- The binary predictions derived from the Ensemble-based predictor showed good accuracy.
- When compared against top ranking predictors REVEL, VEST4, BayesDel and CADD our binary predictions were very close. Although we did not outperform these classifiers, obtaining close performances indicates that we are able to represent the binary status of missense variants given the current stage of our model.
- Globally, our work confirms that it is presently possible to develop a fundamental approach to address the variant interpretation problem that competes with symbol-based approaches.

8.

Bibliography

- Adebali, O. *et al.* (2016) 'Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations', *Genetics in Medicine*, 18(10), pp. 1029–1036. Available at: <https://doi.org/10.1038/gim.2015.208>.
- Adzhubei, I.A. *et al.* (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods*, 7(4), pp. 248–249. Available at: <https://doi.org/10.1038/nmeth0410-248>.
- Ahmed, Z. *et al.* (2020) 'Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine', *Database*. Oxford University Press. Available at: <https://doi.org/10.1093/database/baaa010>.
- Alan Fersht (1985) *Enzyme Structure and Mechanism*. 2nd edn. W.H. Freeman.
- Alber, M. *et al.* (2019) 'Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences', *npj Digital Medicine*. Nature Research. Available at: <https://doi.org/10.1038/s41746-019-0193-y>.
- Alirezaie, N. *et al.* (2018) 'ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants', *American Journal of Human Genetics*, 103(4), pp. 474–483. Available at: <https://doi.org/10.1016/j.ajhg.2018.08.005>.
- Aljarf, R. *et al.* (2022) 'Understanding and predicting the functional consequences of missense mutations in BRCA1 and BRCA2', *Scientific Reports*, 12(1), pp. 1–12. Available at: <https://doi.org/10.1038/s41598-022-13508-3>.
- Altschul, S.F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Research*, 25(17), pp. 3389–3402. Available at: <https://doi.org/10.1093/nar/25.17.3389>.
- Amendola, L.M. *et al.* (2016) 'Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium', *The American*

- Journal of Human Genetics*, 98(6), pp. 1067–1076. Available at: <https://doi.org/10.1016/j.ajhg.2016.03.024>.
- Ancien, F. *et al.* (2018) 'Prediction and interpretation of deleterious coding variants in terms of protein structural stability', *Scientific Reports*, 8(1), pp. 1–11. Available at: <https://doi.org/10.1038/s41598-018-22531-2>.
- Andreis, T.F. *et al.* (2023) 'Challenges in periodic revision of genetic testing results: Comparison of the main classification guidelines and report of a retrospective analysis involving BRCA1/BRCA2 variants of uncertain significance', *Gene*, 862(January). Available at: <https://doi.org/10.1016/j.gene.2023.147281>.
- Andreoletti, G. *et al.* (2019) 'Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation', *Human Mutation*, 40(9), pp. 1197–1201. Available at: <https://doi.org/10.1002/humu.23876>
- Angarica, V.E., Orozco, M. and Sancho, J. (2015) 'Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: Linking snps with disease phenotypes in familial hypercholesterolemia', *Human Molecular Genetics*, 25(6), pp. 1233–1246. Available at: <https://doi.org/10.1093/hmg/ddw004>.
- Antonarakis, S.E. and Cooper, D.N. (2018) *Human genomic variants and inherited disease: Molecular mechanisms and clinical consequences*. Seventh Ed, *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics: Foundations*. Seventh Ed. Elsevier Inc. Available at: <https://doi.org/10.1016/B978-0-12-812537-3.00006-8>.
- Arbiza, L. *et al.* (2006) 'Selective Pressures at a Codon-level Predict Deleterious Mutations in Human Disease Genes', *Journal of Molecular Biology*, 358(5), pp. 1390–1404. Available at: <https://doi.org/10.1016/j.jmb.2006.02.067>.
- Azodi, C.B., Tang, J. and Shiu, S.H. (2020) 'Opening the Black Box: Interpretable Machine Learning for Geneticists', *Trends in Genetics*, 36(6), pp. 442–455. Available at: <https://doi.org/10.1016/j.tig.2020.03.005>.

- Baldi, P. *et al.* (2000) 'Assessing the accuracy of prediction algorithms for classification: An overview', *Bioinformatics*, 16(5), pp. 412–424. Available at: <https://doi.org/10.1093/bioinformatics/16.5.412>.
- Baptista, D., Ferreira, P.G. and Rocha, M. (2021) 'Deep learning for drug response prediction in cancer', *Briefings in Bioinformatics*, 22(1), pp. 360–379. Available at: <https://doi.org/10.1093/bib/bbz171>.
- Bateman, A. *et al.* (2017) 'UniProt: The universal protein knowledgebase', *Nucleic Acids Research*, 45(D1), pp. D158–D169. Available at: <https://doi.org/10.1093/nar/gkw1099>.
- Bean, L.J.H. and Hegde, M.R. (2017) 'Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines', *Genome Medicine*, 9(1), pp. 1–3. Available at: <https://doi.org/10.1186/s13073-017-0508-z>
- Bepler, T. and Berger, B. (2021) 'Learning the protein language: Evolution, structure, and function', *Cell Systems*, 12(6), pp. 654–669.e3. Available at: <https://doi.org/10.1016/j.cels.2021.05.017>.
- Bhasin, M. and Varadarajan, R. (2021) 'Prediction of Function Determining and Buried Residues Through Analysis of Saturation Mutagenesis Datasets', *Frontiers in Molecular Biosciences*, 8(March), pp. 1–11. Available at: <https://doi.org/10.3389/fmolb.2021.635425>.
- Bhattacharya, R. *et al.* (2017) 'Impact of genetic variation on three dimensional structure and function of proteins', *PLoS ONE*, 12(3), pp. 1–22. Available at: <https://doi.org/10.1371/journal.pone.0171355>.
- Binder, J.L. *et al.* (2022) 'AlphaFold illuminates half of the dark human proteins', *Current Opinion in Structural Biology*, 74(April). Available at: <https://doi.org/10.1016/j.sbi.2022.102372>.
- Bioinformatics at COMAV (no date) *Multiple Sequence Alignment*. Available at: <https://bioinf.comav.upv.es/courses/biotech3/theory/multiple.html> (Accessed: 12 July 2023).

