IN SILICO MISSENSE CLASSIFICATION

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

ROHAN GNANAOLIVU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISOR: DR. STEVEN N HART

CO-ADVISOR: DR. CHAD L MYERS

October: 2025

Acknowledgements

The journey through the post graduate experience with the Bioinformatics and Computational Biology (BICB) from the University of Minnesota for the past seven years has been an extraordinary adventure filled with intellectual challenges, moments of discovery, and profound personal growth. This experience, marked by both triumphs and setbacks, has been enriched by extensive learning through rigorous coursework, independent study, and journal clubs from fellow students, faculty, and visiting researchers.

My scientific aspirations took root during my childhood in Chennai, India, where I sat among 70-80 students in classrooms exceeding 100°F, reading about PhD scientists from weathered, half-torn textbooks. Learning about their discoveries that shaped human understanding had sparked my desire to one day earn a PhD and contribute to the scientific community. The opportunity to pursue doctoral studies in computational biology through the BICB program represents the fulfillment of that dream. I am immensely grateful to the Quantitative Health Sciences (QHS) division at Mayo Clinic for providing employment support throughout this PhD journey, and to the Professional Development Assistance Program (PDAP) at Mayo Clinic for easing the financial burdens of graduate education. I am grateful to Dr. Yuk Sham whose early guidance as program director helped me establish a strategic plan for completing my coursework requirements.

I am deeply indebted to Dr. Steven Hart, who served as my academic advisor over the past seven years. Dr. Hart expertise on missense mutations and predictive modeling is second to none, and his own ability to constantly evolve and gain new knowledge in areas of machine learning and generative Artificial Intelligence has been an inspiration to me. I am deeply appreciative of his unwavering commitment to pushing me beyond my comfort zone, fostering the skills essential for academic research excellence, and providing the support, knowledge, and guidance that have shaped my growth as a scientist. His mentorship has taught me to think critically, formulate meaningful research questions, and develop the analytical rigor necessary for a successful career in Computational Biology.

I would like to express my sincere gratitude to my thesis committee members Dr. Chad Myers, Dr. Yuk Cham, and Dr. Hai Dang Nguyen, whose thoughtful guidance and constructive feedback have significantly enhanced the rigor and quality of my research. Their invaluable insights and expertise have been instrumental in shaping the methodological approaches and key findings presented throughout this work.

I am also grateful to the Principal Investigators and their laboratory teams that have educated me during my academic journey, providing me with essential knowledge and research experience that helped me find my scientific interests and navigate this PhD program. I would particularly like to acknowledge Dr. Chen Wang, Dr. Fergus Couch, Dr. Chunling Hu, and Dr. Nicolas Boddicker, whose education and collaboration have influenced my development as a computational biologist and researcher.

Finally, I am truly grateful to my wife Sarah and our cats, who have been my home and constant sources of comfort during countless long hours of evening and late-night work. Their presence and companionship were a great source of motivation that has helped me in this journey.

Dedication

Dedicated to all my family and friends, whose unwavering support, love and inspiration have fueled my lifelong pursuit of knowledge.

I dedicate my doctoral dissertation to all the people that face physical disabilities. Having lived most of my life as an amputee, I know how pursuing your dreams can feel like chasing distant memories. My hope is that this work serves as proof that our limitations need not define our destination.

Abstract

Clinical interpretation of missense mutations remains a significant challenge, often leading to their classification as Variants of Uncertain Significance (VUS). The American College of Medical Genetics and Genomics (ACMG) provide guidelines for clinicians utilizing *in silico* tools to classify variants as benign or pathogenic based on their predicted impact on protein function. However, these tools face several limitations, such as, learning from the stereochemical properties from incomplete 3D protein structures, overfitting to ClinVar and HGMD databases, and poor modeling of intrinsically disordered regions (IDRs), which harbor over 25% of disease-associated mutations. Unlike structured regions, IDRs lack stable secondary and tertiary structures, are rich in polar residues, and cannot be crystallized, making them difficult to assess using conventional predictors reliant on structural features. Many existing models incorporate stereochemical properties from incomplete 3D structures and fail to accurately predict the effects of mutations in IDRs, leading to reduced classification accuracy. There remains a need to improve the prediction of *in silico* missense classification, especially for variants in IDRs.

In Chapter 1, I provide an overview of the fundamental concepts underlying *in silico* missense variant classification, emphasizing its growing role in the clinical interpretation of genetic variation. This chapter outlines how computational methods utilize diverse biological features particularly those derived from protein structure, to predict whether a missense mutation is likely to be pathogenic or benign. Chapter 2 focuses on the application of AlphaFold2, a state-of-the-art protein structure prediction model, in assessing the effects of missense mutations within ordered (structured) regions of proteins. I evaluate the utility

of the change in Gibbs free energy calculated from AlphaFold2 predicted structures, as predictors of protein function. Additionally, this chapter benchmarks several leading *in silico* missense variant classifiers using high-quality functional assay data from four tumor suppressor genes. In Chapter 3, I address missense variant classification in IDRs, which lack stable structure and are poorly modeled by traditional predictors. Here, I introduce a novel computational framework that integrates protein language models, change in phase separation propensity and change in global IDR confirmation due to a missense variant to predict deleteriousness. This method is evaluated using clinically curated variant annotations from the ClinVar database. Finally, Chapter 4 summarizes the major findings, explores the clinical and translational significance of *in silico* missense classification. I discuss how the advancements presented here can contribute to more accurate and interpretable variant interpretations.

Table of Contents

List of Tables

List of Figures

Abbreviations

LOF: Loss of function

TSBC: Tumor Suppressor Breast Cancer genes

VUS: Variant of Uncertain Significance

AF2: AlphaFold2

PDB: Protein Data Bank

RSA: Relative Solvent Accessibility

MAVE: Multiplexed Assays for Variant Effects

ΔΔG: Change is Gibb's free energy

RMSD: Root Mean Square Deviation

GDT: Global Distance Test

AUC: Area under the curve

PLDDT: Predicted Local Distance Difference Test

MSA: Multiple Sequence Alignment

RSA: Relative Solvent Accessibility

CASP: Critical Assessment of Structure Prediction

IDR: Intrinsically Disordered Regions

HDR: Homologous DNA Repair

UCSC: University of California Santa Cruz

ACMG: American College of Medical Genetics and Genomics

HPO: Hyperparameter Optimization

MLP: Multi-Layer Perceptron

PLM: Protein Language Models

PS: Phase Separation

gIDRc: Change in global IDR conformation

AI: Artificial Intelligence

CI: Confidence Interval

CV: Cross Validation

AUC: Area under the curve

PR-AUC: Area under the precision-recall curve

WT: Wild-Type

ML: Machine Learning

Chapter 1 Fundamental concepts underlying in silico missense variant classification

1.1 Current knowledge of missense mutations

There are varying types of mutations that can occur that lead to disease, and missense mutation is one such mutation type that can lead to diseased phenotypes. Missense mutations represent a critical class of genetic variants that can significantly impact human health by altering protein structure and function. These single nucleotide variants occur within protein-coding regions and result in the substitution of one amino acid for another in the translated protein product. Missense mutations can disrupt protein function through multiple molecular mechanisms, including thermodynamic destabilization of protein folding, perturbation of active sites that impair enzymatic activity, disruption of protein-protein interactions essential for cellular processes, and interference with allosteric regulation, thereby rendering that mutation as deleterious[1, 2].

While missense mutations are abundant in human populations, with each individual carrying thousands of such variants, the vast majority are functionally neutral or benign, representing common polymorphisms that have little to no impact on protein function. Disease-causing missense mutations are relatively rare, typically occurring at population frequencies below 0.1%, and often require large-scale sequencing studies of affected cohorts to establish their pathogenic role. The challenge in identifying truly pathogenic

variants lies in distinguishing them from the background of benign variation, necessitating comprehensive functional studies, segregation analyses in affected families, and integration of computational prediction tools with experimental validation. This complexity has led to the accumulation of numerous variants of uncertain significance in clinical databases, highlighting the ongoing need for improved methods to accurately classify missense mutations and their potential role in human disease.

As sequencing of large cohorts is expensive, clinicians and researchers have relied on *in silico* prediction models in conjunction with family history, *in-vivo* functional assays, and frequency of observation in affected and unaffected individuals to diagnose a patient with a reported missense mutation. Fifty percent of all mutations reported in disease-implicated genes are missense mutations in the ClinVar[3] database, thereby highlighting the urgent need to classify them accurately.

1.2 *In silico* prediction of missense mutations

Traditional *in silico* predictors employ machine and deep learning techniques, including supervised and unsupervised approaches, and rely on features set based on sequence conservation, protein secondary structure and physicochemical properties to assess variant pathogenicity[4]. These predictors have traditionally been developed to predict the deleteriousness of a missense mutation. Missense predictors use features derived from sequence conservation across species (Align-GVGD, MutationAssessor[5], SIFT[6], and MutationTaster[7]), Epigenetics (CADD[8], DANN[9]), sequence context and expression (CADD[27], FitCons[37]), structural features from the protein (MutPred2[38], PolyPhen-2[30]), and amino acid properties (PolyPhen-2[39], VEST[40], CADD[25], Align-GVGD[23]). There are also

predictors that learn from the prediction of individual predictors, these types of predictors are referred to as meta-predictors (BRCA-ML[41], REVEL[42], ClinPred[43], BayesDel[44]). *In silico* missense predictors rely heavily on labelled data from mutational databases such as ClinVar[5], VariBench[45] and Human Gene Mutation Database (HGMD)[6]. These predictors also have published thresholds to classify mutations as deleterious versus neutral.

Different *in silico* predictors have their own strengths and weaknesses and can be disease specific predictors[46,47]. Predictors listed in (*Table 1)* has been shown to have high accuracy in predicting deleterious mutations in genes predisposed to breast cancer[48,49]. Each tool employs distinct training methodologies, testing frameworks, and optimization strategies, with many modern algorithms functioning as meta-predictors that integrate predictions from multiple individual tools. Meta-predictors, exemplified by REVEL, show substantially improved performance when gene-specific thresholds are applied. In REVEL's development, VEST and FATHMM emerged as the most influential features among an ensemble of 18 predictors for determining variant pathogenicity. VEST employs a random forest model utilizing 86 quantitative features derived from SNVbox[51]. The model uses mutational data from HGMD and ESP6500 [52]. FATHMM[50] uses multiple features in ten distinct groups (Conservation, epigenetics, DNA sequence context, GC content, transcription binding site) on mutations collected from the HGMD[6], VariBench[45], UniProt[53] and SwissVar[54]. Most Meta-predictors however suffer from the circularity issue, which is, same variants were used in the training and evaluation set[55]

Table 1: Top missense predictors to predict loss-of-function in breast cancer

Predictor	Features	Model	Training
AlphaMissense	Multiple Sequence Alignment	Transformer based	unsupervised
EVE	Multiple Sequence Alignment	Variational autoencoder	unsupervised
ESM1b	Multiple Sequence Alignment		unsupervised
REVEL	Meta predictor	Random Forest	ClinVar
BayesDel	Meta predictor	Bayes Classifier	ClinVar
MutationTaster	Conservation, secondary protein structure	Bayes Classifier	ClinVar
BRCA-ML	Meta predictor	XG-boost	Mave
VEST	SNVBox	Random Forest	ClinVar
FATHMM	Conservation, epigenetics, DNA sequence context, GC content, transcription binding site	Hidden Markov model	ClinVar
ClinPred	Meta predictor	Random Forest and GBM	ClinVar
MetaRNN	Meta predictor	RNN	ClinVar

In 2015, ACMG approved the use of *in silico* predictors as a form of support to classify

missense mutations[10]. However, these predictors are considered the weakest form of evidence compared to laboratory *in-vivo* functional assays or frequency of observation of the mutation in affected and unaffected individuals[11]. *In silico* predictors have several limitations, such as learning from the pathogenicity of other diseases which involve different mechanistic pathway, overfitting on databases such as ClinVar and HGMD[12] for testing and training, incapability of assessing gene dosage, and overfitting on stereochemical properties derived from partial and low resolutions gene structures. Existing predictors have used features derived from the secondary structure of existing crystalized structures in the Protein Data Bank (PDB), however these account for only 48% of the entire proteome and referred to as the ordered regions or the proteome[13]. Proteins also contain large regions that are disordered, also referred to as intrinsically disordered regions (IDRs). IDRs are largely ignored, as it contains large charged hydrophilic amino acids, which reduces the ability to crystalize. Other features include Sequence conversation, epigenomic factors and sequence context.

1.3 Intrinsically disordered regions

IDRs are protein segments that lack a stable secondary or tertiary structure and exist as dynamic ensembles. it is also known that change in protein structure due to a missense mutation is intimately linked to its stability[14, 15]. The change in structure directly impacts the stability of the entire protein, and this can be measured by estimating $\Delta\Delta G_f$ between the folding states of the mutant and its wild-type form[16]. Missense variants have been shown to cause protein misfolding and destabilization. Many publications have shown

that protein stability is directly associated with its function[17-19], with recent publication showing that gain or loss of energy has an impact protein function. However, stability is not the sole reason that impacts protein function and most stability prediction have been made on ordered structures. Most recently, change in $|\Delta\Delta G_f|$ caused by a missense variant may play a role in disease development when genes are haploinsufficient. $|\Delta\Delta G_f|$ was able to predict pathogenic variants at a precision (>96%)[20] when genes are haploinsufficient.

Determining experimental $\Delta\Delta G_f$ upon a mutation requires labor intensive experiments and capital. Due to this, many stability predictors have been developed, with FoldX recognized as the state-of-the-art tool for $\Delta\Delta G_f$ prediction. FoldX requires a highly accurate wild-type structure to make its prediction. Recent Critical Assessment of Structure Prediction (CASP) 13 and CASP14 challenges have highlighted the giant leap in advancement in protein structure prediction, with results showing predicted models having an accuracy like that of high-resolution crystal structures, with a focus on AF2[21, 22]. AF2 can predict complete models which may contain ordered regions and IDRs. A high proportion of TSBC proteins contains IDRs. Many publications have shown that AF2 quality metric pLDDT can be used as predictive measure to detect IDRs[23, 24]. The role of IDRs in cancer is not well understood.

In-vivo functional assays have often been relied upon to assess the effect of a mutation on protein function and have been used to classify a VUS[25]. ACMG have stated that *in-vivo* functional assays can provide a powerful insight on the effect of a mutation on a

protein[26]. However, performing these *in-vivo* functional assays on all possible missense mutations in TSBC genes would require a significant amount of capital, time, and labor. Recent advances in high throughput functional assays, also known as Multiplexed Assays of Mutation Effect (MAVEs) have provided the ability to screen thousands of mutations simultaneously, including those associated with breast cancer[27]. However, these screens are dependent on the availability of assays based on molecular mechanism of the cell and do not easily scale to the entire protein or across proteins in the same molecular pathway, and hence there is still a need to build novel sophisticated missense prediction algorithms to account for every possible missense variant.

1.4 Multiplex assays of mutation effect (MAVE)

MAVE's can be described as high throughput *in-vivo* functional assays that can systematically test all missense mutation present in a gene. MAVE technology can screen thousands of mutations and has been shown to perform better at predicting pathogenic mutations than *in silico* prediction methods[54]. The results from these high throughput experiments (in conjunction with other evidence) are reviewed by an expert panel, such as ClinGen[56] to classify a mutation as deleterious or neutral. However, there are several limitations, such as data sparsity[57,58], inability to test entire protein with reliable reproducibility, limited to type of available *in-vivo* functional assay based on molecular mechanism, and do not account for the variability across assays.

Recently published high throughput functional assays of missense mutations in BRCA1 and BRCA2 that measure homologous DNA repair (HDR) activity in the cell has enabled

gene specific models to use supervised Machine Learning (ML) and Deep learning methods to learn from features in high dimensionality space and make predictions across these genes[27,41,59]. These assays use a haploid cell line (HAP1 cells) that require functional BRCA1 and BRCA2 for cell survival. Similar assays that measure HDR activity in RAD51C, ATM, and PALB2 have provided a unique opportunity to evaluate *in silico* missense prediction model for these genes.

Table 2: Published functional mutations in breast cancer genes

Genes	Total Mutations	Assay
BRCA1	2086	saturation genome editing
BRCA1	1086	Homology-Directed Repair Assay
BRCA2	244	Mano-B
BRCA2	252	Homology-directed repair assay
PALB2	84	Homology-directed repair assay
RAD51C	174	Homology-directed repair assay

1.5 Protein structure predictors

Protein structure has come a long way since the initial predictions made by Pauling and Corey in 1951, where they predicted helical and sheet conformations. Since then, giant strides have been made in this field by multiple research groups namely Rosetta[65,66], Zhang lab[67] and DeepMind[68]. Rosetta employs a two-stage de-novo approach to make its protein predictions. First, it samples the amino acid sequence to create the fragment library to capture the local confirmational space and then it combines different fragments to create the folds by optimizing the energy function to find the global minima by employing a Metropolis Monte Carlo algorithm to generate a native protein conformation[66]. The Zhang lab utilizes a template-based approach that first queries the sequence to find similar evolutionary relatives to create the secondary structure, followed by querying the secondary structure to find representative PDB templates. After ranking the templates based on alignment, fragments are created by isolating aligned fragments from the template and using an *ab initio* model to build unaligned fragments. The full-length structure is then created using parallel tempering Monte Carlo simulations to reassemble the structure fragments[67]. DeepMind's AF2 is neural network model that has been trained on all the structures in the PDB. It uses both template and non-template approaches to create a native protein conformation. In the non-template approach, AF2 first uses the sequences to create a Multiple Sequence Alignment (MSA) matrix using highly similar sequences from other organisms and then creates a two-dimensional representation of every pairwise residue using induction bias. This is then fed into a transformer with attention called Evoformer as

it contains evolutionary information. An Evoformer is then used to exchange information between the MSA representation and the pair representation to build the 3D model using 48 iterations [68]. Our initial assessment of AF2 performance comparing with experimentally-derived BRCT domain isoforms located in BRCA1 (Figure 1) yielded low RMSD metrics, highlight strong performance AF2.

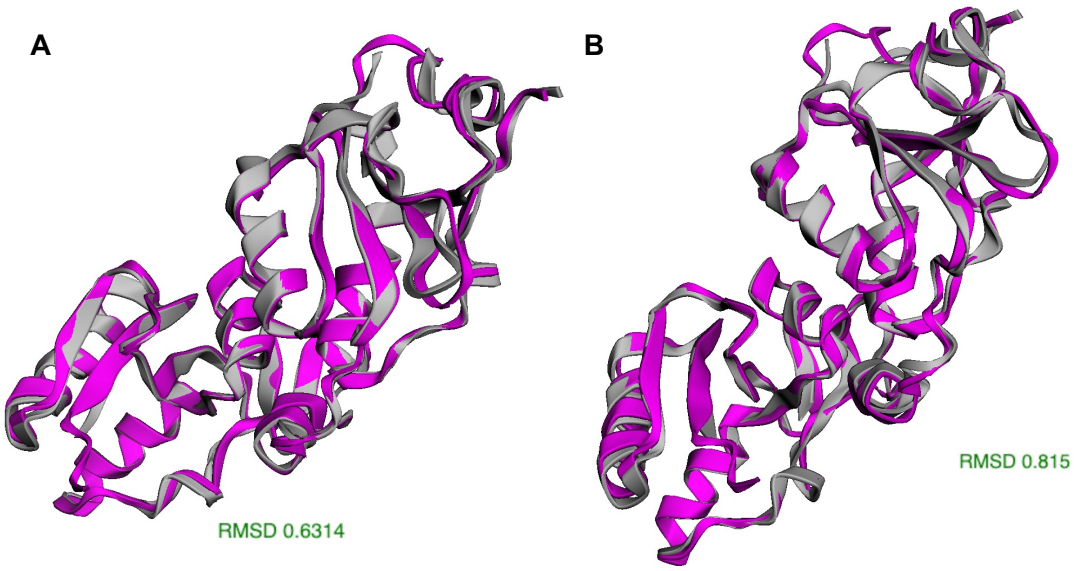

Figure 1: Validation of AlphaFold2 prediction using crystal structure Overlay*. **(A)** Superposition of the experimentally determined high-resolution crystal structure (PDB:1T15) with the AlphaFold2-predicted structure of the BRCT domains (amino acids 1646-1863) in BRCA1, demonstrating structural concordance. **(B)** Alignment between the crystal structure template (PDB:1T15) and AlphaFold2 structure of its top ranked prediction, validating the accuracy of* ab initio *protein structure prediction for this functionally important domain.*

CASP challenges aim at evaluating the accuracy of predicted protein structures submitted by multiple competing groups. The challenge aims at assessing the accuracy of atomic coordinate locations and distance with surrounding atoms. One of the standard metrics of accuracy in CASP is the Global Distance Test Total Score (GDT), which corresponds to the average percentage of connected C-alpha pairs[69,70]. The closer the GDT is to 100%, the

more accurate the backbone of a predicted model. GDT values above 80% denote accurate models and values below 20% denoting mostly random models. In 2018, in the CASP13 challenge, a huge jump in accuracy was seen due to the inclusion of deep-learning methods for contact prediction. For the first time, exceedingly difficult targets were modeled with an average GDT of 70% by DeepMind's AF method. In CASP14, AF2 made a further leap in model accuracy, with the best model reaching a GDT above 90%, a range at the level of experimental accuracy[15].