- Bishop, C.M. (2011) *Pattern recognition and machine learning*. Vol. 4, No. New York: Springer.
- Bondi, A. (1964) 'van der Waals Volumes and Radii', *The Journal of Physical Chemistry*, 68(3), pp. 441–451. Available at: <https://doi.org/10.1021/j100785a001>.
- Botstein, D. and Shortle, D. (1985) 'Strategies and Applications of in Vitro Mutagenesis', *Science*, 229(4719). Available at: <https://doi.org/10.1126/science.2994214>.
- Breiman, L. (2001) 'Random Forest', *Machine Learning*, 45(1), pp. 5–32. Available at: <https://doi.org/10.1023/A:1010933404324>
- Brzovic, P.S. et al. (2001) 'Structure of a BRCA1 – BARD1 heterodimeric RING – RING complex', 8(10), pp. 833–837. Available at: <https://doi.org/10.1038/nsb1001-833>
- Cao, Y. et al. (2020) 'Ensemble deep learning in bioinformatics', *Nature Machine Intelligence*, 2(9), pp. 500–508. Available at: <https://doi.org/10.1038/s42256-020-0217-y>.
- Capriotti, E., Fariselli, P. and Casadio, R. (2005) 'I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure', *Nucleic Acids Research*, 33, pp. W306–W310. Available at: <https://doi.org/10.1093/nar/gki375>.
- Carrell, R.W. and Lomas, D.A. (1997) 'Conformational disease', *Lancet*, 350(9071), pp. 134–138.
- Carretero-Puche, C. et al. (2020) 'How can bioinformatics contribute to the routine application of personalized precision medicine?', *Expert Review of Precision Medicine and Drug Development*, 5(3), pp. 115–117. Available at: <https://doi.org/10.1080/23808993.2020.1758062>.
- Carter, H. et al. (2013) 'Identifying Mendelian disease genes with the variant effect scoring tool.', *BMC genomics*, 14 Suppl 3(Suppl 3), p. S3. Available at: <https://doi.org/10.1186/1471-2164-14-s3-s3>.
- Chen, R. and Snyder, M. (2013) 'Promise of personalized omics to precision medicine', *Wiley Interdisciplinary Reviews: Systems*

- Biology and Medicine*, 5(1), pp. 73–82. Available at: <https://doi.org/10.1002/wsbm.1198>.
- Chen, S., Hu, X. and Shen, Y. (2015) 'Sequence variant interpretation 2.0: Perspective on new guidelines for sequence variant classification', *Clinical Chemistry*. American Association for Clinical Chemistry Inc., pp. 1317–1319. Available at: <https://doi.org/10.1373/clinchem.2015.240812>.
- Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(1), pp. 1–13. Available at: <https://doi.org/10.1186/s12864-019-6413-7>.
- Cline, M.S. *et al.* (2019) 'Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants', *Human Mutation*, 40(9), pp. 1546–1556. Available at: <https://doi.org/10.1002/humu.23861>.
- Conchillo-Solé, O. *et al.* (2007) 'AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides', *BMC Bioinformatics*, 8, p. 65. Available at: <https://doi.org/10.1186/1471-2105-8-65>
- Cooper, D.N. *et al.* (2010) 'Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics', *Human Mutation*, 31(6), pp. 631–655. Available at: <https://doi.org/10.1002/humu.21260>.
- Cover, T.M. and Thomas, J.A. (2005) 'Entropy, Relative Entropy, and Mutual Information', in *Elements of Information Theory*, pp. 13–55. Available at: <https://doi.org/https://doi.org/10.1002/047174882X.ch2>.
- Crockett, D.K. *et al.* (2012) 'Consensus: a framework for evaluation of uncertain gene variants in laboratory test reporting', *Genome Medicine*, 4(5), p. 48. Available at: <https://doi.org/10.1186/gm347>.
- Cubuk, C. *et al.* (2021) 'Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes', *Genetics in Medicine*, 23(11), pp.

- 2096–2104. Available at: <https://doi.org/10.1038/s41436-021-01265-z>.
- Dai, X. and Shen, L. (2022) 'Advances and Trends in Omics Technology Development', *Frontiers in Medicine*, 9(July). Available at: <https://doi.org/10.3389/fmed.2022.911861>.
- David, A. and Sternberg, M.J. (2015) 'The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease', *Journal of Molecular Biology*, 427(17), pp. 2886–2898. Available at: <https://doi.org/10.1016/j.jmb.2015.07.004>
- Dehouck, Y. *et al.* (2011) 'PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality', *BMC Bioinformatics*, 12, p. 151. Available at: <https://doi.org/10.1186/1471-2105-12-151>
- Denny, J.C. and Collins, F.S. (2021) 'Precision medicine in 2030—seven ways to transform healthcare', *Cell*, 184(6), pp. 1415–1419. Available at: <https://doi.org/10.1016/j.cell.2021.01.015>.
- Deo, R.C. (2015) 'Machine learning in medicine', *Circulation*, 132(20), pp. 1920–1930. Available at: <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>.
- Dorrity, M.W. *et al.* (2018) 'Preferences in a trait decision determined by transcription factor variants', *Proceedings of the National Academy of Sciences of the United States of America*, 115(34), pp. E7997–E8006. Available at: <https://doi.org/10.1073/pnas.1805882115>.
- Dunham, A.S. and Beltrao, P. (2021) 'Exploring amino acid functions in a deep mutational landscape', *Molecular Systems Biology*, 17(7), pp. 1–14. Available at: <https://doi.org/10.15252/msb.202110305>.
- Echave, J., Spielman, S.J. and Wilke, C.O. (2016) 'Causes of evolutionary rate variation among protein sites', *Nature Reviews Genetics*, 17(2), pp. 109–121. Available at: <https://doi.org/10.1038/nrg.2015.18>.

- Edgar, R.C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic acids research*, 32(5), pp. 1792–7. Available at: <https://doi.org/10.1093/nar/gkh340>.
- Ellis, R.J. and Minton, A.P. (2006) 'Protein aggregation in crowded environments', *Biological Chemistry*, 387(5), pp. 485–497. Available at: <https://doi.org/10.1515/BC.2006.064>
- Emmert-Streib, F. and Dehmer, M. (2018) 'A Machine Learning Perspective on Personalized Medicine: An Automized, Comprehensive Knowledge Base with Ontology for Pattern Recognition', *Machine Learning and Knowledge Extraction*, 1(1), pp. 149–156. Available at: <https://doi.org/10.3390/make1010009>.
- Eraslan, G. *et al.* (2019) 'Deep learning : new computational modelling techniques for genomics', *Nature Reviews Genetics*, 20(July). Available at: <https://doi.org/10.1038/s41576-019-0122-6>.
- Ernst, C. *et al.* (2018) 'Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics', *BMC Medical Genomics*, 11(1), p. 35. Available at: <https://doi.org/10.1186/s12920-018-0353-y>.
- Esteva, A. *et al.* (2017) 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, 542(7639), pp. 115–118. Available at: <https://doi.org/10.1038/nature21056>.
- Fang, M. *et al.* (2022) 'VIPPID: A gene-specific single nucleotide variant pathogenicity prediction tool for primary immunodeficiency diseases', *Briefings in Bioinformatics*, 23(5), pp. 1–10. Available at: <https://doi.org/10.1093/bib/bbac176>.
- Fauchere, J. (1983) 'Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides', *Eur. J. Med. Chem.-Chim. Ther.*, 18, pp. 369–375.
- Faure, A.J. *et al.* (2020) 'DiMSum: An error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies', *Genome Biology*, 21(1). Available at: <https://doi.org/10.1186/s13059-020-02091-3>.

- Feng, B.J. (2017) 'PERCH: A Unified Framework for Disease Gene Prioritization', *Human Mutation*, 38(3), pp. 243–251. Available at: <https://doi.org/10.1002/humu.23158>.
- Fernald, G.H. *et al.* (2011) 'Bioinformatics challenges for personalized medicine', *Bioinformatics*, pp. 1741–1748. Available at: <https://doi.org/10.1093/bioinformatics/btr295>.
- Fernandez-Escamilla, A.M. *et al.* (2004) 'Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins', *Nature Biotechnology*, 22(10), pp. 1302–1306.
- Fernández-Recio, J. (2011) 'Prediction of protein binding sites and hot spots', *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5), pp. 680–698. Available at: <https://doi.org/10.1002/wcms.45>.
- Ferrer-Costa, C., Orozco, M. and De La Cruz, X. (2002) 'Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties', *Journal of Molecular Biology*, 315(4), pp. 771–786. Available at: <https://doi.org/10.1006/jmbi.2001.5255>.
- Ferrer-Costa, C., Orozco, M. and de la Cruz, X. (2004) 'Sequence-based prediction of pathological mutations', *Proteins: Structure, Function and Genetics*, 57, pp. 811–819. Available at: <https://doi.org/10.1002/prot.20252>.
- Findlay, G.M. *et al.* (2014) 'Saturation editing of genomic regions by multiplex homology-directed repair.', *Nature*, 513(7516), pp. 120–123. Available at: <https://doi.org/10.1038/nature13695>.
- Findlay, G.M. *et al.* (2018) 'Accurate classification of BRCA1 variants with saturation genome editing', *Nature*, 562(7726), pp. 217–222. Available at: <https://doi.org/10.1038/s41586-018-0461-z>.
- Findlay, G.M. (2021) 'Linking genome variants to disease: Scalable approaches to test the functional impact of human mutations', *Human Molecular Genetics*, 30(R2), pp. R187–R197. Available at: <https://doi.org/10.1093/hmg/ddab219>.
- Firnberg, E. *et al.* (2014) 'A comprehensive, high-resolution map of a Gene's fitness landscape', *Molecular Biology and Evolution*,

- 31(6), pp. 1581–1592. Available at: <https://doi.org/10.1093/molbev/msu081>.
- Fogel, A.L. and Kvedar, J.C. (2018) 'Artificial intelligence powers digital medicine', *npj Digital Medicine*. Nature Publishing Group. Available at: <https://doi.org/10.1038/s41746-017-0012-2>.
- Fowler, D.M. *et al.* (2010) 'High-resolution mapping of protein sequence-function relationships', *Nature Methods*, 7(9), pp. 741–746. Available at: <https://doi.org/10.1038/nmeth.1492>.
- Fowler, D.M. and Fields, S. (2014) 'Deep mutational scanning: A new style of protein science', *Nature Methods*, 11(8), pp. 801–807. Available at: <https://doi.org/10.1038/nmeth.3027>.
- Frazer, J. *et al.* (2021) 'Disease variant prediction with deep generative models of evolutionary data', *Nature*, 599(7883), pp. 91–95. Available at: <https://doi.org/10.1038/s41586-021-04043-8>.
- Fröhlich, H. *et al.* (2018) 'From hype to reality: Data science enabling personalized medicine', *BMC Medicine*, 16(1), pp. 1–15. Available at: <https://doi.org/10.1186/s12916-018-1122-7>.
- Galano-Frutos, J.J., Garcíá-Cebollada, H. and Sancho, J. (2019) 'Molecular dynamics simulations for genetic interpretation in protein coding regions: Where we are, where to go and when', *Briefings in Bioinformatics*, 22(1), pp. 3–19. Available at: <https://doi.org/10.1093/bib/bbz146>.
- Gao, M., Zhou, H. and Skolnick, J. (2015) 'Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis', *Structure*, 23(7), pp. 1362–1369. Available at: <https://doi.org/10.1016/j.str.2015.03.028>.
- Garcia, F.A. de O., Andrade, E.S. de and Palmero, E.I. (2022) 'Insights on variant analysis in silico tools for pathogenicity prediction', *Frontiers in Genetics*, 13(November), pp. 1–14. Available at: <https://doi.org/10.3389/fgene.2022.1010327>.
- Gasparini, M., Starita, L. and Shendure, J. (2016) 'The power of multiplexed functional analysis of genetic variants', *Nature Protocols*. Nature Publishing Group, pp. 1782–1787. Available at: <https://doi.org/10.1038/nprot.2016.135>.

- Gelman, H. *et al.* (2019) 'Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation', *Genome Medicine*, 11(1), pp. 1–11. Available at: <https://doi.org/10.1186/s13073-019-0698-7>.
- Geng, C. *et al.* (2019) 'Finding the $\Delta\Delta G$ spot: are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it?', *WIREs Computational Molecular Science*, 9(5), p. e1410. Available at: <https://doi.org/10.1002/wcms.1410>
- Ghosh, R., Oak, N. and Plon, S.E. (2017) 'Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines', *Genome Biology*, 18, p. 225. Available at: <https://doi.org/10.1186/s13059-017-1353-5>.
- Ginneken, B. *et al.* (2015). 'Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans' *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Brooklyn, NY, USA, pp. 286–289. Available at: 10.1109/ISBI.2015.7163869.
- Goecks, J. *et al.* (2020) 'How Machine Learning Will Transform Biomedicine', *Cell*, 181(1), pp. 92–101. Available at: <https://doi.org/10.1016/j.cell.2020.03.022>.
- Gomez-Lopez, G. *et al.* (2017) 'Precision medicine needs pioneering clinical bioinformaticians', *Briefings in Bioinformatics*, 20(3), pp. 752–766. Available at: <https://doi.org/10.1093/bib/bbx144>.
- Gong, S. and Blundell, T.L. (2010) 'Structural and Functional Restraints on the Occurrence of Single Amino Acid Variations in Human Proteins', *PLOS on*, 5(2), p. e9186. Available at: <https://doi.org/10.1371/journal.pone.0009186>.
- Graña-Montes R *et al.* (2017) 'Prediction of protein aggregation and amyloid formation', in *From Protein Structure to Function with Bioinformatics: Second Edition*, pp. 205–63.
- Gray, V.E. *et al.* (2018) 'Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data', *Cell Systems*, 6(1), pp. 116–124.e3. Available at: <https://doi.org/10.1016/j.cels.2017.11.003>.

- Greener, J.G. *et al.* (2022) 'A guide to machine learning for biologists', *Nature Reviews Molecular Cell Biology*, 23(1), pp. 40–55. Available at: <https://doi.org/10.1038/s41580-021-00407-0>.
- Grimm, D.G. *et al.* (2015) 'The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity', *Human Mutation*, 36(5), pp. 513–523. Available at: <https://doi.org/10.1002/humu.22768>.
- Guerois, R., Nielsen, J.E. and Serrano, L. (2002) 'Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations', *Journal of Molecular Biology*, 320(2), pp. 369–387. Available at: [https://doi.org/10.1016/S0022-2836\(02\)00442-4](https://doi.org/10.1016/S0022-2836(02)00442-4)
- Gupta, K. and Varadarajan, R. (2018) 'Insights into protein structure, stability and function from saturation mutagenesis', *Current Opinion in Structural Biology*, 50, pp. 117–125. Available at: <https://doi.org/10.1016/j.sbi.2018.02.006>.
- Gyulkhandanyan, A. *et al.* (2020) 'Analysis of protein missense alterations by combining sequence- and structure-based methods', *Molecular Genetics and Genomic Medicine*, 8(4), pp. 1–28. Available at: <https://doi.org/10.1002/mgg3.1166>.
- Hart, S.N. *et al.* (2018) 'Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models', *Genetics in Medicine*, 0(0), pp. 1–10. Available at: <https://doi.org/10.1038/s41436-018-0018-4>.
- Hart, S.N. *et al.* (2020) 'Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML', *npj Breast Cancer*, 6(1), pp. 1–4. Available at: <https://doi.org/10.1038/s41523-020-0159-x>.
- Hartl, F.U. (2017) 'Protein misfolding diseases', *Annu. Rev. Biochem*, 86(1), pp. 21–16. Available at: <https://doi.org/10.4155/fso.15.38>.
- Hasanin, T. *et al.* (2020) 'Investigating class rarity in big data', *Journal of Big Data*, 7(1). Available at: <https://doi.org/10.1186/s40537-020-00301-0>.