Recent publications have shown that generating mutant structures from AF2 do not show the ability to predict missense mutations by using the distance measure of c-alpha pairs[71]. However, what is unknown is whether using stability predictors from AF2 structures can predict loss-of-function activity in TSBC genes. The pLDDT scores generated by AF2 structures are in the range of (0, 100). High pLDDT scores (e.g., > 80) indicate high confidence of the residue structure, and low pLDDT scores (e.g., < 50) may indicate that the residues are in IDRs[69]. IDRs have no single defined tertiary structure in native conditions and are known to mediate crucial signaling processes and control pathways where high-specificity/low-affinity interactions with other compounds play a crucial role[73]. For this reason, we will use prediction scores from IUPRED3[74], pLDDT and RSA to identify IDRs and ANCHOR2[75] to find disorder binding regions[74], as IDRs have been shown to undergo oligomerization in TSBC proteins.

Using AF2, we created the BRCA1 predicted 3D protein structure, which comprises of 1863 amino acids. Currently, only the BRCT domains and ring domain exists as a high-resolution crystal structure in the PDB, which is only 213 and 117 amino acids, respectively. BRCA1 protein has been hypothesized to contain IDRs, which we observe in the generated complete AF2 monomer. To evaluate the accuracy, we generated AF2 structures using PDB:IT15 protein sequence in template and non-template mode and evaluated the RMSD with the high-resolution crystal structure using visualization tool chimera (*see Figure 1A* and *Figure 1B)* and found highly similar structures with RMSD 0.631 and 0.815, respectively.

1.6 Protein stability predictors

Protein stability predictors predict the gain or loss of energy of the protein upon presence of a mutation. Incorrect folding and decreased stability are one of the consequences of pathogenic missense mutations[9,10]. Protein stability perturbation are expressed as the change in Gibbs free energy between the mutant and wild-type protein[63]. As experimentally deriving Gibbs free energy of the unfolding state (ΔG) is labor intensive to generate for every possible missense mutation, many stability predictors were developed over the past decade[76–80]. Recent publications have also shown that protein stability predictors can be used to predict loss-of-function activity[57].

Structure based stability predictors use the energy features derived from the 3D protein structure, such as Van der Waals, electrostatics, solvent accessibility, etc. The most

universally recognized structure-based stability predictor is FoldX[63,81] ,as multiple studies have shown that FoldX prediction of $|\Delta\Delta G_f|$ from a crystalized structure is currently the most accurate method[13,57,71]. The core function of FoldX, is to calculate the change of ΔG in kcal mol^{-1}, which is the linear combination of the dot product of various energy terms with the relative weight [81,82]. FoldX uses a function called "RepairDB" to first reduce the energy content of a protein-structure model to a minimum by rearranging side chains. The function called "BuildModel" is used to introduce mutations and optimizes the structure of the new protein.

Recent publication showed that there was significant Pearson correlation coefficient between experimentally-derived $\Delta\Delta G_f$ from 24 BRCA1 mutation with the $|\Delta\Delta G_f|$ derived from the FoldX using high-resolution crystal structure 1JNXX[15]. We further evaluated the Pearson correlation coefficient with $\Delta\Delta G_f$ derived from FoldX using high-resolution crystal structure 4OFB and AF2 predicted structure of 4OFB amino acid sequence. We observe a Pearson correlation coefficient value of 0.93, with a significant p-value, showing that $\Delta\Delta G_f$ from FoldX derived from AF2 structures is highly comparable to that generated from high-resolution crystal structure (Figure 2).

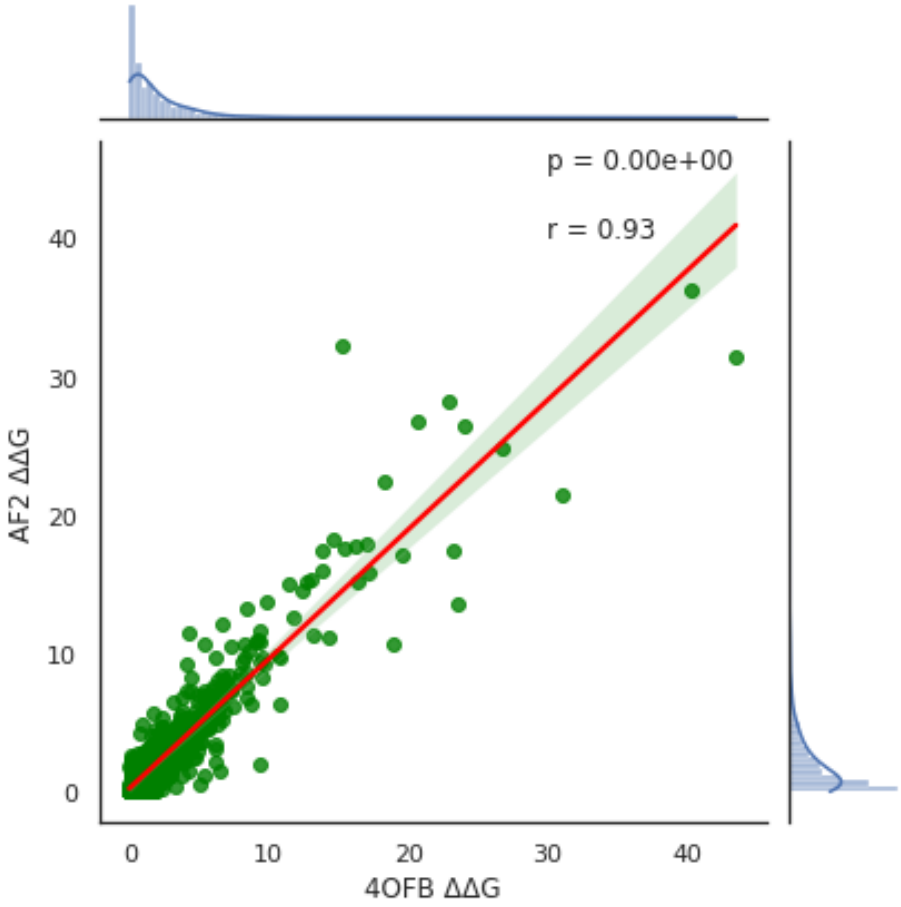

Figure 2: *Linear regression comparing ΔΔG values obtained from FoldX using experimentally-derived crystal structure PDB:4OFB versus AlphaFold2 prediction of 4OFB structure. The regression correlation coefficient (R), and P-value suggest strong correlation between the predicted ΔΔG values.*

1.7 Outline of Chapters

My dissertation research addresses the limitations of *in silico* missense prediction by incorporating generative AI methods to enhance the functional classification of missense variants. The research moves beyond current constraints by examining how generative AI models can contribute to the functional classification of missense variants. The study draws upon experimentally-derived proteins, generative AI models, functionally classified variants from MAVE assays, and classified missense variants from the ClinVar database.

Each chapter presents a methodology that contributes to enhanced accuracy in *in silico* missense prediction.

Chapter 2 examines the application of protein structures created through generative AI technology to analyze protein perturbation resulting from missense variants and their relationship to function. This analysis draws upon functionally validated MAVE assays across four tumor suppressor breast cancer genes: BRCA1, BRCA2, PALB2, and RAD51C. The chapter also investigates the influence of changes in Gibbs energy as a feature for explaining function, comparing this approach to purpose-built missense classification prediction models.

Chapter 3 presents a novel methodology that enhances *in silico* missense classification prediction by improving classification accuracy for variants located in predicted intrinsically disordered regions. The chapter demonstrates the application of protein language models and novel features that extend unsupervised *in silico* missense predictors, resulting in notable improvements in classification accuracy.

Chapter 4 provides a comprehensive summary of the research on generative AI applications in *in silico* missense prediction and the enhancement of *in silico* missense classification prediction. The chapter discusses potential future directions for extending this work into clinical applications.

Chapter 2: *In silico* missense variant classification in ordered regions and the importance of protein structure

Rohan Gnanaolivu[1] gnanaolivu.rohandavid@mayo.edu

Steven N Hart[1,2*] hart.steven@mayo.edu

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America

[2]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America

* Corresponding Author

E-mail: hart.steven@mayo.edu

Contributions: R.G. and S.H. conceived the study and designed the analysis. R.G. developed the code and analyzed the data. R.G. and S.H. wrote and edited the manuscript.

The following chapter has been adapted from the above publication: Rohan Gnanaolivu, Hart SN. Using AI-predicted protein structures as a reference to predict loss-of-function

activity in tumor suppressor breast cancer genes. *Computational and Structural Biotechnology Journal*. 2024;23:3472-3480. doi:https://doi.org/10.1016/j.csbj.2024.10.008

The supplementary methods and tables referenced in this chapter are included in the Appendix.

2.1 Abstract

Background: The loss-of-function (LOF) classification of most missense variants in tumor suppressor breast cancer genes *BRCA1, BRCA2, PALB2*, and *RAD51C* remains unclassified and confounds clinical actionability. Classifying these variants is challenging due to their rarity, leading clinicians to rely on *in silico* predictive methods. Protein stability changes are associated with function, making stability predictors valuable. Stability predictions upon missense variant perturbations require high-resolution protein structures. However, the availability of these high-resolution structures is lacking. This study explores using generative AI to predict high-resolution protein structures, which can then be analyzed with *in silico* protein stability prediction methods to assess LOF activity in ordered regions of the protein. This study also determines the appropriate *in silico* protein stability and dedicated *in silico* missense prediction methods in dbNSFP v4.7 database to predict LOF activity in ordered regions of these four genes. Functional classifications from homology recombination DNA repair (HDR) assays and variant classifications from the ClinVar database provide a reliable dataset for evaluating the performance of these *in silico* prediction methods.

Results: Complex AlphaFold2 structures of the BRCA1-C terminal (BRCT) domain and the DNA-binding domain of *BRCA2,* analyzed with FoldX predicts LOF activity from missense variants significantly better than experimentally-derived structures in ordered regions. The BRCT domain achieved an Area Under the Curve (AUC)=0.861 (95% CI:0.858-0.863) and AUC=0.842 (95% CI:0.840-0.845), while the DB domain achieved an AUC=0.836 (95% CI:0.8322-0.841), compared to AUC=0.847 (95% CI:0.844-0.850)

and AUC=0.835 (95% CI:0.832-0.837) from the BRCT domain, and AUC=0.830 (95% CI:0.821-0.8320) from the DB domain from experimentally-derived structures. Protein stability does not predict LOF activity from missense variants better than dedicated *in silico* missense predictors. Overall, we find that AlphaMissense ranks highly, with an average AUC=0.890 (95% CI 0.886-0.895) from ordered regions across these four cancer genes, compared to all other *in silico* missense predictors present in the dbNSFP database.

Conclusions: The study reveals that generative AI protein predicted structures can outperform experimentally-derived structures in evaluating LOF activity from predicted protein stability in ordered regions of genes BRCA1, BRCA2, PALB2 and RAD51C. The study also highlights the predictive performance of AlphaMissense as the premier *in silico* missense prediction method to predict LOF activity from missense variants in these four-tumor suppressor breast cancer genes. The code for this study can be downloaded for free on GitHub (https://github.com/rohandavidg/CarePred)

2.2 Background

Tumor suppressor genes *BRCA1, BRCA2, PALB2*, and *RAD51C* play crucial roles in HDR activity, and mutations in these genes have been implicated in breast cancer [27]. While the functional impact of truncating mutations in these genes is well characterized, the clinical impact of >95% of all possible missense mutation in genes *BRCA2, PALB2* and *RAD51C* and > 80% of all possible missense mutations in *BRCA1* remain unclassified, therefore, most are commonly referred to as variants of uncertain significance[3]. Various *in silico* and *in vitro* functional assays have been employed to evaluate function. Commonly used

in silico methods primarily use features such as sequence conversation, protein conformation and stereochemical properties of the amino acid as features to infer functional outcomes[10, 28, 29]. Researchers and Clinician have also utilized *in silico* stability predictors, which predict the destabilization or over-stabilization of the protein to infer loss or gain of function. Consequently, the absolute value of predicted ΔΔG is used to evaluate protein function [18, 19, 30].

Calculating protein stability involves measuring the effect and change in free energy of the protein based on the presence of a mutation. The change in protein structure directly impacts overall stability, particularly in haploinsufficient genes like *BRCA1, BRCA2, PALB2*, and *RAD51C* [31-34]. One method to assess protein stability is by estimating the difference in free energy of unfolding of the proteins (ΔG) between the wild-type and variant protein: ΔΔG = ΔG_variant – ΔG_wildtype. FoldX [35], Rosetta[36], and DDGun3D[37] are well-known stability predictors that induce a mutation from a given wild-type protein structure and predict ΔΔG. Recent publications have demonstrated that FoldX, Rosetta and DDGun3D had the highest correlation with functional measurements from deep mutational scanning data utilizing experimentally-derived protein complexes from the PDB compared to other stability predictors [38]. However, the impact of stability on function is protein dependent, as there are several factors such as sequence composition, post-translation modification, haploinsufficiency, and binding partners that influence stability of the protein[39, 40]. Therefore, the predictive performance of these stability methods to predict LOF in these tumor suppressor breast cancer genes is still relatively unknown. The study of protein function using predicted protein stability involves several

limitations as well, such as the inherent variability of these methods[41] and the availability of experimentally-derived crystalized structures of the complete protein.

Advances in generative AI, particularly in protein prediction models like AF2[42] and ESMFold

[43], have shown promise in predicting structures that are highly similar to structures found in the PDB. AF2 and ESMFold were shown to predict protein structures that are highly similar to experimentally-derived structures in the CASP15 challenge[44]. Recent CASP13[45] and CASP14[46] challenges highlighted the giant leap in advancement in protein structure prediction, with results highlighting the accuracy of predicted protein structures. Recent studies have demonstrated that features extracted from AF2 structures are effective in predicting the functional classification of missense variants [47]. However, the ability of using generative AI in HDR pathway genes to predict protein stability is relatively unknown and the impact of features derived from these generative AI models on protein stability is not well understood, with recent publications underscoring the reliability of protein stability predictions from structures generated by homology-based tools, particularly when sequence identity is at least 40%, highlighting the need for accurate wild-type structures in the prediction of protein stability from a missense variant [48, 49]

Experimental methods, such as MAVE and functional assays provide insight on LOF activity of *BRCA1, BRCA2, RAD51C*, and *PALB2*[50-55]. MAVE assays in these genes measure the HDR activity based on survival or growth of the cells upon the induced mutations. Results from these assays provide a reliable set of mutations to evaluate protein

stability on the functional activity in our genes of interest [56]. However, these MAVE assays are shown to have high stochasticity, and hence multiple replicates are often required. MAVE assays are also limited based on the biological mechanism involved and do not provide mutational classification on all possible missense mutations in these genes, with some assays being domain specific to the gene of interest.

In silico missense predictors are considered the weakest form of evidence to classify a missense variant [10]. REVEL [57] and BayesDel [58] are two *in silico* missense prediction tools that have been historically cited to accurately predict deleteriousness[59]. Newer, deep learning models such as AlphaMissense [60] and MetaRNN [61] have gained attention in predicting LOF activity. The effectiveness of these newer predictors, as well as those predicting protein stability based on functional classifications from MAVE experiments in these genes, is still largely unexplored.

The goal of this study was to assess how protein stability impacts LOF activity from missense variants in BRCA1, BRCA2, PALB2 and RAD51C. We used predicted protein structures from generative AI tools as the baseline wild-type structural templates, which were analyzed with protein stability methods FoldX, Rosetta, and DDGun3D to evaluate functional outcomes. Based on the AUC, our results demonstrate that FoldX, Rosetta, and DDGun3D can predict LOF from missense variants using AF2 wild-type structures, with performance comparable to the prediction derived from crystallized structures found in the PDB. Furthermore, AF2 wild-type structures enhance LOF predictions analyzed with FoldX in the BRCT domain of *BRCA1* and the DNA-Binding domain in BRCA2, compared

with the experimentally-derived structures. FoldX was also shown to significantly outperform Rosetta and DDGun3D to predict HDR activity from the MAVE assays in the BRCT domain of *BRCA1* and the DB domain in *BRCA2*. Finally, *in silico* predictor AlphaMissense ranked as the top predictor to predict LOF activity based on the AUC in the ordered regions of *BRCA1* and *BRCA2*, as well as it was found to be one of the top predictors in *PALB2* and *RAD51C*.

2.3 Methods

Our methodology employed a tiered comparative framework to evaluate the utility of generative AI protein models for missense variant analysis in tumor suppressor genes. First, we generated protein structures for established domains in four breast cancer susceptibility genes (BRCA1, BRCA2, PALB2, and RAD51C) using both AlphaFold2 and ESMFold, comparing their structural similarity to experimentally-derived structures. A stringent threshold of <3.8Å RMSD for the carbon-alpha backbone was implemented to identify highly similar structures. For structures meeting this similarity criterion, we proceeded to the second tier, which was the protein stability comparison. Here, missense variants with available functional data from MAVE assays were computationally introduced into both the generative AI-predicted and experimentally-derived structures. The predicted change in ΔΔG were calculated using three established computational predictors (FoldX, DDGun3D, and Rosetta). Finally, we compared these structural predictions with *in silico* missense classification tools from the dbNSFP database to perform comprehensive statistical analysis comparing ClinVar classifications with functional assay results. This multi-layered approach enabled systematic evaluation of how generative AI protein models

perform relative to experimental structures in predicting the functional consequences of cancer-associated missense variants (Figure 3).