- Hastie, T. *et al.* (2009) *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer. Available at: <https://doi.org/10.1007/978-0-387-21606-5>.
- Henikoff, S. and Henikoff, J.G. (1992) 'Amino acid substitution matrices from protein blocks.', *Proceedings of the National Academy of Sciences*, 89(22), pp. 10915–10919. Available at: <https://doi.org/10.1073/pnas.89.22.10915>.
- Ho, D.S.W. *et al.* (2019) 'Machine learning SNP based prediction for precision medicine', *Frontiers in Genetics*, 10(MAR), pp. 1–10. Available at: <https://doi.org/10.3389/fgene.2019.00267>.
- Hopf, T.A. *et al.* (2017) 'Mutation effects predicted from sequence co-variation', *Nature Biotechnology*, 35(2), pp. 128–135. Available at: <https://doi.org/10.1038/nbt.3769>.
- Hoskins, R.A. *et al.* (2017) 'Reports from CAGI: The Critical Assessment of Genome Interpretation', *Human Mutation*, 38(9), pp. 1039–1041. Available at: <https://doi.org/10.1002/humu.23290>.
- Hoskinson, D.C., Dubuc, A.M. and Mason-Suares, H. (2017) 'The current state of clinical interpretation of sequence variants', *Current Opinion in Genetics and Development*, 42, pp. 33–39. Available at: <https://doi.org/10.1016/j.gde.2017.01.001>.
- Hsieh, Y.Y. *et al.* (2016) 'A linear regression model with dynamic pulse transit time features for noninvasive blood pressure prediction', *Proceedings - 2016 IEEE Biomedical Circuits and Systems Conference, BioCAS 2016*, pp. 604–607. Available at: <https://doi.org/10.1109/BioCAS.2016.7833867>.
- Hu, Z. *et al.* (2019) 'VIPdb, a genetic Variant Impact Predictor Database', *Human Mutation*, 40(9), pp. 1202–1214. Available at: <https://doi.org/10.1002/humu.23858>.
- Ioannidis, N.M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants', *American Journal of Human Genetics*, 99(4), pp. 877–885. Available at: <https://doi.org/10.1016/j.ajhg.2016.08.016>.
- Iqbal, S. *et al.* (2020) 'Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of

- missense variants', 117(45). Available at: <https://doi.org/10.1073/pnas.2002660117>.
- Itan, Y. *et al.* (2016) 'The mutation significance cutoff (MSC): gene-level thresholds for variant-level predictions', *Nature methods*, 13(2), pp. 109–110. Available at: <https://doi.org/10.1038/nmeth.3739>.The.
- Jensen, H.H. *et al.* (2018) 'Human calmodulin mutations', *Frontiers in Molecular Neuroscience*, 11(November). Available at: <https://doi.org/10.3389/fnmol.2018.00396>.
- Jordan, D.M., Ramensky, V.E. and Sunyaev, S.R. (2010) 'Human allelic variation: perspective from protein function, structure, and evolution', *Current Opinion in Structural Biology*, 20(3), pp. 342–350. Available at: <https://doi.org/10.1016/j.sbi.2010.03.006>.
- Jubb, H.C. *et al.* (2017) 'Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures', *Journal of Molecular Biology*, 429(3), pp. 365–371. Available at: <https://doi.org/10.1016/j.jmb.2016.12.004>.
- Jumper, J. *et al.* (2021a) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589. Available at: <https://doi.org/10.1038/s41586-021-03819-2>.
- Jumper, J. *et al.* (2021b) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589. Available at: <https://doi.org/10.1038/s41586-021-03819-2>.
- Jumper, J. *et al.* (2022) *Statin-interacting protein 1*, *AlphaFold database*. Available at: <https://alphafold.ebi.ac.uk/entry/Q5VSL9> (Accessed: 12 July 2023).
- K. Mazandu, G. *et al.* (2019) 'Designing Data-Driven Learning Algorithms: A Necessity to Ensure Effective Post-Genomic Medicine and Biomedical Research', *Artificial Intelligence - Applications in Medicine and Biology*, pp. 1–16. Available at: <https://doi.org/10.5772/intechopen.84148>.
- Karczewski, K.J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581(7809),

pp. 434–443. Available at: <https://doi.org/10.1038/s41586-020-2308-7>.

Katsonis, P. *et al.* (2022) 'Genome interpretation using in silico predictors of variant impact', *Human Genetics*. Springer Science and Business Media Deutschland GmbH, pp. 1549–1577. Available at: <https://doi.org/10.1007/s00439-022-02457-6>.

Kircher, M. *et al.* (2014) 'A general framework for estimating the relative pathogenicity of human genetic variants', *Nature Genetics*, 46(3), pp. 310–315. Available at: <https://doi.org/10.1038/ng.2892>.

Knecht, C. *et al.* (2017) 'IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants', *Nucleic acids research*, 45(3), p. e13. Available at: <https://doi.org/10.1093/nar/gkw886>.

Kucukkal, T.G. *et al.* (2015) 'Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins', *Current Opinion in Structural Biology*, 32, pp. 18–24. Available at: <https://doi.org/10.1016/j.sbi.2015.01.003>.

Kumar, P., Henikoff, S. and Ng, P.C. (2009) 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nature Protocols*, 4(7), p. 1073. Available at: <https://doi.org/10.1038/nprot.2009.86>.

Kumar, S. *et al.* (2009) 'Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations', *Genome Research*, 19(9), pp. 1562–1569. Available at: <https://doi.org/10.1101/gr.091991.109>.

de la Campa, E. Álvarez, Padilla, N. and de la Cruz, X. (2017) 'Development of pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence', *BMC Genomics*, 18. Available at: <https://doi.org/10.1186/s12864-017-3914-0>.

Lakhani, P., Flanders, A. and Gorniak, R. (2021) 'Endotracheal tube position assessment on chest radiographs using deep learning', *Radiology: Artificial Intelligence*, 3(1). Available at: <https://doi.org/10.1148/ryai.2020200026>.

- Landrum, M.J. *et al.* (2016) 'ClinVar: Public archive of interpretations of clinically relevant variants', *Nucleic Acids Research*, 44(D1), pp. D862–D868. Available at: <https://doi.org/10.1093/nar/gkv1222>.
- Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Physiology & behavior*, 7616, pp. 285–291. Available at: <https://doi.org/10.1038/nature19057>.
- León, A. and Pastor, Ó. (2021) 'Enhancing Precision Medicine: A Big Data-Driven Approach for the Management of Genomic Data', *Big Data Research*, 26, p. 100253. Available at: <https://doi.org/10.1016/j.bdr.2021.100253>.
- Liu, X. *et al.* (2020) 'dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs', *Genome Medicine*, 12(1), pp. 1–8. Available at: <https://doi.org/10.1186/s13073-020-00803-9>.
- Liu, X., Jian, X. and Boerwinkle, E. (2011) 'dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions', *Human Mutation*, 32(8), pp. 894–899. Available at: <https://doi.org/10.1002/humu.21517>.
- Liu, Y. *et al.* (2022) 'Computational approaches for predicting variant impact: An overview from resources, principles to applications', *Frontiers in Genetics*. Frontiers Media S.A. Available at: <https://doi.org/10.3389/fgene.2022.981005>.
- Livesey, B.J. and Marsh, J.A. (2020) 'Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations', *Molecular Systems Biology*, 16(7), pp. 1–12. Available at: <https://doi.org/10.15252/msb.20199380>.
- Livesey, B.J. and Marsh, J.A. (2022) 'Interpreting protein variant effects with computational predictors and deep mutational scanning', *DMM Disease Models and Mechanisms*, 15(6). Available at: <https://doi.org/10.1242/DMM.049510>.
- Livesey, B.J. and Marsh, J.A. (2023) 'Updated benchmarking of variant effect predictors using deep mutational scanning', *Molecular Systems Biology*, 44, pp. 1–15. Available at: <https://doi.org/10.15252/msb.202211474>.

- López-Ferrando, V. *et al.* (2017) 'PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update', *Nucleic Acids Research*, 45(W1), pp. W222–W228. Available at: <https://doi.org/10.1093/nar/gkx313>.
- Luo, Y. *et al.* (2021) 'ECNet is an evolutionary context-integrated deep learning framework for protein engineering', *Nature Communications*, 12(1). Available at: <https://doi.org/10.1038/s41467-021-25976-8>.
- MacArthur, D.G. *et al.* (2014) 'Guidelines for investigating causality of sequence variants in human disease', *Nature*, 508(7497), pp. 469–476. Available at: <https://doi.org/10.1038/nature13127>.
- Maceachern, S.J. and Forkert, N.D. (2021) 'Machine learning for precision medicine', *Genome*, 64(4), pp. 416–425. Available at: <https://doi.org/10.1139/gen-2020-0131>.
- Malley, J.D., Malley, K.G. and Pajevic, S. (2011) *Statistical learning for biomedical data*. Cambridge University Press.
- Martínez-García, M. and Hernández-Lemus, E. (2022) 'Data Integration Challenges for Machine Learning in Precision Medicine', *Frontiers in Medicine*, 8(January), pp. 1–21. Available at: <https://doi.org/10.3389/fmed.2021.784455>.
- Masica, D.L. *et al.* (2012) 'Phenotype-optimized sequence ensembles substantially improve prediction of disease-causing mutation in cystic fibrosis', *Human Mutation*, 33(8), pp. 1267–1274. Available at: <https://doi.org/10.1002/humu.22110>.
- Masica, D.L. and Karchin, R. (2016) 'Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants', *PLOS Computational Biology*, 12(5), p. e1004725. Available at: <https://doi.org/10.1371/journal.pcbi.1004725>.
- Matreyek, K.A. *et al.* (2018) 'Multiplex assessment of protein variant abundance by massively parallel sequencing', *Nature Genetics*, 50(6), pp. 874–882. Available at: <https://doi.org/10.1038/s41588-018-0122-z>.

- McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), pp. 1–14. Available at: <https://doi.org/10.1186/s13059-016-0974-4>.
- McLaughlin, R.N. *et al.* (2012) 'The spatial architecture of protein function and adaptation', *Nature*, 491(7422), pp. 138–142. Available at: <https://doi.org/10.1038/nature11500>.
- Mehlhoff, J.D. and Ostermeier, M. (2020) 'Biological fitness landscapes by deep mutational scanning', in *Methods in Enzymology*. Academic Press Inc., pp. 203–224. Available at: <https://doi.org/10.1016/bs.mie.2020.04.023>.
- Miller, M. and Kumar, S. (2001) 'Understanding human disease mutations through the use of interspecific genetic variation.', *Human Molecular Genetics*, 10(21), pp. 2319–28. Available at: <https://doi.org/10.1093/hmg/10.21.2319>
- Minton, A.P. (2000) 'Implications of macromolecular crowding for protein assembly', *Current Opinion in Structural Biology*, 10(1), pp. 34–39. Available at: [https://doi.org/10.1016/S0959-440X\(99\)00045-7](https://doi.org/10.1016/S0959-440X(99)00045-7)
- Miyazawa, S. and Jernigan, R.L. (1996) 'Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading', *Journal of Molecular Biology*, 256(3), pp. 623–644. Available at: <https://doi.org/10.1006/jmbi.1996.0114>
- Moynahan, M.E. *et al.* (1999) 'Brca1 controls homology-directed DNA repair', *Molecular Cell*, 4(4), pp. 511–518. Available at: [https://doi.org/10.1016/S1097-2765\(00\)80202-6](https://doi.org/10.1016/S1097-2765(00)80202-6).
- Musolf, A.M. *et al.* (2022) 'What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics', *Human Genetics*, 141(9), pp. 1515–1528. Available at: <https://doi.org/10.1007/s00439-021-02402-z>.
- Navío, D. *et al.* (2019) 'Structural and computational characterization of disease-related mutations involved in protein-protein interfaces', *International Journal of Molecular Sciences*, 20(7). Available at: <https://doi.org/10.3390/ijms20071583>.

- Ng, P.C. and Henikoff, S. (2001) 'Predicting deleterious amino acid substitutions', *Genome Research*, 11(5), pp. 863–874. Available at: <https://doi.org/10.1101/gr.176601>.
- Ng, P.C. and Henikoff, S. (2006) 'Predicting the effects of amino acid substitutions on protein function', *Annual Review of Genomics and Human Genetics*, 7, pp. 61–80. Available at: <https://doi.org/10.1146/annurev.genom.7.080505.115630>.
- Nguyen, V.L. *et al.* (2021) 'Comparison of multivariate linear regression and a machine learning algorithm developed for prediction of precision warfarin dosing in a Korean population', *Journal of Thrombosis and Haemostasis*, 19(7), pp. 1676–1686. Available at: <https://doi.org/10.1111/jth.15318>.
- Nicora, G. *et al.* (2022) 'A machine learning approach based on ACMG / AMP guidelines for genomic variant classification and prioritization American Association of Molecular Pathology', *Scientific Reports*, 12(1), pp. 1–12. Available at: <https://doi.org/10.1038/s41598-022-06547-3>
- Niroula, A., Urolagin, S. and Vihinen, M. (2015) 'PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants', *Plos One*, 10(2), p. e0117380. Available at: <https://doi.org/10.1371/journal.pone.0117380>.
- Niroula, A. and Vihinen, M. (2016) 'Variation Interpretation Predictors: Principles, Types, Performance, and Choice', *Human Mutation*, 37(6), pp. 579–597. Available at: <https://doi.org/10.1002/humu.22987>.
- Niroula, A. and Vihinen, M. (2017) 'Predicting Severity of Disease-Causing Variants', *Human Mutation*, 38, pp. 357–364. Available at: <https://doi.org/10.1002/humu.23173>
- Özkan, S. *et al.* (2021) 'The computational approach to variant interpretation: principles, results, and applicability', in C. Lázaro, J. Lerner-Ellis, and A. Spurdle (eds) *Clinical DNA Variant Interpretation: Theory and Practice*. San Diego: Elsevier Inc./Academic Press, pp. 89–119. Available at: <https://doi.org/10.1016/B978-0-12-820519-8.00007-7>