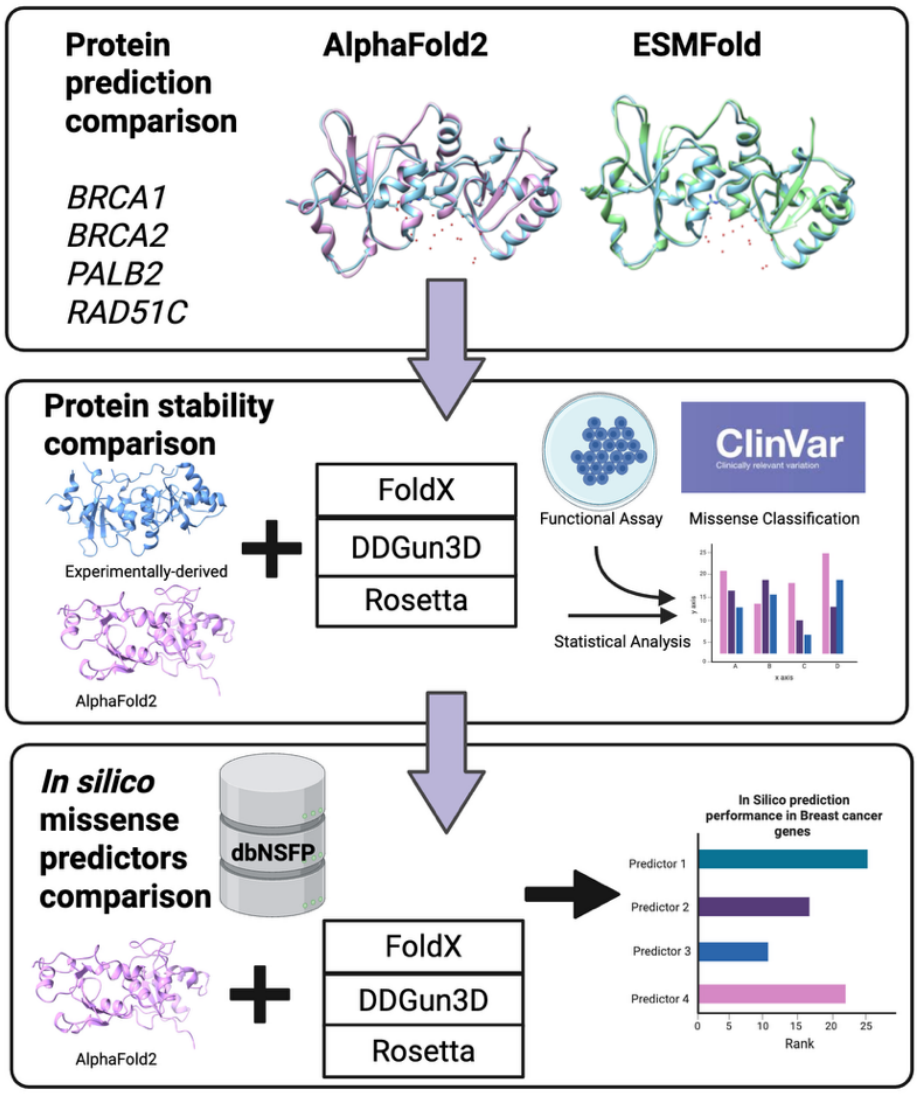

Figure 3: Computational workflow for protein structure prediction and missense variant analysis. The analysis pipeline consists of three main components*: (A) Protein prediction comparison between AlphaFold2 and ESMFold structures for cancer susceptibility genes (BRCA1, BRCA2, PALB2, RAD51C), showing structural representations with confidence coloring. (B) Protein stability comparison integrating experimentally-derived structures with AlphaFold2 predictions, utilizing computational tools (FoldX, DDGun3D, Rosetta) for stability analysis, functional assays, and ClinVar missense variant classification data, followed by statistical analysis and visualization. (C) In silico missense predictor comparison leveraging the dbNSFP database combined with structural analysis tools (FoldX, DDGun3D, Rosetta) and*

AlphaFold2 predictions to evaluate predictor performance across breast cancer genes, with results ranked by predictive accuracy. The workflow demonstrates the integration of structural predictions, stability calculations, and variant effect prediction for comprehensive missense variant interpretation in cancer predisposition genes. Created with BioRender.com

2.3.1 Data selection

We used the functional classification derived from the MAVE assays in genes *BRCA1, BRCA2,*

PALB2 and *RAD51C* respectively and the classification from the ClinVar database (Table S4). The MAVE assay for *BRCA1* was based on saturation Genome editing in HAP1 cell lines that had a total of 2086 mutations over the RING and BRCT functional domains that aim to measure the LOF activity in this cell line [51]. For *BRCA2*, we used the functional classification from a HDR cell-based assay from 462 missense variants affecting the *BRCA2* DNA binding domain [52]. For *RAD51C*, we used the functional classification from a HDR reporter assay, which introduced 174 missense variants in mammalian hRAD51C expression constructs using site-directed mutagenesis [53]. For *PALB2*, we used the functional classifications from 91 missense variants evaluated from a HDR assay from two separate studies [53].

We further supplemented our dataset with the mutational classification from the ClinVar database listed in dbNSFP 4.7 release [62]. We group all classification of pathogenic and likely pathogenic into the “deleterious” category, and classification of benign and Likely benign into the

“neutral” category.

2.3.2 PDB selection

We chose eleven structures from the PDB (Table S5), that represent different ordered regions of genes *BRCA1, BRCA2, RAD51C* and *PALB2*. We chose these structures based on their high resolution and coverage. These structures exist as a subunit and as a complex, with two or more protein chains representing a complex or as a single chain representing a subunit. For *BRCA1*, we downloaded 4OFB[63], 1JNX[64], 1T15[65] and 7LYB[66], as these structures represent the two ordered domains of *BRCA1* (BRCT and RING). 1JNX is a single chain subunit structure, 1T15 and 4OFB are complex structures that contains 2 protein chains each, representing the BRCT domain of *BRCA1*. 7LYB is a 7-protein chain complex structure, where chain 'M' represents the RING domain of *BRCA1*. For *BRCA2*, we downloaded 1MJE[67], 1IYJ[67] and 1MIU[67] which are two-protein chain complex structures and represents the ordered regions of the DB domain of *BRCA2*. Chain A in 1MJE and 1MIU represents the DB domain of BRCA2, whereas chain B in 1IYJ represents the DB of BRCA2. For *RAD51C*, we downloaded structures 8FAZ[68] and 8OUZ[69] , which are complexes with 4 protein chains, In 8FAZ, chain "C" represents the entire gene of RAD51C, whereas chain B represents entire gene of RAD51C. For *PALB2*, we chose the 3EU7[70] structure, a complex with two chains, with chain "A" representing the entire *PALB2* gene. Additionally, we downloaded the 2W18[70] structure, a subunit representing the *PALB2* gene.

2.3.3 AlphaFold2 and ESMFold prediction

AF2 structures of the 8 high resolution PDB structures were generated using ColabFold v1.5.5 [71] using the pdb100 templates. Default settings were used for the remaining configurations. Using BioPython [72], the FASTA sequence from each experimentally derived structure was extracted and used as inputs to generate five separate replicates, ranked based on model confidence. For the eight experimentally-derived structures 8FAZ, 8FOUY, 7LYB, 4OFB, 1JM7, 1IYJ, 1MJE, 3EU7, and 2W18, we generated predicted protein structures utilizing the AF2 multimer prediction tool with ColabFold (Table S6).

Using the Python implementation of ESMFold, we generated predicted structures for six out of the eight experimentally-derived structures using their FASTA sequences as input. We employed the ESM-2 pretrained model "esm2_t33_650M_UR50D," which comprises 650 million parameters and 33 layers, running on a single T4 GPU. ESMFold has a limitation of generating structures with a maximum of 1024 amino acids. Consequently, for the complex structures 7LYB and 1IYJ, which exceed 1024 amino acids, complete structures were not created.

2.3.4 Protein stability

FoldX, Rosetta and DDGun3D predictors were used to predict ΔΔG with both AF2 and experimentally-derived structures for all possible missense mutations, leading to a total of 246,910 mutations. Using the python module pyFoldX [73], which uses FoldX 5.0 the structures were passed through the "RepairPDB " function prior to ΔΔG calculations, with

default settings. For every mutation, five replicates were computed and the mean ΔΔG value was calculated. DDGun3D python package was used to generate predicted ΔΔG using the default settings, five replicates were computes and the mean ΔΔG value was calculated. The Cartesian ΔΔG application from Rosetta suite (Linux build 2021.16.61629) was used to generate the ΔΔG predictions, following standard protocols published in several publications[74] while using the Ref2015 scoring function. The structures were initially relaxed and ΔΔG predictions were made over three iterations, which was averaged to generate mean ΔΔG in (Rosetta energy per unit). To convert the units into kcal/mole scale, we used a scaling factor of 2.94, as shown in the literature [38, 74].

2.3.5 Statistical analysis

A Mann-Whitney U statistical test was employed from the SciPy [27] stats python package to understand the association of predicted |ΔΔG| generated from FoldX, Rosetta and DDGun3D to protein LOF activity. The metric Root Mean Square Deviation (RMSD) calculations were computed using the *superimposer* method from the BioPython PDB python package. RMSD was used as the metric to evaluate the similarity between the AF2 structure compared to the experimentally-derived structure downloaded from the PDB. Similarity was calculated over the entire structure and over the carbon-alpha (Cα) backbone chain. We use SciPy stats package to compute the spearman correlation coefficient between predicted |ΔΔG| from AF2 structures and experimentally-derived structures analyzed with FoldX, Rosetta and DDGun3d to evaluate the linearity. We also compute the spearman correlation of ΔΔG from AF2 structures and experimentally-derived structures analyzed with FoldX, Rosetta and DDGun3d, compared to the continuous functional

MAVE activity score. To compute the confidence interval of the spearman correlation, we do a Fischer transformation of these correlation coefficients. Utilizing these Fisher transformations, we calculate the combined standard errors, the z-scores of the difference and then using the z-scores, we calculate the p-value based on the normal distribution. No corrections were applied for multiple testing.

To calculate the predictive ability of the *in silico* missense predictors and stability predictors to predict LOF activity from missense variants from ordered regions in BRCA1, BRCA2, PALB2 and RAD51C, we used the metric AUC from the *roc_auc_score* method from the scikit-learn v1.0.2 [75] package. We computed the AUC, the AUC under the precision-recall curve and the false positive rates, along with their 95% confidence intervals. This was achieved by sampling 200 times across the dataset while balancing out the class labels. The python code used for this study is available at https://github.com/rohandavidg/CarePred.

2.4 Results

2.4.1 Comparison of Generative AI protein prediction structures

From the literature, two protein structures are considered similar if the RMSD between Cα back-bone chains that are superimposed is <3.8Å [76]. Using this threshold as the metric for similarity, AF2 generates 3D protein structures comparatively similar to the experimentally-derived structures compared to ESMFold from the 9 structures analyzed representing the domains in BRCA1, BRCA2, PALB2 and RAD51C. A RMSD of <3.8Å threshold was generated for only 3 structures (4OFB, 1JNX and 2W18) by ESMFold, and

due to the size limitation of 1024 residues, 2 complex structures (7LYB, 1IYJ) were not analyzed. AF2 predicts 7 structures less than the RMSD threshold and only one 1IYJ fails the threshold (Table 3). 1IYJ could not be predicted without significant error and hence it was excluded from further analysis. Calculating the per residue distance (Å) when the predicted structure was superimposed on the experimentally-derived structure, >98% of the residues was <3.8Å in 6 out of 8 structures, highlighting the prediction accuracy of AF2. However, 7LYB had a mean residue distance of 19Å, even though the distances between the *BRCA1* Cα back-bone chain and its experimentally-derived counterpart was <3.8Å (Figure S9)

Table 3: Root Mean Square Deviation values of AlphaFold2 and ESMFold structures superimposed on experimentally-derived structures from the PDB.

GENE	**PDB ID**	**Residue Count**	**AF2 subunit backbone (RMSD) Å**	**ESMFold Subunit backbone (RMSD) Å**
RAD51C	8FAZ	926	1.04	7.06
BRCA1	7LYB	1389	1.16	NA
BRCA1	4OFB	223	0.4	1.53
BRCA1	1JNX	207	0.8	1
BRCA2	1IYJ	1272	9.28	NA
BRCA2	1MJE	648	2	7.46
PALB2	3EU7	327	0.66	9.79

PALB2	2W18	306	1.08	1.36

2.4.2 Association of protein function from predicted |ΔΔG| generated from experimentally derived structures

Utilizing the Mann-Whitney U statistical test, the association of |ΔΔG| derived from 7 experimentally derived structures PDB ID:1JNX, 4OFB, 1JME, 3EU7, 2W18, and 8FAZ representing regions of *BRCA1, BRCA2, PALB2* and *RAD51C*, analyzed with protein stability predictors FoldX, Rosetta and DDGun3D, was tested against the LOF classification from the mutations found in the ClinVar database and the MAVE assays. Our results show that there was a significant association of predicted |ΔΔG| to LOF activity in all 4 genes (Figure S10).

2.4.3 Linearity of |ΔΔG| generated from protein stability predictions using AF2 structures compared to experimentally-derived structures.

Utilizing the 7 AF2 structures that had an RMSD of the Cα backbone < 3.8Å, the spearman correlation coefficient (rho) was calculated from the |ΔΔG| analyzed with the AF2 wild-type structures and experimentally-derived structures. There is a strong correlation between the |ΔΔG| predictions from FoldX, Rosetta, and DDGun3D using both AF2 structures and experimentally-derived structures (Figure 4). FoldX and Rosetta had a rho value > 0.75 in 5 out of the 7 structures, indicating strong correlations, with 7LYB and IMJE being the only exception. FoldX generated a rho value of 0.63 (95% Confidence Interval (CI) 0.60-0.66) with 7LYB and 0.70 (95% CI 0.690.70) with 1MJE. Rosetta had a rho value of 0.71 (95% CI 0.68-0.73) with 7LYB and 0.75 with 1MJE (95% CI: 0.74-0.75). DDGun3D showed a strong correlation for all 7 structures, with a mean rho value >0.97.

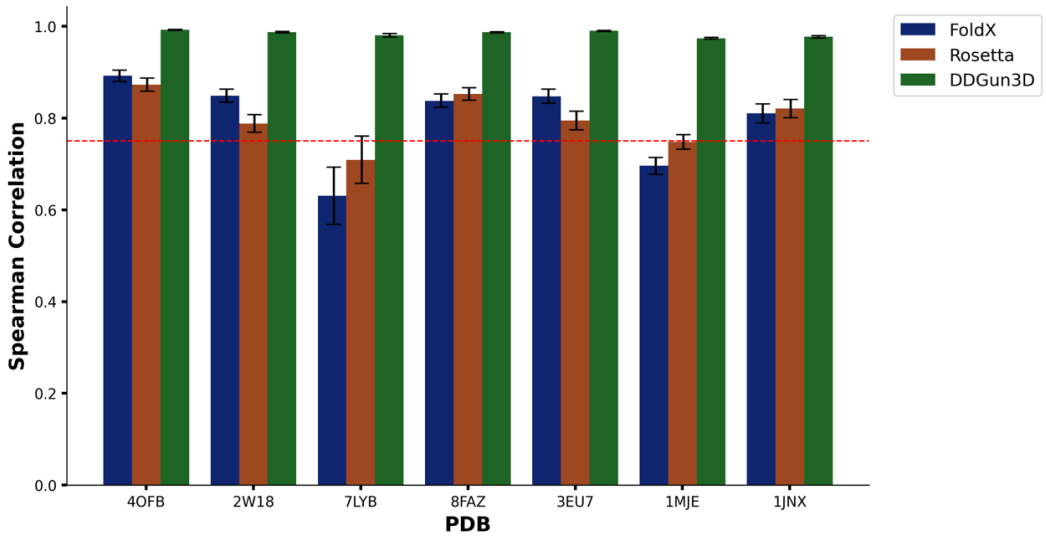

***Figure 4**: Spearman correlation denoting the linearity between the predicted |ΔΔG| derived from experimentally-derived structures vs AlphaFold2 structures as the wild-type template analyzed with protein stability predictors FoldX, Rosetta and DDGun3D. The red line at y=0.75 indicates the threshold for strong correlation.*

To assess whether the heterogeneity of the rho value was impacted by the features from AF2 structures, we tested the monotonic relationship of the |ΔΔG| differences between the predicted |ΔΔG| derived from AF2 structures and experimentally-derived structures, analyzed with FoldX, Rosetta and DDGun3D with the features from AF2 structures. A rho value was calculated, and no monotonic relationship was observed between the |ΔΔG| differences and the AF2 features (Figure S11A). We further examined whether the heterogeneity of the |ΔΔG| differences was dependent on the per-residue distance (Å) between the superimposed AF2 structure and the experimentally-derived structure. The results showed that the rho was not significantly impacted by per-residue distances for up to 4Å (Figure S11B). However, there was a positional effect on heterogeneity, with certain amino acids at specific positions contributing disproportionately to the variability in the prediction from FoldX and Rosetta**.**

2.4.4 Comparison of predicted |ΔΔG| from experimentally-derived structures vs AF2 structures to predict LOF

The predicted |ΔΔG| from FoldX analyzed with the AF2 structures were significantly better than the experimentally-derived structures in the BRCT domain of *BRCA1* and the DBD binding domain in BRCA2 to predict the categorical LOF activity using the Delong test to compare the AUC (Figure 5). In the BRCT domain of *BRCA1*, FoldX generated an AUC=0.861 (95% CI:0.858-0.863) from the predicted AF2 complex structure of 4OFB, compared to AUC=0.847 (95% CI: 0.844-0.850) from the experimentally-derived complex structure of 4OFB. FoldX also generated an AUC=0.836 (95% CI:0.833-0.839) from the AF2 subunit structure of 1JNX, compared to an AUC=0.792 (95% CI:0.789-0.795) from the experimentally-derived subunit structure of 1JNX, showing that utilizing the complex AF2 structure enhances the prediction of LOF activity using FoldX in the BRCT domain of *BRCA1* (Table S7). However, in the RING domain of *BRCA1*, FoldX using complex experimentally-derived structure of 7LYB was significantly better than the predictions from FoldX with an AF2 wild-type template of 7LYB, with an AUC=0.835 (95% CI=0.831-0.840), compared to AUC=0.741 (95% CI:0.735-0.748) (Table S8). The AF2 complex structure of 7LYB was noted to have a mean per residue distance of 19Å when superimposed on the experimentally-derived structure, which was outlier compared to the other 6 structures that had a mean per residue distance of 3Å (Figure S9).

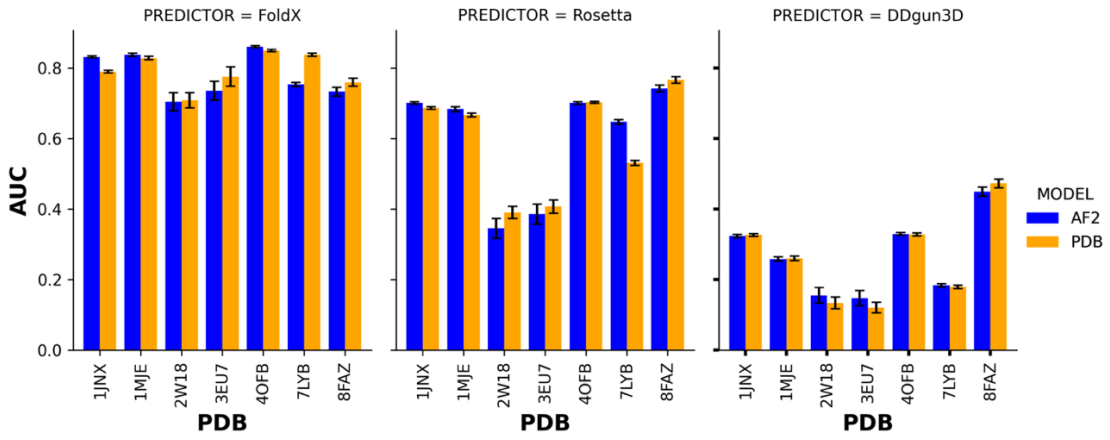

Figure 5*: Area under curve denoting the predictive ability of |ΔΔG| from experimentally-derived structures vs AlphaFold2 structures as the wild-type template to predict loss-of-function activity in BRCA1, BRCA2, PALB2 and RAD51C analyzed with protein stability predictors FoldX, Rosetta and DDGun3D.*

In *BRCA2*, Rosetta analyzed with the AF2 complex structure of 1MJE representing the BRCA2DSS1-ssDNA complex, generated an AUC=0.681 (95% CI:0.675-0.687), which was significantly better that an AUC=0.665 (95% CI:0.659-0.672) which was generated when analyzed by the experimentally-derived complex structure of 1MJE. FoldX predictions of LOF activity were significantly better than Rosetta and DDGun3D, and sub-setting to the mutations to the DBD domain of BRA2, FoldX analyzed with AF2 complex structure of 1MJE generated an AUC=0.8364 (95% CI:0.8322-0.8407), which was significantly better than an AUC=0.8299 (95% CI:0.8213-0.8320) that was generated by FoldX analyzed with the experimental-derived complex structure of 1MJE (Table S9).

In *PALB2*, there was no significant difference between the predicted |ΔΔG| from the AF2 complex structure and the experimentally-derived complex structure analyzed with FoldX. Among the stability predictors, FoldX predictions of LOF activity were significantly better

than Rosetta and DDGun3D. Analyzing the AF2 complex structure of 3EU7, FoldX generated an AUC=0.721 (95% CI:0.694-0.747), compared to Rosetta and DDGun3D, which generated an AUC=0.401 (95% CI:0.375-0.427) and AUC=0.157 (95% CI:0.135-0.179) respectively. With the AF2 subunit structure of 2W18, FoldX generated an AUC=0.719 (95% CI: 0.695-0.743), which was significantly better than Rosetta and DDgun3D, with an AUC=0.359 (95% 0.329-0.388) and AUC=0.159 (95% CI:0.138-0.181) respectively (Table S10).

In *RAD51C*, we find that predicted |ΔΔG| analyzed from the experimentally-derived complex structure of 8FAZ was significantly better than the complex AF2 structure of 8FAZ analyzed with Rosetta and FoldX to predict LOF activity. An AUC=0.766 (95% CI:0.757-0.775) was generated from experimentally-derived complex, which was significantly better than AUC=0.741 (95% CI:0.731-0.751) generated from the AF2 structure analyzed with Rosetta. With FoldX as well, an AUC=0.759 (95% CI: 0.748-0.771) was generated from experimentally-derived complex, which was significantly better than AUC=0.732 (95% CI: 0.720-0.745) from the AF2 complex (Table S11).

Using rho as the metric, we evaluated the prediction of |ΔΔG| using the experimentally-derived structures and AF2 structures with the continuous measurement of HDR functional activity from MAVE assays and found no significant difference in the correlation of functional activity with |ΔΔG| from the 7 AF2 structures or its experimentally derived counterparts analyzed with FoldX, Rosetta or DDGun3D (Figure 6). However, there was heterogeneity from the prediction among the stability predictors. The rho values derived

using the predicted |ΔΔG| analyzed from the AF2 complex structure of 4OFB representing *BRCA1* BRCT domain and AF2 structure of 1MJE representing BRCA2-DSS1-ssDNA complex using FoldX was significantly better than the rho from the AF2 structures or the experimentally-derived structure analyzed with Rosetta and DDGun3D. A rho=0.497 (95% CI:0.455-0.537) was observed from the AF2 structure of 4OFB, compared to rho=0.379 (95% CI:0.331-0.424) from experimentally-derived structure and rho=0.360 (95% CI:0.312-0.407) from the AF2 structure analyzed by Rosetta, and rho=0.33 (95% CI:0.25-0.382) from experimentally-derived structure and rho=0.334 (95% CI:0.285-0.381) from the AF2 structure analyzed by DDGun3D. A rho=0.634 (95% CI:0.562-0.696) was observed from the AF2 structure of 1MJE, compared to rho=0.455 (95% CI:0.362-0.539) from experimentally-derived structure and rho=0.470 (95% CI:0.378-0.553) from the AF2 structure analyzed by Rosetta, and rho=0.462 (95% CI:0.369-0.545) from experimentally-derived structure and rho=0.470 (95% CI:0.357-0.535) from the AF2 structure analyzed by DDGun3D. In *PALB2*, and *RAD51C*, there was no significant difference between FoldX, Rosetta and DDGun3D to predict the continuous HDR functional activity.

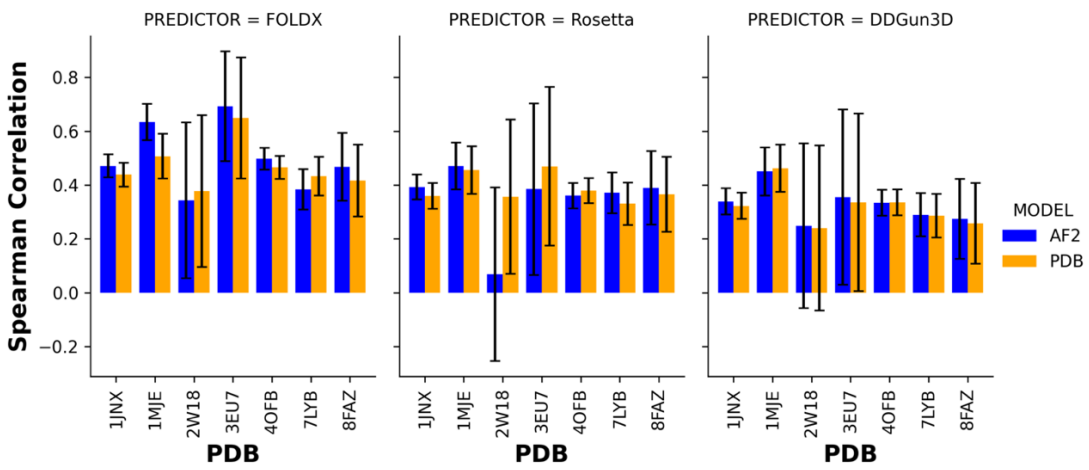

***Figure 6**: Spearman correlation denoting the linearity of predicted |ΔΔG| from experimentally-derived structures vs AlphaFold2 structures with the continuous measurement of functional HDR activity in genes BRCA1, BRCA2, PALB2 and RAD51C utilizing the |ΔΔG| predictions from protein stability predictors FoldX, Rosetta and DDGun3D*

2.4.5 Comparison of protein stability predictors with purposefully developed *in silico* missense predictors of function

We find that existing *in silico* missense predictors significantly outperform protein stability predictors in predicting LOF activity caused by missense variants, based on the AUC (Figure 7). Additionally, no single *in silico* missense predictor emerges as a clear choice for predicting LOF activity in all four genes based on the AUC. When considering the average AUC from all four genes, we find that MetaRNN and AlphaMissense are the leading predictors, with MetaRNN achieving an average AUC=0.895 (95% CI:0.891-0.902) and AlphaMissense reaching an average AUC=0.890 (95% CI 0.886-0.896). To determine the most appropriate *in silico* missense or stability predictor for predicting LOF activity in ordered regions across all four genes, we rank ordered the predictions for all predictors in the dbNSFP v4.7 database and stability predictors, calculating the average rank across all four genes based on the AUC.

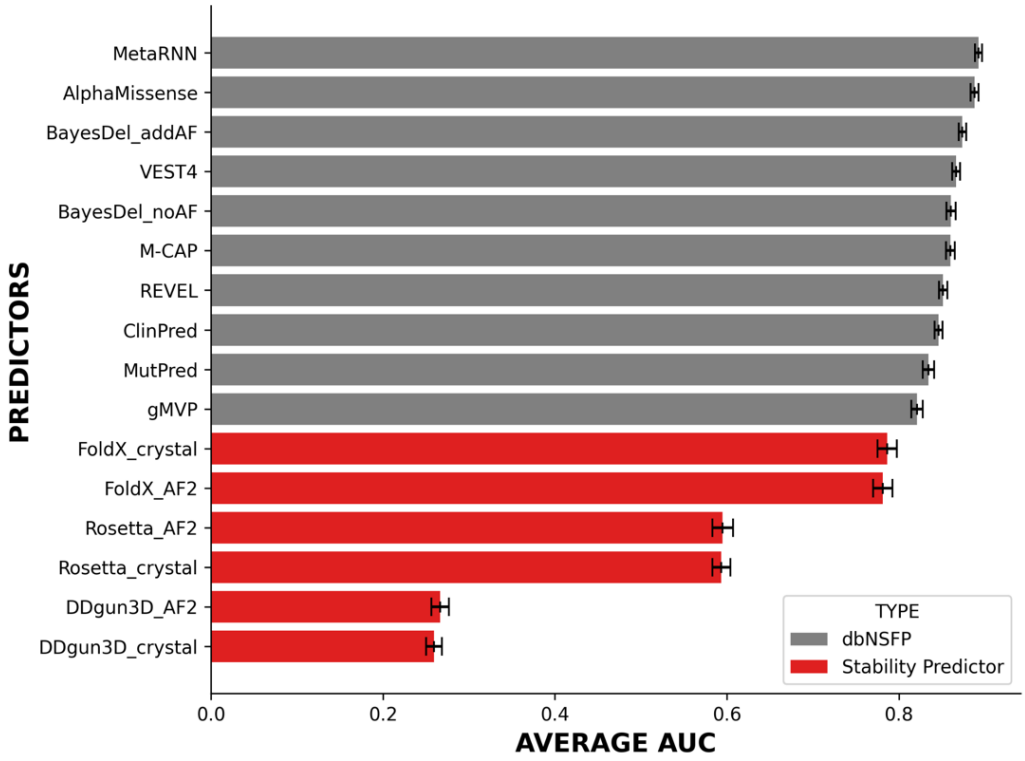

***Figure 7**: Area under the curve of in silico missense predictors vs stability predictors to predict loss-of-function activity in BRCA1, BRCA2, RAD51C and PALB2 stratified by in silico missense predictors found in the dbNSFP database (in grey) vs protein stability predictors (in red).*

From the rank ordered results (Figure 8), we find that AlphaMissense had the highest average rank, followed by MetaRNN, and BayesDel_addAF. This result would imply that AlphaMissense is the best *in silico* missense predictor in predicting LOF activity in these breast cancer genes. The best stability predictor is FoldX using AF2 wild-type structures which had a rank of 19. This suggests that predicted |ΔΔG| predicts LOF better than 38 other predictors in the dbNSFP database. However, we should note that the mutations used for this evaluation are enriched with mutations that are in ordered regions of the protein.

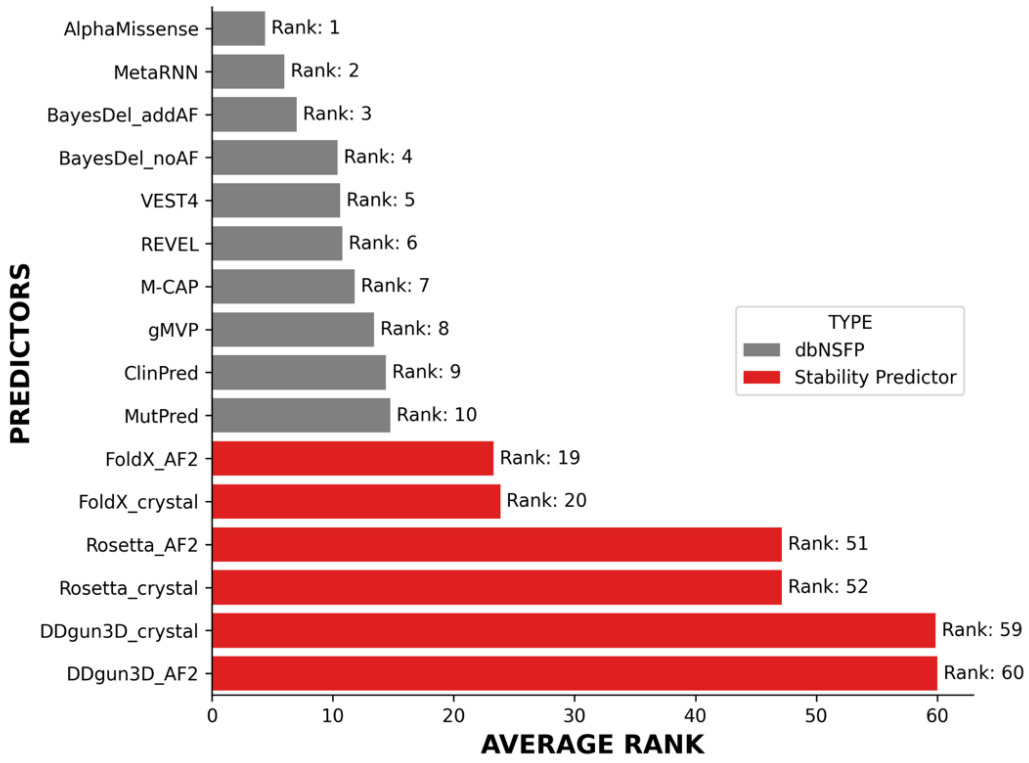

***Figure 8**: Rank ordered performance of in silico missense predictors vs stability predictors to predict loss-of-function activity in BRCA1, BRCA2, RAD51C and PALB2 stratified by in silico missense predictors found in the dbNSFP database (in grey) vs protein stability predictors (in red).*

We also evaluated the false positive rates (FPR) for all dbNSFP predictors and stability predictors. Our analysis showed that AlphaMissense, MetaRNN, and gMVP had the lowest FPRs, with values comparable to one another (Figure S12). Specifically, AlphaMissense had an average FPR of 0.139 (95% CI: 0.130-0.148), MetaRNN had an average FPR of 0.143 (95% CI: 0.134-0.152), and gMVP had an average FPR of 0.146 (95% CI: 0.137-0.156) across the ordered regions for all four genes.

2.5 Discussion

2.5.1 AF2 predictions are better than ESMFold

AF2 and ESMFold are considered to two best protein structural predictors and consistent with existing literature, AF2 multimer predicts protein structures that are highly similar to experimentally-derived structures compared to ESMFold in *BRCA1, BRCA2, PALB*2, and *RAD51C*. ESMFold predictions are comparatively faster than AF2, but the predictions are limited

to 1024 amino acids, thereby larger structures, such as 7LYB were not compared. AF2 is more suited for this study for the high similarity and its ability to predict large protein structures.

2.5.2 Association of predicted protein stability to protein loss-of-function

Consistent with existing literature, we found that there is a strong association of stability towards LOF activity using all three protein stability predictors, suggesting that protein destabilization or over-stabilization is an important factor in the contribution to LOF activity in *BRCA1, BRCA2, PALB2* and *RAD51C* [19, 30]. These genes have been reported to be haploinsufficient, suggesting that minor perturbation is required to completely destabilize the protein to impact loss-of-function. We also chose FoldX, Rosetta and DDGun3D as these predictors were shown to perform the best in terms of correlation results with deep mutational scan data using complex experimentally-derived protein structures and these predictors require a wild-type protein template to generate its predictions [38].

2.5.3 Prediction from DDGun3D is less dependent on wild-type protein templates

Overall, strong correlation was observed between the |ΔΔG| prediction using either AF2 structure of the experimentally-derived structure, with the exception for 7LYB and 1MJE. We find that DDGun3D predictions showed stronger correlations compared to FoldX and Rosetta. This could be due to the fact that only 33% of DDGun3D predictions is based on the wild-type protein structure, while the remaining contribution comes from Blosum62 matrix substitution scores, difference in statistical potential score from the linear chain of the amino acid between the wild-type and mutant, and the difference between the hydrophobicity of the wild-type and mutant [37], whereas FoldX and Rosetta use a mixture of physics and statistical methods to compute protein stability based on the clashes caused by introducing a mutant within the wild-type structures.

2.5.4 Protein structural features have no impact on protein stability predictions.

Recent publications have highlighted the use of protein structural features from AF2 structures to study LOF activity [47]. Features such as Predicted Local Distance Difference Test (PLDDT), Relative Solvent Accessibility (RSA), distance between residues, physicochemical properties such as atomic weight, isoelectric points and aromaticity were shown to be useful in predicting LOF activity. We tested the dependence of the difference in |ΔΔG| prediction from experimentally-derived structure and AF2 structures with these features. Our results showed no influence on the difference in stability prediction to the

difference in features values, thus suggesting that these features are independent of protein stability and hence have independent effects on the relationship with LOF. Thereby we can make a claim that stability, along with features such as PLDDT, RSA, distance between residue, physicochemical properties such as atomic weight, isoelectric points and aromaticity can be used to study protein LOF in *BRCA1, BRCA2, PALB2* and *RAD51C*.

2.5.5 FoldX prediction using AF2 wild-type structures predicts loss-of-function better than other stability predictors.

Overall, our results support the choice of using FoldX as the protein stability predictor in these genes using complex structures of AF2 predicted structures, compared to Rosetta and DDGun3D. More importantly, our results suggest that using FoldX with AF2 structures that are highly similar to experimentally-derived structures predict LOF activity just as good as prediction from FoldX with complex experimentally-derived structures. Our results also suggest that there is a domain specific effect on the predictiveness of stability towards function in *BRCA1*, suggesting that the degree of perturbation is heterogeneous across domains. Even though the predictions are based on several replicates, we acknowledge that there is inherent noise in the predictions from FoldX, Rosetta and DDGun3d. These predictors were not inherently built to predict protein LOF, but experimental protein stability.

2.5.6 AlphaMissense prediction of LOF activity in breast cancer genes

The analysis performed, based on the AUC metric, suggest that *in silico* missense predictors predicts functional consequences better than protein stability predictors in *BRCA1, BRCA2, RAD51C* and *PALB2*. AlphaMissense consistently ranks highly in the prediction of LOF activity from missense variants in ordered regions in all four genes, and on average AlphaMissense emerges as one of the top-ranking predictors. BayesDel is currently considered the *in silico* model of preference by the ClinGen Variant Curation Expert Panel [52] (https://cspec.genome.network/cspec/ui/svi/doc/GN092), however newer predictors, such as AlphaMissense and MetaRNN predict function significantly better than BayesDel in the ordered regions of these four breast cancer genes based on the AUC metric. AlphaMissense utilizes a transformer based multiple sequence alignment of protein sequences, whereas MetaRNN is an ensemble neural network model that uses many existing *in silico* missense predictors as features and hence there is an argument that the predictions might be overfit to the training data to the individual features.

In summary, Generative AI tools such as AF2 multimer predictions can be used in the prediction of protein stability in haploinsufficient genes *BRCA1, BRCA2, PALB2* and *RAD51C* to study LOF activity from a missense variant. However, these predictions are primarily focused on mutations in ordered regions and do not account for disordered regions of the protein complex. The application of generative AI to study functional consequence in disordered regions was beyond the scope of this study. While we acknowledge the limitations of ClinVar data, particularly the circularity issue where

training data used to develop *in silico* missense predictors may be used in their evaluation, this concern is especially pronounced in ensemble based methods employed in the development of these *in silico* missense predictors [77]. Additionally, MAVE assays provide an independent dataset for evaluating LOF activity from missense variants, though they are not without limitations. These assays are known to have errors due to the type of assay used, experimental artifacts, and variability in replicates. While these assays will not replace computational methods in the near future, they contribute to refining our understanding of the factors causing damaging effects on protein function. [78]

Chapter 3: *In silico* missense variant classification in intrinsically disordered regions

Rohan Gnanaolivu[1] gnanaolivu.rohandavid@mayo.edu

Steven N Hart[1,2*] hart.steven@mayo.edu

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America

[2]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America

* Corresponding Author

E-mail: hart.steven@mayo.edu

Contributions: R.G. and S.H. conceived the study and designed the analysis. R.G. developed the code and analyzed the data. R.G. and S.H. wrote and edited the manuscript.

3.1 Abstract

Accurate classification of missense variants is a fundamental challenge in genomics, particularly for those within intrinsically disordered regions (IDRs) where the performance of existing computational predictors is suboptimal. To address this, we developed a machine learning model that extends traditional missense tools with properties that infer globular IDR conformation, phase separation, and protein embeddings. Using ClinVar variant classifications as ground truth, AlphaMissense, EVE, and ESM1b were the highest scoring unsupervised *in silico* missense predictors for IDR variants. Our baseline model, using only IDR-specific features achieved competitive performance on the hold-out test set with a PR-AUC of 0.800. Critically, when these IDR features were combined with these methods we saw significant overall improvement. The AlphaMissense-Enhanced model increased its PR-AUC from 0.807 to 0.931. Similarly, ESM1b-Enhanced improved PR-AUC from 0.679 to 0.878 and EVE increased from 0.591 to 0.918. These results demonstrate the effectiveness of our enhancements for classifying missense variants in IDRs and highlight its ability to complement existing *in silico* missense predictors.

3.2 Introduction

Missense variants are single nucleotide polymorphisms (SNPs) that result in the substitution of a single amino acid in the protein sequence. These changes can have a wide range of effects on protein function, from benign alterations to severe pathogenic consequences. For example, the Glu6Val (E6V) variant in *HBB* is responsible for sickle cell disease[79], while the Gly380Arg (G380R) variant in *FGFR3* leads to achondroplasia[80]. Understanding the functional impact or pathogenic potential of missense variants is essential for accurate disease diagnosis, prognosis, and therapeutic decision-making. Databases such as ClinVar[3] and Human Gene mutation database (HGMD)[12] contain curated missense variants that are classified for their role in causing disease. However, a large portion of missense variants remain unclassified due to the rarity in the population and accurately predicting the deleteriousness of missense variants remains a significant challenge[81].

Many computational *in silico* missense predictors have been developed to aid in classifying missense variants[6, 8, 29, 58, 82, 83]. These models employ machine and deep learning techniques, including supervised and unsupervised approaches, and rely on features set based on sequence conservation, protein secondary structure and physicochemical properties to assess variant pathogenicity[4]. The American College of Medical Genetics and Genomics (ACMG) guidelines also recognize the importance of computational missense predictors, categorizing them as supportive evidence for pathogenic (PP3) and benign (BP4) classifications[84], under the assumption that a prediction of altered function

is equivalent to pathogenicity[85]. Recent predictors such as AlphaMissense[60], ESM1B[86], and EVE[82] has demonstrated strong performance from missense variants in structured regions. However, the performance in IDRs is suboptimal compared to their performance in ordered regions [60, 87, 88]. For example, AlphaMissense reports an AUC of 0.94 for missense variants in ordered regions but only about 0.85 for missense variants in disordered regions.

Traditional *in silico* missense classification predictors predominantly rely on sequence conservation-based features, assessing variant tolerance in highly conserved genomic regions and their impact on secondary structure. Even though 15-20% of IDRs are in regions with low sequence conservation, they are generally less conserved than structured domains and exhibit greater tolerance to variants[89]. Unlike ordered protein domains, IDRs lack stable conformation, can adopt multiple structural states, and exist in a low-energy state, making them challenging to model using conventional computational approaches. Despite this, IDRs play essential roles in protein function, with approximately 25% of disease-associated variants localized within these regions[88].