- Özkan, S., Padilla, N. and de la Cruz, X. (2021) 'Towards a new, endophenotype-based strategy for pathogenicity prediction in *brca1* and *brca2*: In silico modeling of the outcome of HDR/SGE assays for missense variants', *International Journal of Molecular Sciences*, 22(12), p. 6226. Available at: <https://doi.org/10.3390/ijms22126226>.
- Padilla, N. *et al.* (2019) 'BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge', *Human Mutation*, pp. 1–47. Available at: <https://doi.org/10.1002/humu.23802>.
- Parsons, M.T. *et al.* (2019) 'Large scale multifactorial likelihood quantitative analysis of BRCA1 and BRCA2 variants: An ENIGMA resource to support clinical variant classification', *Human Mutation*, 40(9), pp. 1557–1578. Available at: <https://doi.org/10.1002/humu.23818>.
- Patrinos, G.P. (2021) *Clinical DNA Variant Interpretation: Theory and Practice*. Academic Press.
- Patwardhan, R.P. *et al.* (2009) 'High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis', *Nature Biotechnology*, 27(12), pp. 1173–1175. Available at: <https://doi.org/10.1038/nbt.1589>.
- Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: <https://doi.org/10.1016/j.molcel.2012.08.019>.
- Pejaver, V. *et al.* (2022) 'Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria', *American Journal of Human Genetics*, 109(12), pp. 2163–2177. Available at: <https://doi.org/10.1016/j.ajhg.2022.10.013>.
- Pepe, M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.
- Petch, J., Di, S. and Nelson, W. (2022) 'Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology', *Canadian Journal of Cardiology*, 38(2), pp. 204–213. Available at: <https://doi.org/10.1016/j.cjca.2021.09.004>.

- Petrosino, M. *et al.* (2021) 'Analysis and interpretation of the impact of missense variants in cancer', *International Journal of Molecular Sciences*, 22(11). Available at: <https://doi.org/10.3390/ijms22115416>.
- Petukh, M., Kucukkal, T.G. and Alexov, E. (2015) 'On human disease-causing amino acid variants: Statistical study of sequence and structural patterns', *Human Mutation*, 36(5), pp. 524–534. Available at: <https://doi.org/10.1002/humu.22770>.
- Pires, D.E.V., Ascher, D.B. and Blundell, T.L. (2014) 'MCSM: Predicting the effects of mutations in proteins using graph-based signatures', *Bioinformatics*, 30(3), pp. 335–342. Available at: <https://doi.org/10.1093/bioinformatics/btt691>.
- Qorri, E. *et al.* (2022) 'A Comprehensive Evaluation of the Performance of Prediction Algorithms on Clinically Relevant Missense Variants', *International Journal of Molecular Sciences*, 23(14). Available at: <https://doi.org/10.3390/ijms23147946>.
- Raraigh, K.S. *et al.* (2018) 'Functional Assays Are Essential for Interpretation of Missense Variants Associated with Variable Expressivity', *American Journal of Human Genetics*, 102(6), pp. 1062–1077. Available at: <https://doi.org/10.1016/j.ajhg.2018.04.003>.
- Ray, P.P., Dash, D. and De, D. (2019) 'A Systematic Review and Implementation of IoT-Based Pervasive Sensor-Enabled Tracking System for Dementia Patients', *Journal of Medical Systems*. Springer New York LLC. Available at: <https://doi.org/10.1007/s10916-019-1417-z>.
- Reeb, J., Wirth, T. and Rost, B. (2020) 'Variant effect predictions capture some aspects of deep mutational scanning experiments', *BMC Bioinformatics*, 21(1), pp. 1–12. Available at: <https://doi.org/10.1186/s12859-020-3439-4>.
- Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genetics in Medicine*, 17(5), pp. 405–424. Available at: <https://doi.org/10.1038/gim.2015.30>

- Richardson, M.E. *et al.* (2021) 'Strong functional data for pathogenicity or neutrality classify BRCA2 DNA-binding-domain variants of uncertain significance', *American Journal of Human Genetics*, 108(3), pp. 458–468. Available at: <https://doi.org/10.1016/j.ajhg.2021.02.005>.
- Riera, C. *et al.* (2015) 'Molecular damage in Fabry disease: Characterization and prediction of alpha-galactosidase A pathological mutations', *Proteins: Structure, Function and Bioinformatics*, 83(1), pp. 91–104. Available at: <https://doi.org/10.1002/prot.24708>.
- Riera, C., Lois, S. and De la Cruz, X. (2014) 'Prediction of pathological mutations in proteins: The challenge of integrating sequence conservation and structure stability principles', *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3), pp. 249–268. Available at: <https://doi.org/10.1002/wcms.1170>.
- Riera, C., Padilla, N. and de la Cruz, X. (2016) 'The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions', *Human Mutation*, 37(10), pp. 1013–1024. Available at: <https://doi.org/10.1002/humu.23048>.
- Riesselman, A.J., Ingraham, J.B. and Marks, D.S. (2018) 'Deep generative models of genetic variation capture the effects of mutations', *Nature Methods*, 15(10), pp. 816–822. Available at: <https://doi.org/10.1038/s41592-018-0138-4>.
- Rivas, G. and Minton, A.P. (2016) 'Macromolecular Crowding In Vitro, In Vivo, and In Between', *Trends in Biochemical Sciences*, 41(11), pp. 970–981. Available at: <https://doi.org/10.1016/j.tibs.2016.08.013>.
- Roy, R., Chun, J. and Powell, S.N. (2012) 'BRCA1 and BRCA2: different roles in a common pathway of genome protection', *Nature Reviews Cancer*, 12(1), p. 68. Available at: <https://doi.org/10.1038/nrc3181>.
- Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, 1(5), pp. 206–215. Available at: <https://doi.org/10.1038/s42256-019-0048-x>.