IDRs are highly abundant in the eukaryotic proteome, accounting for nearly 30% of all proteins, and play critical roles in processes such as transcriptional regulation, DNA replication, and signal transduction[90]. IDRs are protein segments that lack a stable secondary or tertiary structure and exist as dynamic ensembles. Their structural flexibility

makes them challenging to study using X-ray crystallography or cryo-electron microscopy[91]. Instead, IDRs are typically identified based on their amino acid composition, and various disorder prediction algorithms have been developed to assess their propensity for disorder[92]. AlphaFold2 demonstrated a strong correlation between low-confidence predictions and IDRs. A study found that a combination of the confidence score pLDDT from AlphaFold2 and Relative Solvent Accessibility (AlphaFold-RSA) provides a robust approach for predicting disordered regions within proteins[93, 94].

IDRs do not adopt a single stable structure; instead, they exist in multiple conformations, influencing their flexibility and interaction potential. Recent advances in the prediction of global protein conformation (gIDRc) can predict biophysical properties of IDRs[90], providing insight into their structural adaptability altered behavior. Beyond structural flexibility, IDRs often drive biomolecular phase separation (PS), forming dynamic membrane-less organelles that regulate cellular organization. Variants within IDRs can disrupt PS, leading to loss or gain of function and contributing to diseases such as neurodegeneration and cancer[88]. Several novel computational methods exist that can predict PS using different approaches. PSAP[95] uses 55 features from amino acid composition, trained on 90 manually curated human proteins. PSPHunter[96] employs 123 features including sequence embeddings and evolutionary data from large multi-species databases. catGRANULE 2.0[97] integrates 128 features combining physicochemical properties with AlphaFold2 structural information. These methods specifically predict protein condensate formation and PS propensity, that differs from traditional *in silico*

missense predictors that focus on protein stability, folding, and general functional disruption. PS predictors instead assess the capacity for proteins to form dynamic, reversible liquid-like assemblies that are critical for cellular organization and regulation. In addition to biophysical modeling, protein language models (pLMs) offer a powerful approach to understanding missense variant effects in IDRs. Recent studies have demonstrated that pLMs can effectively capture sequence-based features relevant to protein function, making them valuable tools for missense variant prediction[98]

To address the challenge of missense variant classification in IDRs, we developed a machine learning model that integrates biophysical features unique to these regions. Specifically, our model incorporates predictions of gIDRc, PS, and contextual protein embeddings derived from pLMs. We first establish a baseline model using only these IDR-specific features. We then demonstrate that this framework can be used to significantly enhance the performance of existing state-of-the-art unsupervised predictors, such as AlphaMissense, EVE, and ESM1b, by integrating their scores as additional features. This work presents a novel framework that provides a more nuanced and accurate classification of missense variants in disordered regions, addressing a critical gap in variant interpretation.

3.3 Materials and methods

3.3.1 Model generation

To analyze the impact of missense variants in IDRs, we utilized protein coordinate predictions of disorder from AlphaFold-RSA, downloaded from MobiDB[99]. The dataset was filtered for human proteins (NCBI Taxon ID 9606), and reference protein FASTA sequences for all proteins were downloaded from UniProt using their corresponding UniProt IDs (Figure. 13A). These reference sequences represent the full-length protein sequences, incorporating regions predicted to be disordered by AlphaFold-RSA. We then identified all missense variants from the ClinVar database that overlapped with these disordered regions. For each of these variants, a corresponding mutant FASTA sequence was generated by introducing the mutation within the predicted disordered segment of the reference sequence (Figure. 13B). This approach ensured that only variants occurring within IDRs were considered for further analysis. The reference and mutant FASTA sequences were then used as inputs for feature and embedding extraction using tools ALBATROSS[90], PSAP[95], and ProtTrans[98]. ALBATROSS is used to predict biophysical properties that can be used to infer global protein conformation from an IDR. PSAP is a predictor that predicts PS from an IDR protein sequence. For ALBATROSS and PSAP, the absolute delta change between the reference and mutant sequences was computed to quantify the structural and PS alterations caused by the variant. With ProtTrans, protein embeddings were initially generated for each amino acid across the entire IDR protein sequence input, subsequently, the mean of the embeddings of size 1024 across all amino acids in the sequence was calculated to create a single scalar representation for the protein sequence. This process was performed separately for the reference and

mutant sequences. To capture mutation induced shifts in feature representation, the average of the reference and mutant embeddings was computed. This aggregated representation effectively summarizes the contextual changes introduced by the mutation and was used as the input feature set for downstream analysis. These embeddings of size 1024 was further refined to only the top 20 embeddings that were most predictive to deleterious function. The extracted features from ALBATROSS, PSAP and ProtTrans were then concatenated. A gradient boosting classifier (XGBoost) was trained, optimized using hyperparameter tuning with Optuna[100] and validated on a hold-out test set that resulted from randomized splitting train and test to assess its performance in distinguishing the impact of missense variants in IDRs (Figure. 13C). The code for this study is available at https://github.com/rohandavidg/IFP-MIDR

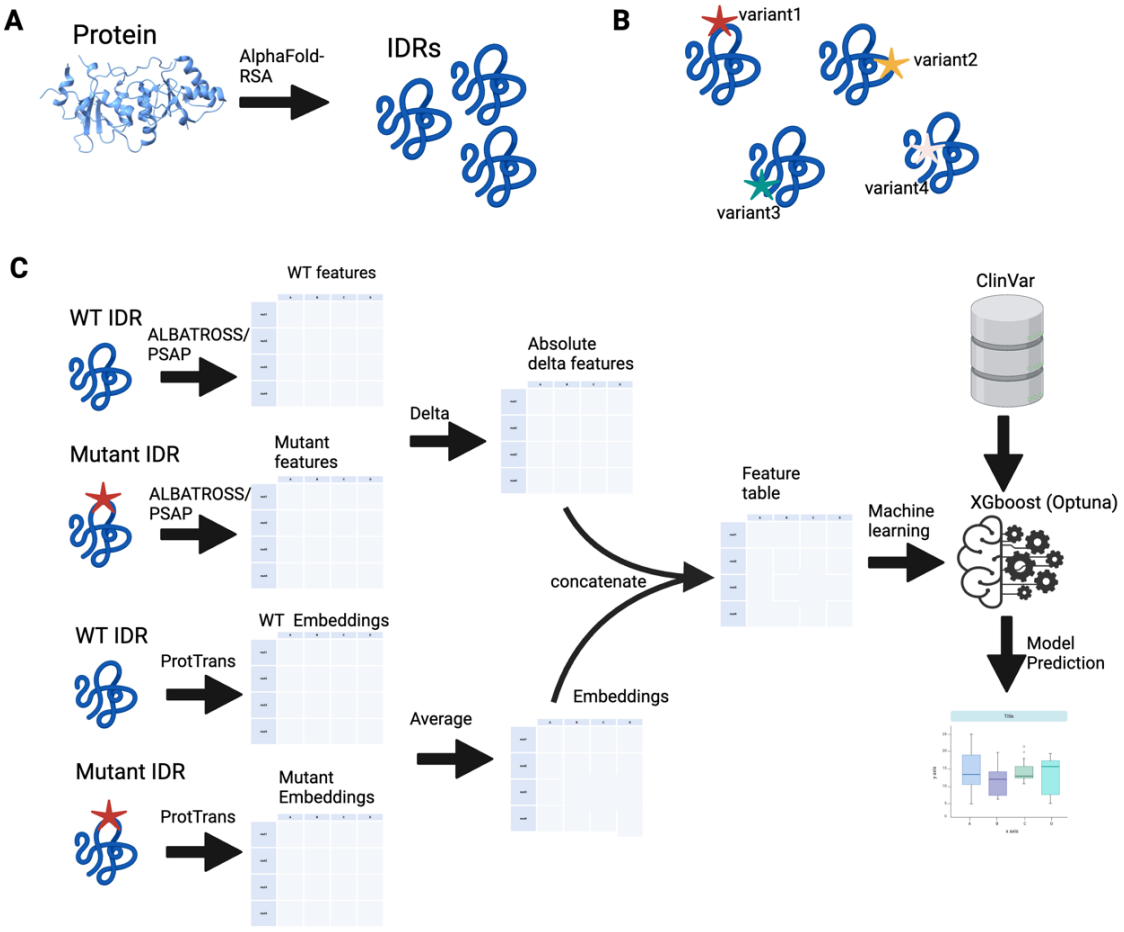

Figure 9: Illustration figure of the computational framework for predicting the impact of missense variants in IDRs*. (A) Predictions from AlphaFold-RSA were used to predict the IDRs from AlphaFold structures. (B) Missense variants were introduced into the FASTA sequence of the predicted IDR sequence after mapping the protein coordinates with the genomic coordinates listed in the ClinVar database. Different variants (variant1, variant2, variant3, variant4) represent variants occurring at various positions of the IDRs. (C) Features were extracted from both wild-type (WT) and mutant IDRs using ALBATROSS and PSAP to capture biophysical properties predictive of global conformation and phase separation. The absolute delta between WT and mutant features was computed. Additionally, ProtTrans embeddings were generated for both WT and mutant sequences. The embeddings were combined using the average. The extracted features and embeddings were concatenated into a feature table and used as input for an XGBoost model optimized with Optuna. The baseline model was trained using ClinVar classifications as ground truth for predictions. The final model outputs classification results evaluated using performance metrics such as AUC and PR-AUC. Created with BioRender.com.*

3.3.2 Data selection

The variants used in this study were downloaded from the ClinVar 2024-09-17 release, with variants located in Pfam domains filtered out as per the coordinates downloaded from University of California Santa Cruz (UCSC) resources[101]. The ClinVar VCF file was annotated with CAVA v2[102] to determine the protein substitution from genomic variants, leveraging transcripts from NCBI and EMBL-EBI (MANE) transcripts that corresponds to UniProt protein FASTA reference. IDRs predicted by AlphaFold2-RSA were observed in 834 genes that mapped to human NCBI Taxon ID 9606. To align these predictions with ClinVar transcript annotations, Ensembl transcripts associated with UniProt IDs were mapped to NCBI transcripts. This process ensured that protein coordinates were consistently aligned with genomic coordinates based on transcript information from the annotated ClinVar VCF, resulting in a total of 15,999 variants located in predicted IDR regions. Based on the clinical classification in ClinVar, 85.7% of the variants in IDRs are classified as Variant of Uncertain Significance (VUS) (Figure. S18A).

The ClinVar classification of missense variants was grouped into three categories, which were Deleterious, Neutral and VUS. Variants labeled as ‘Likely_pathogenic’, ‘Pathogenic/Likely_pathogenic’, ‘Pathogenic|drug_response’, ‘Pathogenic’, ‘Pathogenic|other’, ‘Pathogenic/Likely_pathogenic|other’, ‘Likely_pathogenic|other’, and ‘Likely_pathogenic/Likely_risk_allele’ were classified as Deleterious, while those categorized as ‘Likely_benign’, ‘Benign’, and ‘Benign/Likely_benign’ were classified as Neutral, and the remaining was categorized as VUS. Further refinement to include only

genes with Ensembl transcripts that could be mapped to an orthogonal RefSeq transcript reduced the dataset to a total of 2,203 variants (Figure. S18B).

To evaluate the performance of existing predictors for variants found in predicted IDRs, we utilized data from dbNSFP v4.8[62], which includes a comprehensive set of *in silico* missense prediction scores from 56 predictors, including un-supervised models AlphaMissense, ESM1b, and EVE. Mapping all variants from the dbNSFP database to ClinVar variants located within predicted IDR regions resulted in a final dataset of 2,104 missense variants with known classifications in 290 genes (Table S12). In total, the dataset comprised of 316 variants classified as deleterious and 1788 variants classified as neutral. The predictive performance of each *in silico* missense predictor was assessed using its normalized rank scores available in the dbNSFP database.

3.3.3 Feature generation

Features such as radius of gyration, end-to-end distance, Asphericity, and Prefactor predicting gIDRc were generated using ALBATROSS within Sparrow v0.2.3 for both mutant and WT using the input FASTA sequences representing IDRs. Features predicting PS were evaluated using PSAP v1.0.7, which computed 15 features for wild-type (WT) and mutant sequences. To incorporate sequence-level representations, we used the pLM ProtTrans, generating 2048-dimensional embeddings for both mutant and WT sequences for every amino acid. To derive a single representative embedding per sequence, the mean

of the amino acid-level embeddings across the entire protein chain was computed. Prior to model training, we applied supervised feature selection using F-statistic scoring to reduce protein embedding dimensions from their original high-dimensional space to 20 dimensions, retaining only the most predictive embedding features while preserving all other features and the specific standalone predictor for each model. These combined features of gIDRc, PS, and deep learning-based embeddings provided a comprehensive representation of the potential deleterious impact of missense variants in IDRs.

3.3.4 Data preprocessing

The features generated by ALBATROSS and PSAP for both mutant and WT IDR sequences were used to compute the delta change, capturing the difference between the two sequence states. The absolute values of these deltas were then calculated to ignore the direction of change and quantify the magnitude of change independent of direction. Additionally, to evaluate the underlying representation of different embedding combination approaches to protein function, the pLM embeddings generated by ProtTrans for the mutant and WT sequences were calculated using four different combination strategies, which were L1 (absolute difference), L2 (Euclidean distance), average (mean of embeddings), and Hadamard product (element-wise dot product).

Let E_m and E_w represent the embeddings from the mutant and WT, respectively. We define Hadamard, Average, L1 and L2 as follows:

$$Hadamard\ (E_m, E_w) = f(E_m).f(E_w)$$

$$Average\ (E_m, E_w) = \frac{f(E_m) + f(E_w)}{2}$$

$$L1\ (E_m, E_w) = |f(E_m) - f(E_w)|$$

$$L2(E_m, E_w) = |f(E_m) - f(E_w)|^2$$

3.3.5 Model creation

The features derived from ALBATROSS, PSAP, and the ProtTrans embeddings were generated for a total of 2,104 variant, comprising 316 deleterious and 1788 benign variants. These features were concatenated to form a comprehensive dataset for model evaluation. To assess predictive performance, we compared different model predictors, which were, Random Forest, Multi-Layer Perceptron (MLP), Naïve Bayes, and XGBoost classifiers using scikit-learn v0.24.2. Prior to model training, the dataset was standardized using the StandardScaler from scikit-learn, and the class labels were binarized using LabelEncoder. We employed 10-fold cross-validation to ensure robust model evaluation from an imbalanced dataset, computing key performance metrics including AUC and PR-AUC to compare model effectiveness in distinguishing deleterious from benign variants.

3.3.6 Hyperparameter optimization (HPO)

We employed the Optuna framework to optimize the hyperparameters of the XGBoost model. The dataset was split into an 80:20 ratio randomly, where 80% of the data (1,683 variants) was used in 10-fold stratified cross-validation to determine the optimal

hyperparameters, with the stratification ensuring that there is even distribution of classes within each fold. In a subsequent experiment, we implemented gene-based splitting to prevent data leakage, ensuring that variants from the same gene were not present in both training and testing sets. This resulted in 1852 variants in the training and 252 variants in the test. A total of 150 optimization trials were conducted, with the objective of maximizing the PR-AUC across the 10-fold stratified cross-validation. Once the optimal hyperparameters were identified, the final tuned model was applied to the hold-out test set to evaluate its performance on unseen data.

3.3.7 Combination with AlphaMissense, ESM1b and EVE

The XGBoost model was retrained, and HPO was performed with AlphaMissense, ESM1b and EVE added separately as additional features to assess their impact on predictive performance. The experimental design-maintained consistency with our initial approach, utilizing both random and gene-based data splitting strategies. For gene-based splitting, the dataset was partitioned to prevent data leakage, ensuring variants from the same gene were not present in both training (1,852 variants) and testing (252 variants) sets. HPO employed the Optuna framework across 150 optimization trials, with the objective of maximizing PR-AUC during 10-fold stratified cross-validation. Following optimization, final models were evaluated on the hold-out test set, with performance confidence intervals calculated through bootstrap resampling over 1,000 iterations to ensure robust statistical assessment.

3.3.8 ClinVar review status classification

To assess whether prediction accuracy of the baseline model depends on variant classification quality, we stratified the hold-out test set according to ClinVar's star-based review status system. Variants labeled as “no assertion criteria provided” or “criteria provided, single submitter” were grouped under the single-star category. Variants with “criteria provided, multiple submitters, no conflicts” were categorized as two stars, while those “reviewed by expert panel” were assigned to the three-star category. This stratification resulted in 292 variants in the one-star category, 100 in the two-star category, and 29 in the three-star category. Within each star category the PR-AUC were calculated, along with its CIs.

3.3.9 Comparative analysis of dbNSFP *in silico* missense predictors and the enhanced predictors

To evaluate whether our methodology demonstrates superior performance on variants where existing missense predictors struggle, we conducted a systematic analysis of variant-level classifications across all *in silico* missense predictors in the dbNSFP database using our hold-out test set. For each IDR variant, we calculated the misclassification rate among existing predictors relative to ClinVar annotations and stratified variants by error rate to identify those that consistently challenge current computational approaches. We then assessed the performance of our baseline model and feature-enhanced variants (incorporating AlphaMissense, ESM1b, or EVE) specifically on these difficult-to-classify

variants. Additionally, we compared the predictions of each enhanced model against its corresponding standalone predictor to quantify the improvement gained through integration with our IDR-specific features.

3.3.10 Statistical analysis

Baseline model performance was assessed using AUC and PR-AUC metrics across four machine learning algorithms: XGBoost, Random Forest, Naïve Bayes, and multilayer perceptron (MLP). The evaluation framework employed 10-fold cross-validation for model training and hyperparameter optimization, followed by performance assessment on an independent hold-out test set to ensure unbiased evaluation. The association between individual biophysical features and variant pathogenicity was assessed using the non-parametric Mann-Whitney U test. To evaluate statistical significance of performance differences between models, we compared the distributions of AUC and PR-AUC scores across cross-validation folds using the Mann-Whitney U test. This approach enabled pairwise comparisons between the baseline model, enhanced models and standalone *in silico* predictors (AlphaMissense, ESM1b, EVE), and final enhanced models. For all performance metrics on the independent test set, 95% confidence intervals were calculated using bootstrap resampling with 1,000 iterations.

3.4 Results

3.4.1 Evaluation of dbNSFP predictors with functions

A significant number of *in silico* missense predictors are trained on ClinVar and HGMD datasets, which exhibit substantial overlap in their training data and the data used in this study. However, 23 models in the dbNSFP database are not trained on ClinVar and HGMD. Among these, when focusing on variants located in predicted IDR regions, AlphaMissense, an unsupervised model, demonstrated the highest predictive performance with an AUC of 0.908 (95% CI:0.889-0.924) (Figure 2A) and PR-AUC of 0.733 (95% CI:0.690-775) (Figure 2B), followed by ESM1b with an AUC of 0.847 (95% CI: 0.819–0.872) and PR-AUC of 0.606 (95% CI:0.548-0.662). Additionally, EVE, the other unsupervised autoencoder model using sequence conservation, achieved an AUC of 0.715 (95% CI: 0.679–0.751) and PR-AUC of 0.50 (95% CI:0.446-0.556), highlighting its moderate performance for missense classification in predicted IDRs. These results, derived from ClinVar-labeled variants located in AlphaFold-RSA predicted disordered regions, suggest that AlphaMissense currently represents the state-of-the-art benchmark for predicting deleterious missense variants in IDRs (Supplementary Table S21).