- Sadowski, C.E. *et al.* (2017) 'BRCA1/2 missense mutations and the value of in-silico analyses', *European Journal of Medical Genetics*, 60(11), pp. 572–577. Available at: <https://doi.org/10.1016/j.ejmg.2017.08.005>.
- Sagi, O. and Rokach, L. (2018) 'Ensemble learning: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), pp. 1–18. Available at: <https://doi.org/10.1002/widm.1249>.
- Sahni, N. *et al.* (2015) 'Widespread macromolecular interaction perturbations in human genetic disorders', *Cell*, 161(3), pp. 647–660. Available at: <https://doi.org/10.1016/j.cell.2015.04.013>.
- Sanavia, T. *et al.* (2020) 'Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine', *Computational and Structural Biotechnology Journal*, 18, pp. 1968–1979. Available at: <https://doi.org/10.1016/j.csbj.2020.07.011>.
- Sánchez, I.E. *et al.* (2006) 'Point Mutations in Protein Globular Domains: Contributions from Function, Stability and Misfolding', *Journal of Molecular Biology*, 363(2), pp. 422–432. Available at: <https://doi.org/10.1016/j.jmb.2006.08.020>.
- Sarkar, A., Yang, Y. and Vihinen, M. (2020) 'Variation benchmark datasets: Update, criteria, quality and applications', *Database*, 2020, pp. 1–16. Available at: <https://doi.org/10.1093/database/baz117>.
- Schmidt, A. *et al.* (2023) 'Predicting the pathogenicity of missense variants using features derived from AlphaFold2', *Bioinformatics*, 39(5), pp. 1–7. Available at: <https://doi.org/10.1093/bioinformatics/btad280>.
- Schuster-Böckler, B. and Bateman, A. (2008) 'Protein interactions in human genetic diseases', *Genome Biology*, 9(1), p. R9. Available at: <https://doi.org/10.1186/gb-2008-9-1-r9>
- Schwarz, J.M. *et al.* (2014) 'Mutationtaster2: Mutation prediction for the deep-sequencing age', *Nature Methods*, 11(4), pp. 361–362. Available at: <https://doi.org/10.1038/nmeth.2890>.

- Shendure, J., Findlay, G.M. and Snyder, M.W. (2019) 'Genomic Medicine—Progress, Pitfalls, and Promise', *Cell*, 177(1), pp. 45–57. Available at: <https://doi.org/10.1016/j.cell.2019.02.003>.
- Shihab, H.A. *et al.* (2013) 'Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models', *Human Mutation*, 34(1), pp. 57–65. Available at: <https://doi.org/10.1002/humu.22225>.
- Shuyan, L. *et al.* (2010) 'Loss of post-translational modification sites in disease', *Biocomputing*, pp. 337–347.
- So, M.K. *et al.* (2019) 'Reinterpretation of BRCA1 and BRCA2 variants of uncertain significance in patients with hereditary breast/ovarian cancer using the ACMG/AMP 2015 guidelines', *Breast Cancer*, 26(4), pp. 510–519. Available at: <https://doi.org/10.1007/s12282-019-00951-w>.
- Sontag, E.M., Samant, R.S. and Frydman, J. (2017) 'Mechanisms and Functions of Spatial Protein Quality Control', *Annual Review of Biochemistry*, 86, pp. 97–122.
- Sruthi, C.K. and Prakash, M. (2020) 'Deep2Full: Evaluating strategies for selecting the minimal mutational experiments for optimal computational predictions of deep mutational scan outcomes', *PLoS ONE*, 15(1), pp. 1–18. Available at: <https://doi.org/10.1371/JOURNAL.PONE.0227621>.
- Sruthi, C.K. and Prakash, M.K. (2021) 'Disentangling the Contribution of Each Descriptive Characteristic of Every Single Mutation to Its Functional Effects', *Journal of Chemical Information and Modeling*, 61(4). Available at: <https://doi.org/10.1021/acs.jcim.0c01223>.
- Starita, L.M. *et al.* (2015) 'Massively parallel functional analysis of BRCA1 RING domain variants', *Genetics*, 200(2), pp. 413–422. Available at: <https://doi.org/10.1534/genetics.115.175802>.
- Starita, L.M. *et al.* (2017) 'Variant Interpretation: Functional Assays to the Rescue', *American Journal of Human Genetics*, 101(3), pp. 315–325. Available at: <https://doi.org/10.1016/j.ajhg.2017.07.014>.

- Stefl, S. *et al.* (2013) 'Molecular mechanisms of disease-causing missense mutations', *Journal of Molecular Biology*, 425(21), pp. 3919–3936. Available at: <https://doi.org/10.1016/j.jmb.2013.07.014>.
- Stein, A. *et al.* (2019) 'Biophysical and Mechanistic Models for Disease-Causing Protein Variants', *Trends in Biochemical Sciences*, 44(7), pp. 575–588. Available at: <https://doi.org/10.1016/j.tibs.2019.01.003>
- Steinbrecher, T. *et al.* (2017) 'Free Energy Perturbation Calculations of the Thermodynamics of Protein Side-Chain Mutations', *Journal of Molecular Biology*, 429(7), pp. 923–929. Available at: <https://doi.org/10.1016/j.jmb.2017.03.002>
- Stenson, P.D. *et al.* (2012) 'The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution', *Current protocols in bioinformatics*, 39(1), pp. 1–13. Available at: <https://doi.org/10.1002/0471250953.bi0113s39>.
- Steward, R.E. *et al.* (2003) 'Molecular basis of inherited diseases: a structural perspective', *Trends in Genetics*, 19(9), pp. 505–513. Available at: [https://doi.org/10.1016/S0168-9525\(03\)00195-1](https://doi.org/10.1016/S0168-9525(03)00195-1).
- Su, R. *et al.* (2020) 'Meta-GDBP: A high-level stacked regression model to improve anticancer drug response prediction', *Briefings in Bioinformatics*, 21(3), pp. 996–1005. Available at: <https://doi.org/10.1093/bib/bbz022>.
- Subramanian, M. *et al.* (2020) 'Precision medicine in the era of artificial intelligence: implications in chronic disease management', *Journal of Translational Medicine*, 18(1), pp. 1–12. Available at: <https://doi.org/10.1186/s12967-020-02658-5>.
- Sun, H. and Yu, G. (2019) 'New insights into the pathogenicity of non-synonymous variants through multi-level analysis', *Scientific Reports*, 9(1), pp. 1–11. Available at: <https://doi.org/10.1038/s41598-018-38189-9>.
- Sundaram, L. *et al.* (2018) 'Predicting the clinical impact of human mutation with deep neural networks', 50(August), pp. 1161–1170. Available at: <https://doi.org/10.1038/s41588-018-0167-z>.

- Sunyaev, S., Ramensky, V. and Bork, P. (2000) 'Towards a structural basis of human non-synonymous single nucleotide polymorphisms', *Trends in Genetics*, 16(5), pp. 198–200. Available at: [https://doi.org/10.1016/S0168-9525\(00\)01988-0](https://doi.org/10.1016/S0168-9525(00)01988-0).
- Suzek, B.E. *et al.* (2007) 'UniRef: Comprehensive and non-redundant UniProt reference clusters', *Bioinformatics*, 23(10), pp. 1282–1288. Available at: <https://doi.org/10.1093/bioinformatics/btm098>.
- Suzek, B.E. *et al.* (2015) 'UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches', *Bioinformatics*, 31(6), pp. 926–932. Available at: <https://doi.org/10.1093/bioinformatics/btu739>.
- Tabet, D. *et al.* (2022) 'Scalable Functional Assays for the Interpretation of Human Genetic Variation', *Annual Review of Genetics*, 56, pp. 441–465. Available at: <https://doi.org/10.1146/annurev-genet-072920-032107>.
- Tang, H. and Thomas, P.D. (2016) 'Tools for predicting the functional impact of nonsynonymous genetic variation', *Genetics*, 203(2), pp. 635–647. Available at: <https://doi.org/10.1534/genetics.116.190033>.
- Tavtigian, S. V. *et al.* (2006) 'Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral', *Journal of Medical Genetics*, 43(4), pp. 295–305. Available at: <https://doi.org/10.1136/jmg.2005.033878>.
- Tavtigian, S. V *et al.* (2008) 'In Silico Analysis of Missense Substitutions Using Sequence-Alignment Based Methods', 29(11), pp. 1327–1336. Available at: <https://doi.org/10.1002/humu.20892>.
- Tian, Y. *et al.* (2019) 'REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification', *Scientific Reports*, 9(1), pp. 1–6. Available at: <https://doi.org/10.1038/s41598-019-49224-8>.
- Towler, W.I. *et al.* (2013) 'Analysis of BRCA1 Variants in Double-Strand Break Repair by Homologous Recombination and Single-Strand Annealing', *Human Mutation*, 34(3), pp. 439–445. Available at: <https://doi.org/10.1002/humu.22251>.