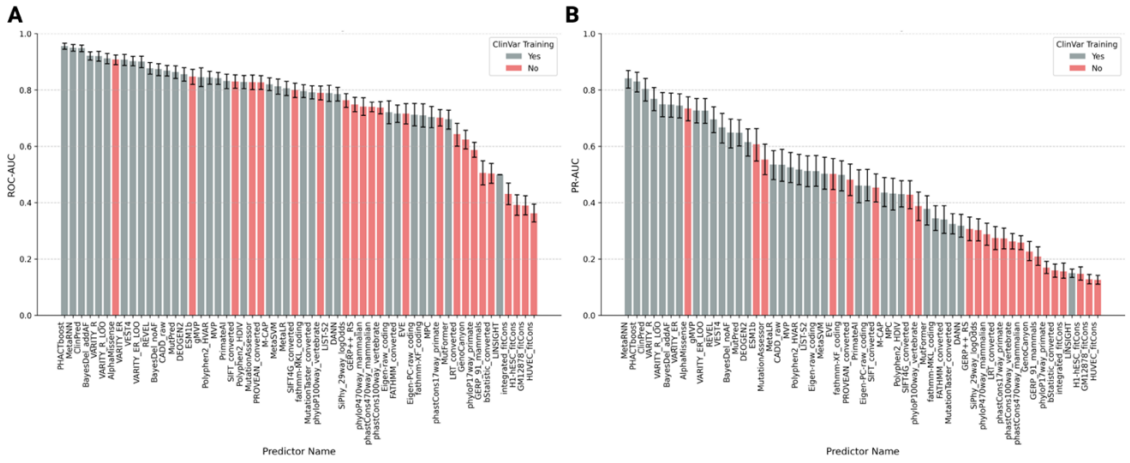

Figure 10: Performance metrics of in silico missense predictors on variants found in IDR regions as per AlphaFold-RSA predictions *A: The area under the curve performance of all in silico missense predictors listed in the dbNSFP databases for variants found in the ClinVar database that is within AlphaFold-RSA predicted IDR regions of the genome. The predictors in red highlight those have not been trained by variants listed in the ClinVar database. B: Precision-Recall performance of all in silico missense predictors listed in the dbNSFP databases for variants found in the ClinVar database that is within AlphaFold-RSA predicted IDR regions of the genome. The predictors in red highlight those have not been trained by variants listed in the ClinVar database*

3.4.2 Association of global IDR conformation with protein function

Using a Mann-Whitney U test, we evaluated the association between protein function and the absolute changes in radius of gyration, end-to-end distance, Asphericity, and Prefactor between the WT and mutant sequences. All four features demonstrated a significant association with protein function. To examine their distribution across deleterious and neutral categories, a square root transformation was applied to all four features (Figure. 15A). Further evaluation of their predictive performance based on the AUC metric, revealed that the absolute change in scaling exponent had the highest AUC of 0.628. Notably, all five features achieved an AUC greater than random prediction (Figure. 15B), highlighting their potential relevance in impact assessment.

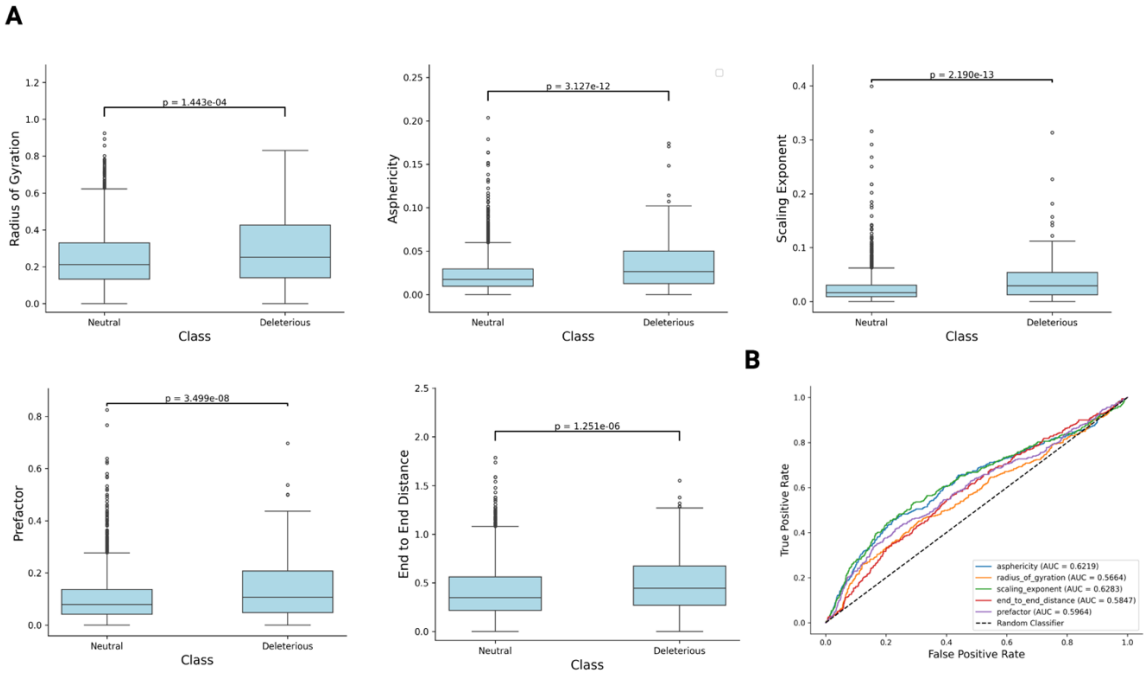

Figure 11: Absolute change in global protein conformation due to a missense variant in an IDR contributes significantly to protein prediction. *(A) Association of the absolute change in Radius of Gyration, Asphericity, Scaling Exponent, Prefactor, End to End distance transformed using the square root between the reference and mutant with protein function. Outliers were removed to observe the overall distribution. (B) Area under the curve performance highlighting the prediction of protein function from absolute change in Radius of Gyration, Asphericity, Scaling Exponent, Prefactor, End to End distance induced due to a missense variant.*

3.4.3 Evaluation of embedding combination method and model performance

We evaluated four different embedding combination methods (L1, L2, average, and Hadamard) across four different machine learning and deep learning models (XGBoost, Random Forest, Naïve Bayes, and MLP) to predict protein function Using PR-AUC and AUC across 10-fold stratified cross-validation, we found that the average and Hadamard combination methods, when integrated with features from PS and gIDRc and modeled using either XGBoost or Random Forest, achieved the highest predictive performance.

Specifically, with XGBoost, both the average and Hadamard methods yielded high mean AUC values of 0.904 and 0.908, respectively, with no statistically significant difference between them (Figure. 16A and S3 table). Similarly, both methods achieved a mean PR-AUC of 0.76 and 0.74 with XGBoost, again showing no significant difference (Figure. 16B). When evaluated using Random Forest, the average and Hadamard methods also exhibited comparable performance. However, models utilizing Hadamard and average embeddings with XGBoost and Random Forest significantly outperformed those using L1 and L2 embeddings, demonstrating statistically significant improvements in both AUC and accuracy. The default hyperparameters for XGBoost included learning_rate=0.3, n_estimators=100, max_depth=6, gamma=0, and subsample=1.0. Given its robust performance, we selected XGBoost trained on the average embedding combination, along with features from gIDRc and PS, as the final model for further optimization.

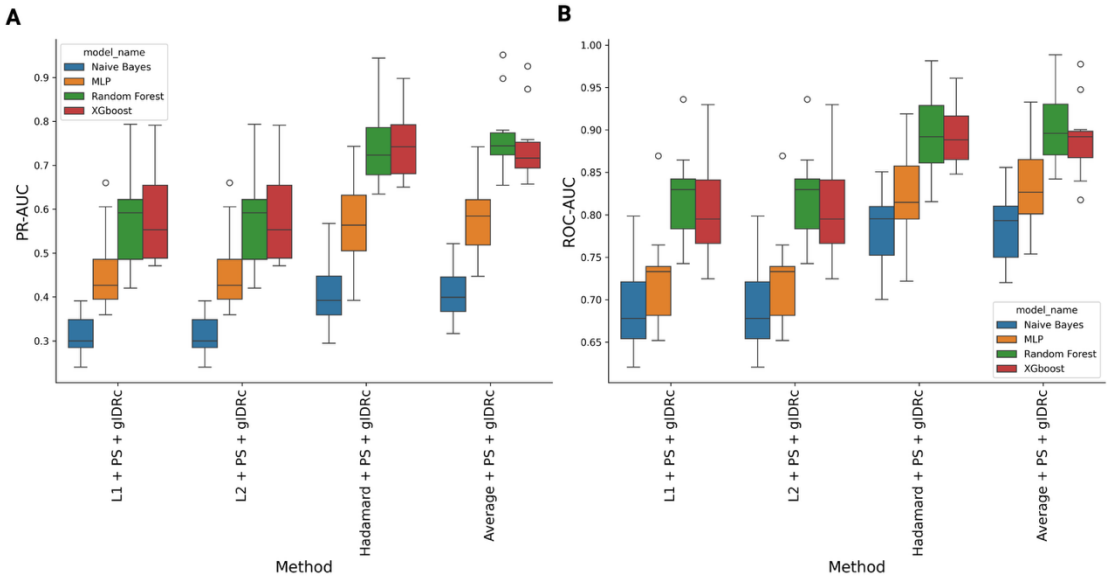

Figure 12: Performance comparison of multiple protein embedding combination methods using multiple machine learning models to predict protein function on classified variants found in ClinVar in predicted IDR regions. *(A) Accuracy performance of XGBoost, Random Forest, Naïve Bayes and Multi-Layer*

Perceptron using protein language model embeddings (L1, L2, average and Hadamard), along with features predicting absolute change in Phase Separation and absolute change in global IDR conformation to predict protein function from a missense variant. (B) Area under the curve performance of XGBoost, Random Forest, Naïve Bayes and Multi-Layer Perceptron using protein language model features (L1, L2, average and Hadamard), along with features predicting absolute change in Phase Separation and absolute change in global IDR conformation to predict protein function from a missense variant.

3.4.4 Hyperparameter optimization (HPO)

Using the Optuna framework, we further optimized the XGBoost model through 150 trials, aiming to maximize the mean AUC across 10-fold cross-validation. Each trial represents a new set of hyperparameters that is tested. HPO resulted in an increase in mean PR-AUC from 0.802 to 0.840, demonstrating a slight but meaningful improvement in predictive performance from the training set. When applied to the hold-out test set, which comprised of 421 unique variants, the optimized model achieved an PR-AUC of 0.800 (95% CI 0.7113-0.876), which we refer to as the baseline model. To ensure there was no data leakage, we computed pairwise sequence identity between sequences in the training and test sets and found that 0% of sequences in the test set had 100% identity with those in the training set. The final set of optimal hyperparameters identified included: n_estimators=330, max_depth=10, learning_rate=0.077, colsample_bytree=0.772, subsample=0.505, gamma=624, min_child_weight=3, reg_alpha=1.68, and reg_lambda=5.14. Among these, “gamma”, “reg_alpha”, and “min child weight” were found to be the most influential and showed the strongest correlations with PR-AUC performance. The negative correlations indicate that lower regularization and reduced constraints on tree splitting improved model performance, suggesting the model benefits from increased flexibility rather than regularization constraints (Figure. S19).

3.4.5 Feature evaluation

Utilizing SHapley Additive exPlanations (SHAP)[103] to assess feature importance in the optimized XGBoost model revealed that the embeddings combined via the average method between the WT and mutant were the most influential (Figure. S20). The analysis demonstrated that PLM embeddings from ProtTrans constituted the most influential features, with the highest-ranking feature achieving a SHAP importance value of approximately 0.09. PS features from PSAP and gIDRc features from ALBATROSS were prominently distributed throughout the importance hierarchy, with several features from each category ranking within the top 20 most influential predictors. The balanced representation of features from ProtTrans, PSAP, and ALBATROSS within the top performing features validates our integrative approach, demonstrating that PLM embeddings, PS propensity, and gIDRc provide distinct yet synergistic information for predicting missense variant pathogenicity.

3.4.6 Improvement with AlphaMissense, ESM1b and EVE

As noted above, AlphaMissense, ESM1b and EVE were the highest performing unsupervised classification models. To improve these metrics, we retrained the XGBoost model and optimized its hyperparameters to create an “Enhanced” version. Enhanced model creation begins with one of AlphaMissense, ESM1b or EVE. Next, PS, gIDRc and the average of the embeddings between the WT and mutant are added. The dataset was

split into 80% training data (1,683 variants), where 10-fold cross validation was performed, and 20% (421 variants) was reserved as a hold-out test set.

EVE incorporation as a feature also demonstrated significant performance gains. The EVE-enhanced model achieved a mean PR-AUC of 0.813 in 10-fold cross-validation, substantially outperforming standalone EVE (PR-AUC: 0.506; 95% CI: 0.475-0.537; $p < 0.001$; Cohen's d = 5.46). Hold-out test validation showed EVE-enhanced achieving an PR-AUC of 0.918 (95% CI: 0.860-0.962) compared to standalone EVE, with a PR-AUC of 0.591 (95% CI: 0.471-0.705; $p < 0.001$; Cohen's d = 6.774) (Figure 5A). Optimal hyperparameters were n_estimators=466, max_depth=10, learning_rate=0.029, colsample_bytree=0.863, subsample=0.602, gamma=0.877, min_child_weight=1, reg_alpha=0.38, and reg_lambda=7.66 (S4 Table).

ESM1b integration demonstrated equally substantial performance enhancement. The ESM1b-enhanced model achieved a mean PR-AUC of 0.840 in 10-fold cross-validation from the training set, significantly outperforming standalone ESM1b (PR-AUC: 0.622; 95% CI: 0.561-0.682; $p < 0.001$; Cohen's d = 3.25). The 35.19% increase in PR-AUC substantially enhances deleterious variant detection capability. Hold-out test validation confirmed superior performance with PR-AUC of 0.877 (95% CI: 0.809-0.929) versus standalone ESM1b PR-AUC of 0.679 (95% CI: 0.563-0.780; $p < 0.001$; Cohen's d = 4.35) (Figure 5B). Optimal hyperparameters included: n_estimators=330, max_depth=10,

learning_rate=0.077, colsample_bytree=0.772, subsample=0.505, gamma=0.624, min_child_weight=3, reg_alpha=1.68, and reg_lambda=5.14.

The AlphaMissense-enhanced model achieved a mean PR-AUC of 0.863 representing a significant improvement over standalone AlphaMissense (PR-AUC: 0.740, 95% CI: 0.680-0.797; p = 0.0017, Cohen's d = 1.84) from the training set. The 16.78% improvement in PR-AUC is clinically significant as it directly measures enhanced detection of deleterious variants. Independent validation on the hold-out test set confirmed this enhancement, with PR-AUC improving from 0.807 (95% CI: 0.715-0.885) to 0.931 (95% CI: 0.881-0.971; p < 0.001; Cohen's d = 3.882), representing a 16.52% improvement (Figure 5C). This performance rivals top *in silico* missense predictors trained on ClinVar variants, including MetaRNN (PR-AUC: 0.841, 95% CI: 0.806-0.869) and PHACTboost (PR-AUC: 0.830, 95% CI: 0.793-0.863). Optimal hyperparameters were determined through Optuna optimization: n_estimators=390, max_depth=3, learning_rate=0.068, colsample_bytree=0.83, subsample=0.595, gamma=1.50, min_child_weight=1, reg_alpha=1.12, and reg_lambda=1.192.

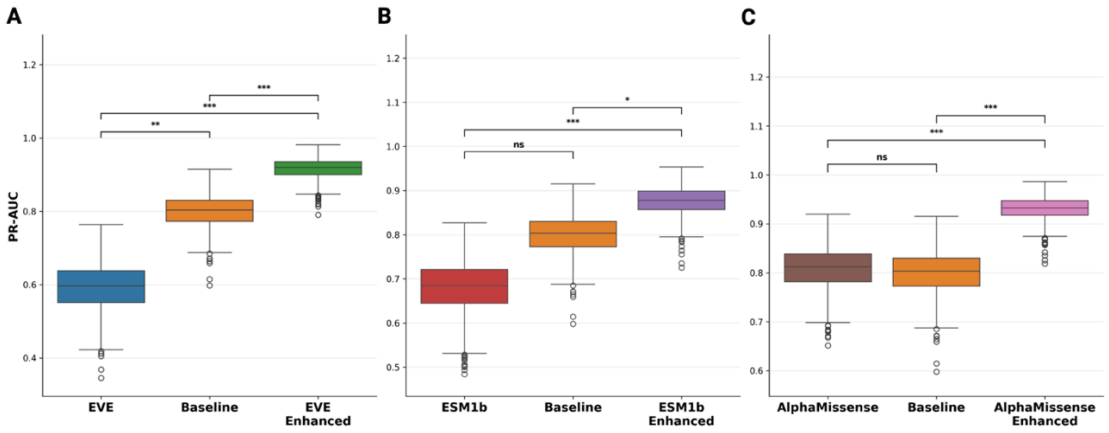

Figure 13: Performance comparison of protein variant effect prediction methods. PR-AUC metrics comparing standalone predictors versus baseline and enhanced models across EVE, ESM1b, and AlphaMissense methods on the hold-out test set*. (A) PR-AUC performance with confidence intervals using bootstrap resampling with 1,000 iterations comparing EVE vs Baseline vs EVE enhanced. (B) PR-AUC performance with confidence intervals using bootstrap resampling with 1,000 iterations comparing ESM1b vs Baseline vs ESM1b enhanced (C) PR-AUC performance with confidence intervals using bootstrap resampling with 1,000 iterations comparing AlphaMissense vs Baseline vs AlphaMissense enhanced. Statistical significance was assessed using pairwise Mann-Whitney U tests: ***p < 0.001, **p < 0.01, *p < 0.05, ns = not significant.*

3.4.7 Model Performance on ClinVar variants based on review status

Overall, the baseline model achieved a PR-AUC of 0.800 (95% CI 0.711-0.876) on a hold-out test set comprising 421 variants. Stratification by ClinVar review status revealed quality-dependent performance patterns, with one-star variants (lowest evidence quality, n=292) achieved PR-AUC of 0.724 (95% CI: 0.592-0.833), two-star variants (moderate evidence quality, n=100) achieved higher PR-AUC of 0.824 (95% CI: 0.639-0.945), and three-star variants (expert panel review, n=29) achieved PR-AUC of 0.700 (95% CI: 0.100-1.000). The performance estimate for three-star variants has substantial uncertainty, as evidenced by the confidence interval spanning nearly the entire possible range due to the limited sample size. Better performance on two-star variants suggests the model performance improved on variants with better evidence quality (Figure. S21).

3.4.8 Comparative Analysis of *In Silico* Missense Predictors and Enhanced Model Performance

Evaluation of our base model against established *in silico* predictors revealed superior performance on challenging variants within the hold-out test dataset. Several variants showed consistently high error rates across most predictors, highlighting the challenge in predicting these variants with high accuracy. Our model correctly classified several variants that were consistently misclassified by the majority of dbNSFP predictors. For instance, likely pathogenic variants R573C in FGA and A654V in HIF1A were incorrectly predicted as neutral by most existing tools, including AlphaMissense, while our model accurately identified them as deleterious. Conversely, likely benign variants T319M and P869S in GLI3, and S186Y in BRCA1, were incorrectly predicted as damaging by many *in silico* predictors, whereas our model correctly classified them as neutral. These results demonstrate our model's enhanced capability to handle challenging variants that represent common failure modes for existing prediction tools (Figure S22).

Enhanced models demonstrated substantial improvements over their standalone counterparts across all three integrated predictors. The AlphaMissense enhanced model, using an optimal cut-point of 0.028, correctly classified 32 variants that were misclassified by standalone AlphaMissense, including 8 pathogenic/likely pathogenic and 24 benign/likely benign variants (Supplementary Table S6). Similarly, the ESM1b enhanced model (optimal cut-point: 0.123) accurately classified 28 variants missed by standalone

ESM1b, comprising 10 pathogenic/likely pathogenic and 18 benign/likely benign variants. The EVE enhanced model showed the most substantial improvement, correctly identifying 47 variants that standalone EVE misclassified (27 pathogenic/likely pathogenic and 20 benign/likely benign variants) using an optimal cut-point of 0.028. These results demonstrate consistent enhancement across all integrated predictors, with strength in reducing false negative classifications of pathogenic variants while maintaining specificity for benign variants.

3.5 Discussion

The relatively lower predictive performance of *in silico* missense classification in IDRs, compared to structured domains is well-documented[60, 87]. This discrepancy may stem from the fact that most known pathogenic missense variants reside in ordered protein regions. Due to their lack of well-defined secondary structures, low evolutionary conservation, and highly dynamic nature, IDRs present significant challenges for accurate prediction of missense classification. Many computational models are overfit to structured regions, leading to biased predictions that underperform for disordered region variants. Given these challenges, new computational strategies incorporating biophysical properties unique to disordered regions are necessary to enhance the classification of missense variants within IDRs.

PLMs like ProtTrans, combined with biophysical predictors such as Albatross and PSAP, offer a powerful approach to capturing biophysical information embedded within protein sequences, providing valuable insights into predicted consequences. The optimal method for integrating embeddings from WT and mutant sequences remains system dependent. In this study, we found that the Hadamard product and average embedding combination methods outperformed L1 and L2 distance-based approaches, demonstrating their effectiveness in assessing the impact of missense variants in IDRs.

Recent studies have shown an association between changes in PS and overall protein function[88]. Our findings further highlight that alterations in gIDRc also contribute significantly to predictions, though it is not the sole predictor. Notably, we demonstrate that gIDRc, PS, and protein sequence embeddings create a robust framework for predicting the consequences of missense variants in IDRs. This approach significantly outperforms leading *in silico* missense models, including AlphaMissense, ESM1b and EVE.

Moreover, incorporating AlphaMissense, ESM1b, or EVE as additional features further enhances predictive accuracy, suggesting that gIDRc, PS, and protein embeddings provide independent and complementary information. Feature importance analysis using SHAP further supports this conclusion, revealing that gIDRc, PS, and protein embeddings rank among the top predictive features, underscoring their critical role in classification of IDR-associated missense variants. The improvement in prediction classification performance is

now comparable to top *in silico* missense classification methods trained directly on ClinVar annotations.

Our study has several limitations that provide avenues for future work. First, our dataset relied on computational predictions of disorder by AlphaFold-RSA rather than experimentally validated IDRs. Although filtering variants within known Pfam domains increased confidence in our disorder predictions, this approach reduced dataset size and precluded comprehensive validation against experimentally confirmed IDRs. Notably, only 18 neutral variants in our hold-out test set were confirmed to reside within validated DisProt regions, all of which were correctly predicted by the three enhanced models. The use our gene-based data splitting strategy, while necessary to prevent information leakage, resulted in a relatively small and heterogeneously distributed test set that limited the statistical power of our evaluation. Future studies would benefit from larger, more balanced datasets that incorporate experimentally validated disordered regions.

Second, the general sparsity of high-confidence, experimentally annotated pathogenic variants in IDRs presents a major challenge for training and robustly evaluating any predictive model. To address this, future studies could employ semi-supervised learning to leverage the vast amount of unlabeled variant data or use generative models to augment training sets with biologically plausible synthetic variants.

Finally, there is a remaining knowledge gap in the specific molecular mechanisms by which IDR variants exert their effects, such as altering post-translational modifications, cellular signaling, or protein self-assembly. Further investigation into these downstream consequences is necessary to fully understand the functional impact of variation in disordered regions.

While computational predictors have significantly improved missense variant classification, major limitations persist in their ability to assess IDRs. Traditional models overemphasize evolutionary conservation and structural stability, resulting in biased predictions that fail to capture the complexity of IDRs. Our baseline model introduces an IDR specific framework that integrates predictions of global conformation changes, PS dynamics, and deep learning-based embeddings to refine missense variant classification in IDRs. By addressing the shortcomings of existing approaches, this study advances the accurate classification of IDR variants and enhances our understanding of their role in disease.

Chapter 4: Conclusions and translational significance of *in silico* missense classification

4.1 Summary

In silico missense prediction tools is one of many methods employed by a clinician to support the functional variant classification in clinical genetics, with modern tools like AlphaMissense, EVE, ESM1b, REVEL, and BayesDel gaining widespread adoption. Despite significant advances, these tools still require substantial improvement in accuracy, particularly for variants in intrinsically disordered protein regions, which account for over 25% of known deleterious variants. Our research highlights the use of generative AI models and its value in study of missense variants, as well as improving the prediction of these well-known *in silico* missense predictors performance in regions of the proteins, that are predicted to be IDRs.

The emergence of generative protein structure prediction models, particularly AlphaFold2 and ESMFold, has revolutionized our ability to investigate both ordered and disordered protein regions. These AI-driven tools have democratized access to high-quality protein structures, enabling researchers to study protein modifications and variant effects without relying solely on expensive experimental crystallography. Our study demonstrates that generative protein structure prediction models achieve remarkable accuracy in ordered regions of four critical tumor suppressor proteins (BRCA1, BRCA2, PALB2, and RAD51C) associated with breast cancer susceptibility. This validates the use of AI-predicted structures for variant effect prediction in clinically relevant genes. We highlight

how AI can significantly accelerate the study of protein complex modifications resulting from missense variants, providing researchers with rapid, cost-effective alternatives to experimental approaches for initial variant assessment.

Through correlation with MAVE experimental data across all four tumor suppressor genes, we demonstrate strong associations between predicted protein perturbations and actual functional outcomes. This validation bridges computational predictions with experimental reality. A key finding shows that predicted $|\Delta\Delta G|$ values from generative AI models perform comparably to, or even better than, those derived from experimentally-determined crystal structures. This superiority likely stems from AI models' ability to capture dynamic side-chain conformations and protein flexibility, whereas crystal structures represent static, artificially constrained protein states that may not reflect physiological conditions.

Our study demonstrates that traditional computational features, including evolutionary conservation, secondary structure descriptors, and physicochemical properties are predominantly optimized for predicting variant effects in ordered protein regions, explaining their suboptimal performance in IDRs. However, recent breakthroughs in protein language models, conformational prediction algorithms, and phase separation modeling have opened new areas for understanding missense variant impacts in IDRs. These emerging computational approaches provide critical insights that traditional methods cannot capture. Protein language models can extract sequence-level representations that encode functional relationships beyond simple conservation patterns. Conformational prediction tools can model the dynamic flexibility and ensemble properties that define IDR behavior. Phase separation predictors can assess how variants affect the

formation of biomolecular condensates, which are increasingly recognized as crucial for cellular organization and disease mechanisms.

The integration of these novel IDR-specific features with established prediction frameworks creates a synergistic effect that substantially enhances overall missense variant classification. Rather than replacing existing tools, our approach demonstrates that combining traditional ordered-region predictors with IDR-specialized features produces superior performance. This strategy leverages the proven strengths of conventional methods for structured domains while addressing their inherent limitations in disordered regions, ultimately advancing the field toward more comprehensive and accurate variant interpretation capabilities.

4.2 Leason learned and future directions

4.2.1 Leason's learned and future direction in Chapter 2

Chapter 2 demonstrates several key findings that advance our understanding of AI-driven protein structure prediction and stability analysis in cancer genetics. AlphaMissense was highlighted show the highest average AUC all the know ordered domains in BRCA1, BRCA2, RAD51C, and PALB2. AF2 consistently outperformed ESMFold in predicting protein structures with lower RMSD metrics compared to experimentally-derived structures across ordered domains of critical breast cancer genes, establishing AF2 as the superior choice for structural biology applications. Among protein stability predictors, FoldX emerged as the most effective tool, showing stronger associations with protein

function and superior AUC performance compared to Rosetta and DDGun3D when validated against functional evidence from MAVE assays. Remarkably, stability predictions generated using AI-predicted AF2 structures demonstrated high correlation with those from experimentally-derived structures, validating the potential of generative AI models to accelerate and democratize study of protein perturbation based on missense variant. However, these findings are constrained to functionally validated loss-of-function mutations in four haploinsufficient tumor suppressor genes within the homologous DNA repair pathway. Future research should expand this framework to encompass genes with diverse functional mechanisms, including gain-of-function and dominant-negative effects, while also exploring disease-associated variants across different biological pathways beyond DNA repair. This expansion would provide a more comprehensive understanding of how AI-predicted structures and stability analysis can be applied broadly across human genetics and precision medicine.

4.2.2 Leason's learned and future direction in Chapter 3

Chapter 3 presents a methodology that significantly enhances the performance of leading *in silico* missense prediction models by incorporating features specifically designed to capture deleterious effects in IDRs, as identified through AlphaFold2-RSA predictions. Among the global IDR conformation features analyzed, asphericity emerged as the most predictive individual feature for functional assessment, highlighting the critical role of protein shape dynamics in determining variant pathogenicity within disordered regions. The study demonstrated that PLM embeddings of wild-type and mutant IDR sequences,

when combined using an averaging method, provided superior predictive power compared to other combination strategies based on AUC and accuracy metrics. The integration of PLM embeddings with phase separation and global IDR conformation features through an XGBoost framework, combined with existing unsupervised models (AlphaMissense, ESM1b, and EVE), yielded substantial improvements in prediction outcomes, establishing a new benchmark for IDR variant classification. However, this work is currently limited to variants catalogued in the ClinVar database rather than comprehensive analysis of all possible missense mutations, representing a significant opportunity for future expansion. Additionally, future research should explore additional feature sets that could provide deeper insights into the functional consequences of IDR variants, while also developing novel functional classification assays specifically designed to validate IDR variant effects experimentally, thereby creating a more robust foundation for computational model training and validation.

Bibliography

[1] W. A. B. B K Shoichet, R Kuroki, B W Matthews, "A relationship between protein stability and protein function.," *Proc. Natl. Acad. Sci. ,* no. 92, pp. 452–456 1995.

[2] A. Stein, D. M. Fowler, R. Hartmann-Petersen, and K. Lindorff-Larsen, "Biophysical and Mechanistic Models for Disease-Causing Protein Variants," *Trends Biochem Sci,* vol. 44, no. 7, pp. 575-588, Jul 2019, doi: 10.1016/j.tibs.2019.01.003.

[3] M. J. Landrum *et al.*, "ClinVar: improving access to variant interpretations and supporting evidence," *Nucleic Acids Res,* vol. 46, no. D1, pp. D1062-D1067, Jan 4 2018, doi: 10.1093/nar/gkx1153.

[4] C. Ernst *et al.*, "Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics," *BMC Med Genomics,* vol. 11, no. 1, p. 35, Mar 27 2018, doi: 10.1186/s12920-018-0353-y.

[5] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, and Z. Zhang, "Assessment of computational methods for predicting the effects of missense mutations in human cancers," *BMC Genomics,* vol. 14 Suppl 3, no. Suppl 3, p. S7, 2013, doi: 10.1186/1471-2164-14-S3-S7.

[6] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res,* vol. 31, no. 13, pp. 3812-4, Jul 1 2003, doi: 10.1093/nar/gkg509.

[7] J. M. Schwarz, C. Rodelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," *Nat Methods,* vol. 7, no. 8, pp. 575-6, Aug 2010, doi: 10.1038/nmeth0810-575.

[8] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res,* vol. 47, no. D1, pp. D886-D894, Jan 8 2019, doi: 10.1093/nar/gky1016.

[9] D. Quang, Y. Chen, and X. Xie, "DANN: a deep learning approach for annotating the pathogenicity of genetic variants," *Bioinformatics,* vol. 31, no. 5, pp. 761-3, Mar 1 2015, doi: 10.1093/bioinformatics/btu703.

[10] S. Richards *et al.*, "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genet Med,* vol. 17, no. 5, pp. 405-24, May 2015, doi: 10.1038/gim.2015.30.

[11] J. Weile and F. P. Roth, "Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas," *Hum Genet,* vol. 137, no. 9, pp. 665-678, Sep 2018, doi: 10.1007/s00439-018-1916-x.

[12] P. D. Stenson *et al.*, "The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting," *Hum Genet,* vol. 139, no. 10, pp. 1197-1207, Oct 2020, doi: 10.1007/s00439-020-02199-3.

[13] E. Porta-Pardo, V. Ruiz-Serra, S. Valentini, and A. Valencia, "The structural coverage of the human proteome before and after AlphaFold," *PLoS Comput Biol,* vol. 18, no. 1, p. e1009818, Jan 2022, doi: 10.1371/journal.pcbi.1009818.

[14] P. Bross, T. J. Corydon, B. S. Andresen, M. M. J□rgensen, L. Bolund, and N. Gregersen, "Protein misfolding and degradation in genetic diseases," *Human Mutation,* vol. 14, no. 3, pp. 186-198, 1999, doi: 10.1002/(sici)1098-1004(1999)14:3<186::Aid-humu2>3.0.Co;2-j.
[15] C. Ferrer-Costa, M. Orozco, and X. de la Cruz, "Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties," *J Mol Biol,* vol. 315, no. 4, pp. 771-86, Jan 25 2002, doi: 10.1006/jmbi.2001.5255.
[16] S. Khan and M. Vihinen, "Performance of protein stability predictors," *Hum Mutat,* vol. 31, no. 6, pp. 675-84, Jun 2010, doi: 10.1002/humu.21242.
[17] R. Puglisi, "Protein Mutations and Stability, a Link with Disease: The Case Study of Frataxin," *Biomedicines,* vol. 10, no. 2, Feb 11 2022, doi: 10.3390/biomedicines10020425.
[18] P. Yue, Z. Li, and J. Moult, "Loss of protein structure stability as a major causative factor in monogenic disease," *J Mol Biol,* vol. 353, no. 2, pp. 459-73, Oct 21 2005, doi: 10.1016/j.jmb.2005.08.020.
[19] M. Petrosino *et al.*, "Analysis and Interpretation of the Impact of Missense Variants in Cancer," *Int J Mol Sci,* vol. 22, no. 11, May 21 2021, doi: 10.3390/ijms22115416.
[20] L. N. Salmena, S, "BRCA1 Haploinsufficiency: Consequences for Breast Cancer," *Womens Health* vol. 8, pp. 127–129, 2012.
[21] J. Pereira, A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan, and A. N. Lupas, "High-accuracy protein structure prediction in CASP14," *Proteins,* vol. 89, no. 12, pp. 1687-1699, Dec 2021, doi: 10.1002/prot.26171.
[22] J. Won, M. Baek, B. Monastyrskyy, A. Kryshtafovych, and C. Seok, "Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning," *Proteins,* vol. 87, no. 12, pp. 1351-1360, Dec 2019, doi: 10.1002/prot.25804.
[23] C. J. Wilson, W. Y. Choy, and M. Karttunen, "AlphaFold2: A Role for Disordered Protein/Region Prediction?," *Int J Mol Sci,* vol. 23, no. 9, Apr 21 2022, doi: 10.3390/ijms23094591.
[24] K. M. Ruff and R. V. Pappu, "AlphaFold and Implications for Intrinsically Disordered Proteins," *J Mol Biol,* vol. 433, no. 20, p. 167208, Oct 1 2021, doi: 10.1016/j.jmb.2021.167208.
[25] P. C. M. Lyra, Jr. *et al.*, "Integration of functional assay data results provides strong evidence for classification of hundreds of BRCA1 variants of uncertain significance," *Genet Med,* vol. 23, no. 2, pp. 306-315, Feb 2021, doi: 10.1038/s41436-020-00991-0.
[26] S. E. Brnich *et al.*, "Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework," *Genome Med,* vol. 12, no. 1, p. 3, Dec 31 2019, doi: 10.1186/s13073-019-0690-2.
[27] R. G. Pauli Virtanen, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman,

Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt & SciPy 1.0 Contributors, "SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python," *Nature Methods,* vol. 17, 261–272, 2020.

[28] H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, and R. Karchin, "Identifying Mendelian disease genes with the variant effect scoring tool," *BMC Genomics,* vol. 14 Suppl 3, no. Suppl 3, p. S3, 2013, doi: 10.1186/1471-2164-14-S3-S3.

[29] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Curr Protoc Hum Genet,* vol. Chapter 7, p. Unit7 20, Jan 2013, doi: 10.1002/0471142905.hg0720s76.

[30] L. Gerasimavicius, X. Liu, and J. A. Marsh, "Identification of pathogenic missense mutations using protein stability predictors," *Sci Rep,* vol. 10, no. 1, p. 15387, Sep 21 2020, doi: 10.1038/s41598-020-72404-w.

[31] S. N. Leonardo Salmena, "BRCA1 haploinsufficiency: consequences for breast cancer," *Women's Health,* vol. 8, no. 2, pp. 127-129, 2012.

[32] S. R. Gunnarsdottir *et al.*, "BRCA2 Haploinsufficiency in Telomere Maintenance," *Genes (Basel),* vol. 13, no. 1, Dec 28 2021, doi: 10.3390/genes13010083.

[33] G. Smeenk *et al.*, "Rad51C is essential for embryonic development and haploinsufficiency causes increased DNA damage sensitivity and genomic instability," *Mutat Res,* vol. 689, no. 1-2, pp. 50-8, Jul 7 2010, doi: 10.1016/j.mrfmmm.2010.05.001.

[34] J. Nikkila *et al.*, "Heterozygous mutations in PALB2 cause DNA replication and damage response defects," *Nat Commun,* vol. 4, p. 2578, 2013, doi: 10.1038/ncomms3578.

[35] J. Delgado, L. G. Radusky, D. Cianferoni, and L. Serrano, "FoldX 5.0: working with RNA, small molecules and a new graphical interface," *Bioinformatics,* vol. 35, no. 20, pp. 4168-4169, Oct 15 2019, doi: 10.1093/bioinformatics/btz184.

[36] R. F. Alford *et al.*, "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design," *J Chem Theory Comput,* vol. 13, no. 6, pp. 3031-3048, Jun 13 2017, doi: 10.1021/acs.jctc.7b00125.

[37] L. Montanucci *et al.*, "DDGun: an untrained predictor of protein stability changes upon amino acid variants," *Nucleic Acids Res,* vol. 50, no. W1, pp. W222-W227, Jul 5 2022, doi: 10.1093/nar/gkac325.

[38] L. Gerasimavicius, B. J. Livesey, and J. A. Marsh, "Correspondence between functional scores from deep mutational scans and predicted effects on protein stability," *Protein Sci,* vol. 32, no. 7, p. e4688, Jul 2023, doi: 10.1002/pro.4688.

[39] J. M. Lee, H. M. Hammaren, M. M. Savitski, and S. H. Baek, "Control of protein stability by post-translational modifications," *Nat Commun,* vol. 14, no. 1, p. 201, Jan 13 2023, doi: 10.1038/s41467-023-35795-8.

[40] G. Birolo, S. Benevenuta, P. Fariselli, E. Capriotti, E. Giorgio, and T. Sanavia, "Protein Stability Perturbation Contributes to the Loss of Function in

Haploinsufficient Genes," *Front Mol Biosci,* vol. 8, p. 620793, 2021, doi: 10.3389/fmolb.2021.620793.
[41] C. Pancotti *et al.*, "Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset," *Brief Bioinform,* vol. 23, no. 2, Mar 10 2022, doi: 10.1093/bib/bbab555.
[42] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature,* vol. 596, no. 7873, pp. 583-589, Aug 2021, doi: 10.1038/s41586-021-03819-2.
[43] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science,* vol. 379, no. 6637, pp. 1123-1130, 2023, doi: doi:10.1126/science.ade2574.
[44] B. Moussad, R. Roche, and D. Bhattacharya, "The transformative power of transformers in protein structure prediction," *Proc Natl Acad Sci U S A,* vol. 120, no. 32, p. e2303499120, Aug 8 2023, doi: 10.1073/pnas.2303499120.
[45] R. E. Andrew W. Senior, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis "Improved protein structure prediction using potentials from deep learning," *Nature,* vol. 577, pp. 706–710, 2020.
[46] J. Jumper *et al.*, "Applying and improving AlphaFold at CASP14," *Proteins,* vol. 89, no. 12, pp. 1711-1721, Dec 2021, doi: 10.1002/prot.26257.
[47] A. Schmidt, S. Roner, K. Mai, H. Klinkhammer, M. Kircher, and K. U. Ludwig, "Predicting the pathogenicity of missense variants using features derived from AlphaFold2," *Bioinformatics,* vol. 39, no. 5, May 4 2023, doi: 10.1093/bioinformatics/btad280.
[48] A. Valanciute, L. Nygaard, H. Zschach, M. Maglegaard Jepsen, K. Lindorff-Larsen, and A. Stein, "Accurate protein stability predictions from homology models," *Comput Struct Biotechnol J,* vol. 21, pp. 66-73, 2023, doi: 10.1016/j.csbj.2022.11.048.
[49] Q. Pan, T. B. Nguyen, D. B. Ascher, and D. E. V. Pires, "Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures," *Brief Bioinform,* vol. 23, no. 2, Mar 10 2022, doi: 10.1093/bib/bbac025.
[50] M. Gasperini, L. Starita, and J. Shendure, "The power of multiplexed functional analysis of genetic variants," *Nat Protoc,* vol. 11, no. 10, pp. 1782-7, Oct 2016, doi: 10.1038/nprot.2016.135.
[51] G. M. Findlay *et al.*, "Accurate classification of BRCA1 variants with saturation genome editing," *Nature,* vol. 562, no. 7726, pp. 217-222, Oct 2018, doi: 10.1038/s41586-018-0461-z.
[52] C. Hu *et al.*, "Functional analysis and clinical classification of 462 germline BRCA2 missense variants affecting the DNA binding domain," *Am J Hum Genet,* vol. 111, no. 3, pp. 584-593, Mar 7 2024, doi: 10.1016/j.ajhg.2024.02.002.
[53] C. Hu *et al.*, "Functional and Clinical Characterization of Variants of Uncertain Significance Identifies a Hotspot for Inactivating Missense Variants in RAD51C,"

Cancer Res, vol. 83, no. 15, pp. 2557-2571, Aug 1 2023, doi: 10.1158/0008-5472.CAN-22-2319.