- Lo Vercio, L. *et al.* (2020) 'Supervised machine learning tools: A tutorial for clinicians', *Journal of Neural Engineering*, 17(6). Available at: <https://doi.org/10.1088/1741-2552/abbff2>.
- Vihinen, M. (2012) 'How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis.', *BMC genomics*, 13 (4), pp. 1-10. Available at: <https://doi.org/10.1186/1471-2164-13-S4-S2>.
- Vihinen, M. (2014) 'Majority vote and other problems when using computational tools', *Human Mutation*, 35(8), pp. 912–914. Available at: <https://doi.org/10.1002/humu.22600>.
- Vihinen, M. (2020) 'Problems in variation interpretation guidelines and in their implementation in computational tools', *Molecular Genetics & Genomic Medicine*, 8, p. e1206. Available at: <https://doi.org/10.1002/mgg3.1206>
- Vihinen, M. (2021) 'Functional effects of protein variants', *Biochimie*, 180, pp. 104–120. Available at: <https://doi.org/10.1016/j.biochi.2020.10.009>.
- Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods*, 17(3), pp. 261–272. Available at: <https://doi.org/10.1038/s41592-019-0686-2>.
- De Visser, J.A.G.M. and Krug, J. (2014) 'Empirical fitness landscapes and the predictability of evolution', *Nature Reviews Genetics*, 15(7), pp. 480–490. Available at: <https://doi.org/10.1038/nrg3744>.
- Wainberg, M. *et al.* (2018) 'Deep learning in biomedicine', *Nature Biotechnology*, 36(9), pp. 829–838. Available at: <https://doi.org/10.1038/nbt.4233>.
- Walsh, I. *et al.* (2014) 'PASTA 2.0: An improved server for protein aggregation prediction', *Nucleic Acids Research*, 42, pp. W301–W307.
- Wang, G. and Dunbrack, R.L. (2003) 'PISCES: A protein sequence culling server', *Bioinformatics*, 19(12), pp. 1589–1591. Available at: <https://doi.org/10.1093/bioinformatics/btg224>.

- Wang, Z. and Moulton, J. (2001) 'SNPs, protein structure, and disease', *Human Mutation*, 17(4), pp. 263–270. Available at: <https://doi.org/10.1002/humu.22>.
- Weile, J. *et al.* (2017) 'A framework for exhaustively mapping functional missense variants', *Molecular Systems Biology*, 13(12), p. 957. Available at: <https://doi.org/10.15252/msb.20177908>.
- Wikipedia (2023) *BLOSUM*. Available at: <https://en.wikipedia.org/wiki/BLOSUM#> (Accessed: 12 July 2023).
- Wilcox, E.H. *et al.* (2022) 'Evaluating the impact of in silico predictors on clinical variant classification', *Genetics in Medicine*, 24(4), pp. 924–930. Available at: <https://doi.org/10.1016/j.gim.2021.11.018>.
- Winter, J.C.F. De, Gosling, S.D. and Potter, J. (2016) 'Comparing the Pearson and Spearman Correlation Coefficients Across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data', *Psychological Methods*, 21(3), pp. 273–290. Available at: <https://doi.org/10.1037/met0000079.supp>.
- Wong, A.K. *et al.* (2021) 'Decoding disease: from genomes to networks to phenotypes', *Nature Reviews Genetics*, 22(12), pp. 774–790. Available at: <https://doi.org/10.1038/s41576-021-00389-x>.
- Wu, Y. *et al.* (2019) 'A web application and service for imputing and visualizing missense variant effect maps', *Bioinformatics*, 35(17), pp. 3191–3193. Available at: <https://doi.org/10.1093/bioinformatics/btz012>.
- Xu, J. *et al.* (2019) 'Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives', *Human Genetics*, 138(2), pp. 109–124. Available at: <https://doi.org/10.1007/s00439-019-01970-5>.
- Yang, Y. *et al.* (2018) 'PON-tstab: Protein variant stability predictor. Importance of training data quality', *International Journal of Molecular Sciences*, 19(4), p. e1009.
- Yazar, M. and Özbek, P. (2021) 'In Silico Tools and Approaches for the Prediction of Functional and Structural Effects of Single-Nucleotide Polymorphisms on Proteins: An Expert Review',

- OMICS *A Journal of Integrative Biology*, 25(1), pp. 23–37.
Available at: <https://doi.org/10.1089/omi.2020.0141>.
- Yeger-Lotem, E. and Sharan, R. (2015) 'Human protein interaction networks across tissues and diseases', *Frontiers in Genetics*, 6(Aug), pp. 1–5. Available at: <https://doi.org/10.3389/fgene.2015.00257>.
- Yin, S., Ding, F. and Dokholyan, N. V. (2007) 'Eris: An automated estimator of protein stability [2]', *Nature Methods*, 4(6), pp. 466–467. Available at: <https://doi.org/10.1038/nmeth0607-466>.
- Yu, K.H., Beam, A.L. and Kohane, I.S. (2018) 'Artificial intelligence in healthcare', *Nature Biomedical Engineering*, 2(10), pp. 719–731. Available at: <https://doi.org/10.1038/s41551-018-0305-z>.
- Yue, P., Li, Z. and Moulton, J. (2005) 'Loss of protein structure stability as a major causative factor in monogenic disease', *Journal of Molecular Biology*, 353(2), pp. 459–473.
- Yue, P. and Moulton, J. (2006) 'Identification and analysis of deleterious human SNPs', *Journal of Molecular Biology*, 356(5), pp. 1263–1274. Available at: <https://doi.org/10.1016/j.jmb.2005.12.025>.
- Zhou, Y. *et al.* (2019) 'An optimized prediction framework to assess the functional impact of pharmacogenetic variants', *Pharmacogenomics Journal*, 19(2), pp. 115–126. Available at: <https://doi.org/10.1038/s41397-018-0044-2>.
- Zitnik, M. and Leskovec, J. (2017) 'Predicting multicellular function through multi-layer tissue networks', *Bioinformatics*, 33(14), pp. i190–i198. Available at: <https://doi.org/10.1093/bioinformatics/btx252>.