[54] T. Wiltshire *et al.*, "Functional characterization of 84 PALB2 variants of uncertain significance," *Genet Med,* vol. 22, no. 3, pp. 622-632, Mar 2020, doi: 10.1038/s41436-019-0682-z.

[55] R. Boonen *et al.*, "Functional analysis of genetic variants in the high-risk breast cancer susceptibility gene PALB2," *Nat Commun,* vol. 10, no. 1, p. 5296, Nov 22 2019, doi: 10.1038/s41467-019-13194-2.

[56] M. M. Li *et al.*, "Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists," *J Mol Diagn,* vol. 19, no. 1, pp. 4-23, Jan 2017, doi: 10.1016/j.jmoldx.2016.10.002.

[57] N. M. Ioannidis *et al.*, "REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants," *Am J Hum Genet,* vol. 99, no. 4, pp. 877-885, Oct 6 2016, doi: 10.1016/j.ajhg.2016.08.016.

[58] B. J. Feng, "PERCH: A Unified Framework for Disease Gene Prioritization," *Hum Mutat,* vol. 38, no. 3, pp. 243-251, Mar 2017, doi: 10.1002/humu.23158.

[59] Y. Tian *et al.*, "REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification," *Sci Rep,* vol. 9, no. 1, p. 12752, Sep 4 2019, doi: 10.1038/s41598-019-49224-8.

[60] J. Cheng *et al.*, "Accurate proteome-wide missense variant effect prediction with AlphaMissense," *Science,* vol. 381, no. 6664, p. eadg7492, Sep 22 2023, doi: 10.1126/science.adg7492.

[61] C. Li, D. Zhi, K. Wang, and X. Liu, "MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning," *Genome Med,* vol. 14, no. 1, p. 115, Oct 8 2022, doi: 10.1186/s13073-022-01120-z.

[62] X. Liu, C. Li, C. Mou, Y. Dong, and Y. Tu, "dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs," *Genome Med,* vol. 12, no. 1, p. 103, Dec 2 2020, doi: 10.1186/s13073-020-00803-9.

[63] E. R. White *et al.*, "Peptide library approach to uncover phosphomimetic inhibitors of the BRCA1 C-terminal domain," *ACS Chem Biol,* vol. 10, no. 5, pp. 1198-208, May 15 2015, doi: 10.1021/cb500757u.

[64] R. G. a. J. N. M. G. R. Scott Williams, "Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1," *nature structural biology,* vol. 8, 10, pp. 838-842, 2001.

[65] J. A. Clapperton *et al.*, "Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer," *Nat Struct Mol Biol,* vol. 11, no. 6, pp. 512-8, Jun 2004, doi: 10.1038/nsmb775.

[66] Q. Hu, M. V. Botuyan, D. Zhao, G. Cui, E. Mer, and G. Mer, "Mechanisms of BRCA1-BARD1 nucleosome recognition and ubiquitylation," *Nature,* vol. 596, no. 7872, pp. 438-443, Aug 2021, doi: 10.1038/s41586-021-03716-8.

[67] P. D. J. Haijuan Yang, Julie Miller, Elspeth Kinnucan, Yutong Sun, Nicolas H. Thoma, Ning Zheng, Phang-Lang Chen, Wen-Hwa Lee, Nikola P. Pavletich,

"BRCA2 Function in DNA Binding and Recombination from a BRCA2-DSS1-ssDNA structure," *Science,* vol. 297, pp. 1837-1848, 2002.

[68] Y. Rawal *et al.*, "Structural insights into BCDX2 complex function in homologous recombination," *Nature,* vol. 619, no. 7970, pp. 640-649, Jul 2023, doi: 10.1038/s41586-023-06219-w.

[69] L. A. Greenhough *et al.*, "Structure and function of the RAD51B-RAD51C-RAD51D-XRCC2 tumour suppressor," *Nature,* vol. 619, no. 7970, pp. 650-657, Jul 2023, doi: 10.1038/s41586-023-06179-1.

[70] A. W. Oliver, S. Swift, C. J. Lord, A. Ashworth, and L. H. Pearl, "Structural basis for recruitment of BRCA2 by PALB2," *EMBO Rep,* vol. 10, no. 9, pp. 990-6, Sep 2009, doi: 10.1038/embor.2009.126.

[71] M. Mirdita, K. Schutze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," *Nat Methods,* vol. 19, no. 6, pp. 679-682, Jun 2022, doi: 10.1038/s41592-022-01488-1.

[72] P. J. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics,* vol. 25, no. 11, pp. 1422-3, Jun 1 2009, doi: 10.1093/bioinformatics/btp163.

[73] L. G. Radusky and L. Serrano, "pyFoldX: enabling biomolecular analysis and engineering along structural ensembles," *Bioinformatics,* vol. 38, no. 8, pp. 2353-2355, Apr 12 2022, doi: 10.1093/bioinformatics/btac072.

[74] H. Park *et al.*, "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules," *J Chem Theory Comput,* vol. 12, no. 12, pp. 6201-6212, Dec 13 2016, doi: 10.1021/acs.jctc.6b00819.

[75] G. V. Fabian Pedregosa, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel**,** Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[76] I. Kufareva and R. Abagyan, "Methods of protein structure comparison," *Methods Mol Biol,* vol. 857, pp. 231-57, 2012, doi: 10.1007/978-1-61779-588-6_10.

[77] D. G. Grimm *et al.*, "The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity," *Hum Mutat,* vol. 36, no. 5, pp. 513-23, May 2015, doi: 10.1002/humu.22768.

[78] R. Gnanaolivu and S. N. Hart, "Using AI-predicted protein structures as a reference to predict loss-of-function activity in tumor suppressor breast cancer genes," *Comput Struct Biotechnol J,* vol. 23, pp. 3472-3480, Dec 2024, doi: 10.1016/j.csbj.2024.10.008.

[79] K. Esoh and A. Wonkam, "Evolutionary history of sickle-cell mutation: implications for global genetic medicine," *Hum Mol Genet,* vol. 30, no. R1, pp. R119-R128, Apr 26 2021, doi: 10.1093/hmg/ddab004.

[80] X. Zhang *et al.*, "Review of published 467 achondroplasia patients: clinical and mutational spectrum," *Orphanet J Rare Dis,* vol. 19, no. 1, p. 29, Jan 27 2024, doi: 10.1186/s13023-024-03031-1.

[81] B. H. Shirts, A. Jacobson, G. P. Jarvik, and B. L. Browning, "Large numbers of individuals are required to classify and define risk for rare variants in known

cancer risk genes," *Genet Med,* vol. 16, no. 7, pp. 529-34, Jul 2014, doi: 10.1038/gim.2013.187.

[82] J. Frazer *et al.*, "Disease variant prediction with deep generative models of evolutionary data," *Nature,* vol. 599, no. 7883, pp. 91-95, Nov 2021, doi: 10.1038/s41586-021-04043-8.

[83] H. A. Shihab *et al.*, "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models," *Hum Mutat,* vol. 34, no. 1, pp. 57-65, Jan 2013, doi: 10.1002/humu.22225.

[84] E. H. Wilcox *et al.*, "Evaluating the impact of in silico predictors on clinical variant classification," *Genet Med,* vol. 24, no. 4, pp. 924-930, Apr 2022, doi: 10.1016/j.gim.2021.11.018.

[85] T. E. K. Sean D Mooney, "The functional importance of disease-associated mutation," *BMC Bioinformatics,* p. 3:24, 2002, doi: 10.1186/1471-2105-3-24.

[86] N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos, "Genome-wide prediction of disease variant effects with a deep protein language model," *Nat Genet,* vol. 55, no. 9, pp. 1512-1522, Sep 2023, doi: 10.1038/s41588-023-01465-0.

[87] J. B. Zhou, Y. Xiong, K. An, Z. Q. Ye, and Y. D. Wu, "IDRMutPred: predicting disease-associated germline nonsynonymous single nucleotide variants (nsSNVs) in intrinsically disordered regions," *Bioinformatics,* vol. 36, no. 20, pp. 4977-4983, Dec 22 2020, doi: 10.1093/bioinformatics/btaa618.

[88] M. Feng *et al.*, "Decoding Missense Variants by Incorporating Phase Separation via Machine Learning," *Nat Commun,* vol. 15, no. 1, p. 8279, Sep 27 2024, doi: 10.1038/s41467-024-52580-3.

[89] T. R. Alderson, I. Pritisanac, D. Kolaric, A. M. Moses, and J. D. Forman-Kay, "Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2," *Proc Natl Acad Sci U S A,* vol. 120, no. 44, p. e2304302120, Oct 31 2023, doi: 10.1073/pnas.2304302120.

[90] J. M. Lotthammer, G. M. Ginell, D. Griffith, R. J. Emenecker, and A. S. Holehouse, "Direct prediction of intrinsically disordered protein conformational properties from sequence," *Nat Methods,* vol. 21, no. 3, pp. 465-476, Mar 2024, doi: 10.1038/s41592-023-02159-5.

[91] R. Trivedi and H. A. Nagarajaram, "Intrinsically Disordered Proteins: An Overview," *Int J Mol Sci,* vol. 23, no. 22, Nov 14 2022, doi: 10.3390/ijms232214050.

[92] B. Han, C. Ren, W. Wang, J. Li, and X. Gong, "Computational Prediction of Protein Intrinsically Disordered Region Related Interactions and Functions," *Genes (Basel),* vol. 14, no. 2, Feb 8 2023, doi: 10.3390/genes14020432.

[93] B. Zhao, S. Ghadermarzi, and L. Kurgan, "Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins," *Comput Struct Biotechnol J,* vol. 21, pp. 3248-3258, 2023, doi: 10.1016/j.csbj.2023.06.001.

[94] D. Piovesan, A. M. Monzon, and S. C. E. Tosatto, "Intrinsic protein disorder and conditional folding in AlphaFoldDB," *Protein Sci,* vol. 31, no. 11, p. e4466, Nov 2022, doi: 10.1002/pro.4466.

[95] G. van Mierlo, J. R. G. Jansen, J. Wang, I. Poser, S. J. van Heeringen, and M. Vermeulen, "Predicting protein condensate formation using machine learning," *Cell Rep,* vol. 34, no. 5, p. 108705, Feb 2 2021, doi: 10.1016/j.celrep.2021.108705.
[96] J. Sun *et al.*, "Precise prediction of phase-separation key residues by machine learning," *Nat Commun,* vol. 15, no. 1, p. 2662, Mar 26 2024, doi: 10.1038/s41467-024-46901-9.
[97] M. Monti *et al.*, "catGRANULE 2.0: accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution," *Genome Biol,* vol. 26, no. 1, p. 33, Feb 20 2025, doi: 10.1186/s13059-025-03497-7.
[98] A. Elnaggar *et al.*, "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Trans Pattern Anal Mach Intell,* vol. 44, no. 10, pp. 7112-7127, Oct 2022, doi: 10.1109/TPAMI.2021.3095381.
[99] D. Piovesan *et al.*, "MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins," *Nucleic Acids Res,* vol. 53, no. D1, pp. D495-D503, Jan 6 2025, doi: 10.1093/nar/gkae969.
[100] T. S. Akiba, Shotaro; Yanase, Toshihiko; Ohta, Takeru; Koyama, Masanori, "Optuna: A Next-generation Hyperparameter Optimization Framework," *arXiv,* 2019.
[101] G. Perez *et al.*, "The UCSC Genome Browser database: 2025 update," *Nucleic Acids Res,* vol. 53, no. D1, pp. D1243-D1249, Jan 6 2025, doi: 10.1093/nar/gkae974.
[102] M. Munz *et al.*, "CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting," *Genome Med,* vol. 7, no. 1, p. 76, Jul 28 2015, doi: 10.1186/s13073-015-0195-6.
[103] S. L. Lundberg, Su-In, "A Unified Approach to Interpreting Model Predictions," *arXiv,* 2017.

Appendix

Supplementary material for chapter 2

Supplementary Tables

Table S4: Table denoting the total number of classified mutations used from the ClinVar database and MAVE functional assays for evaluation.

GENES	Functional			ClinVar	
	Neutral	Deleterious	Intermediate	Neutral	Deleterious
BRCA1	1476	441	169	720	324
BRCA2	313	137	12	336	935
PALB2	84	7	0	31	271
RAD51C	137	30	7	3	57

Table S5: List of PDB ID from Genes used in this study.

GENE	PDB ID	Atom count	Resolution	Residue count	Protein chains
RAD51C	8FAZ	7331	2.30 Å	926	4
RAD51C	8OUZ	7449	2.20 Å	924	4
BRCA1	7LYB	14699	3.28 Å	1389	7
BRCA1	4OFB	1791	3.05 Å	223	2
BRCA1	1T15	1906	1.85 Å	219	2
BRCA1	1JNX	1660	2.50 Å	207	1
BRCA2	1IYJ	10092	3.40 Å	1272	2
BRCA2	1MJE	5198	3.50 Å	648	2
PALB2	3EU7	2585	2.20 Å	327	2
PALB2	2W18	2543	1.90 Å	306	1

Table S6: Parameters used in ColabFold and ESMFold

ColabFold parameters	
template mode	pdb100
model_type (complex)	alphafold2_multimer_v3
model_type (monomer)	alphafold2_ptm
num_cycles	3
recycle_early_stop_tolerance	0.5
relax_max_iterations	200
pairing_strategy	greedy

ESMFold Parameter	
pretrained model	esm2_t33_650M_UR50D
repr_layers	33
chunk size	128
num_recycles	4

Supplementary Figures

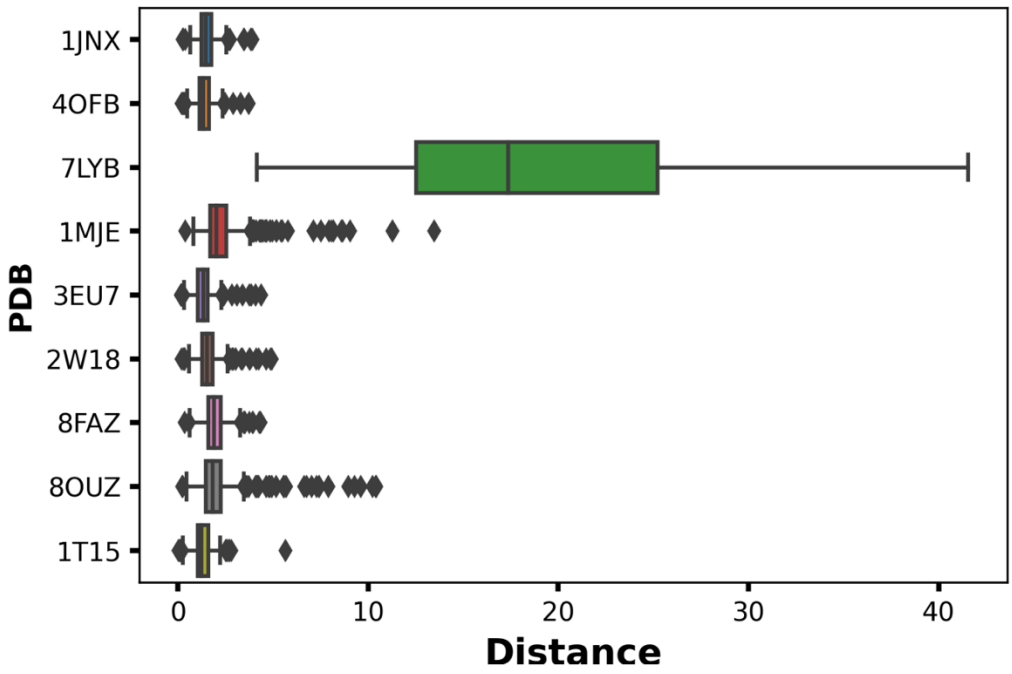

Figure S14*: Distribution of the per residue distance in Å between AlphaFold2 predicted structure and experimentally-derived structure found in the PDB. AlphaFold2 prediction of 7LYB (BRCA1 Ring domain) was an outlier with a mean residue distance of 19Å.*

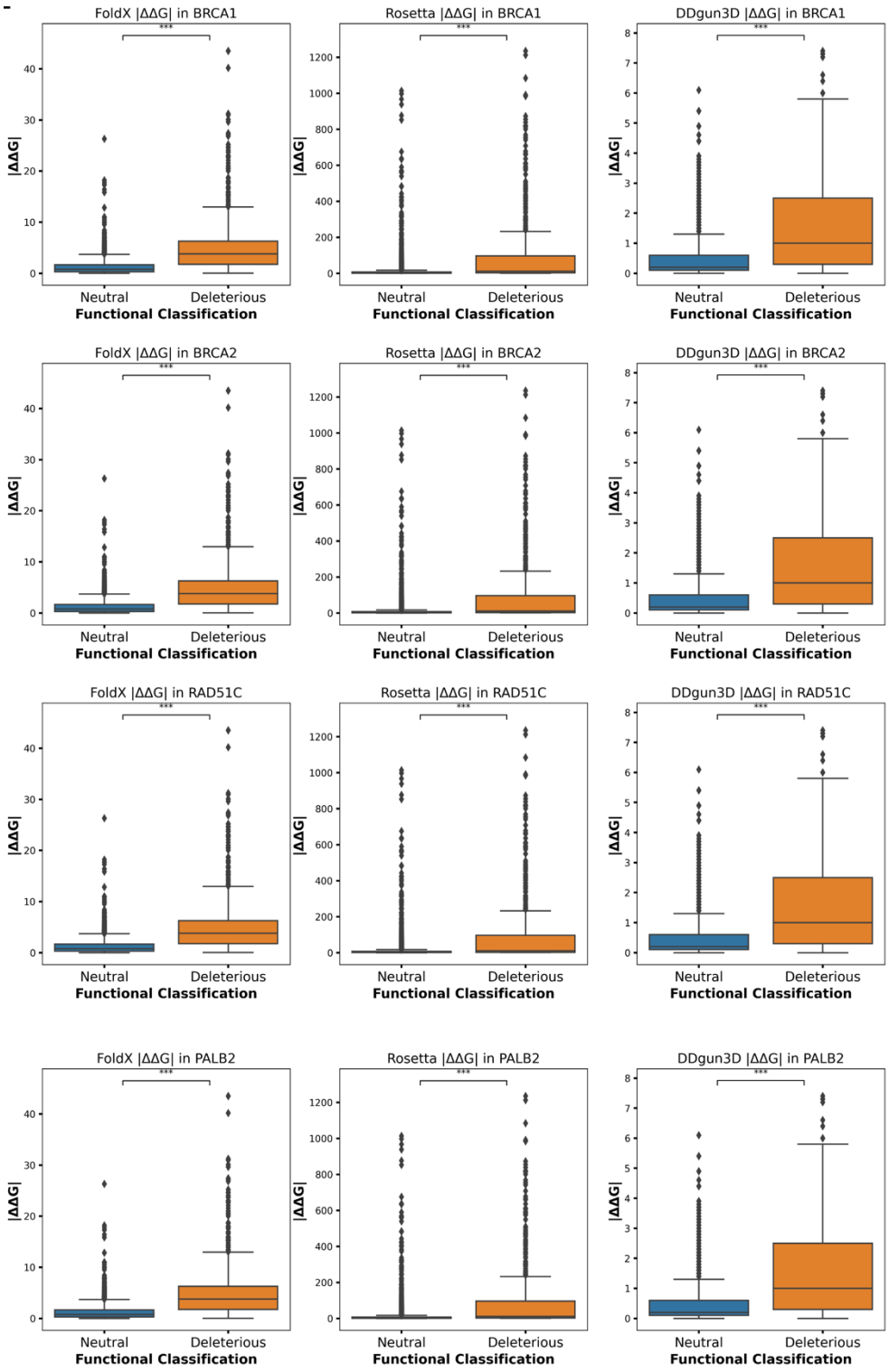

***Figure S15**: Distribution of predicted |ΔΔG| from experimentally-derived structure analyzed with FoldX, Rosetta and DDGun3D stratified by functional classification (Deleterious vs Neutral) in genes BRCA1, BRCA2, PALB2 and RAD51C, with the association between the two groups denoted by the Mann-Whitney U test.*

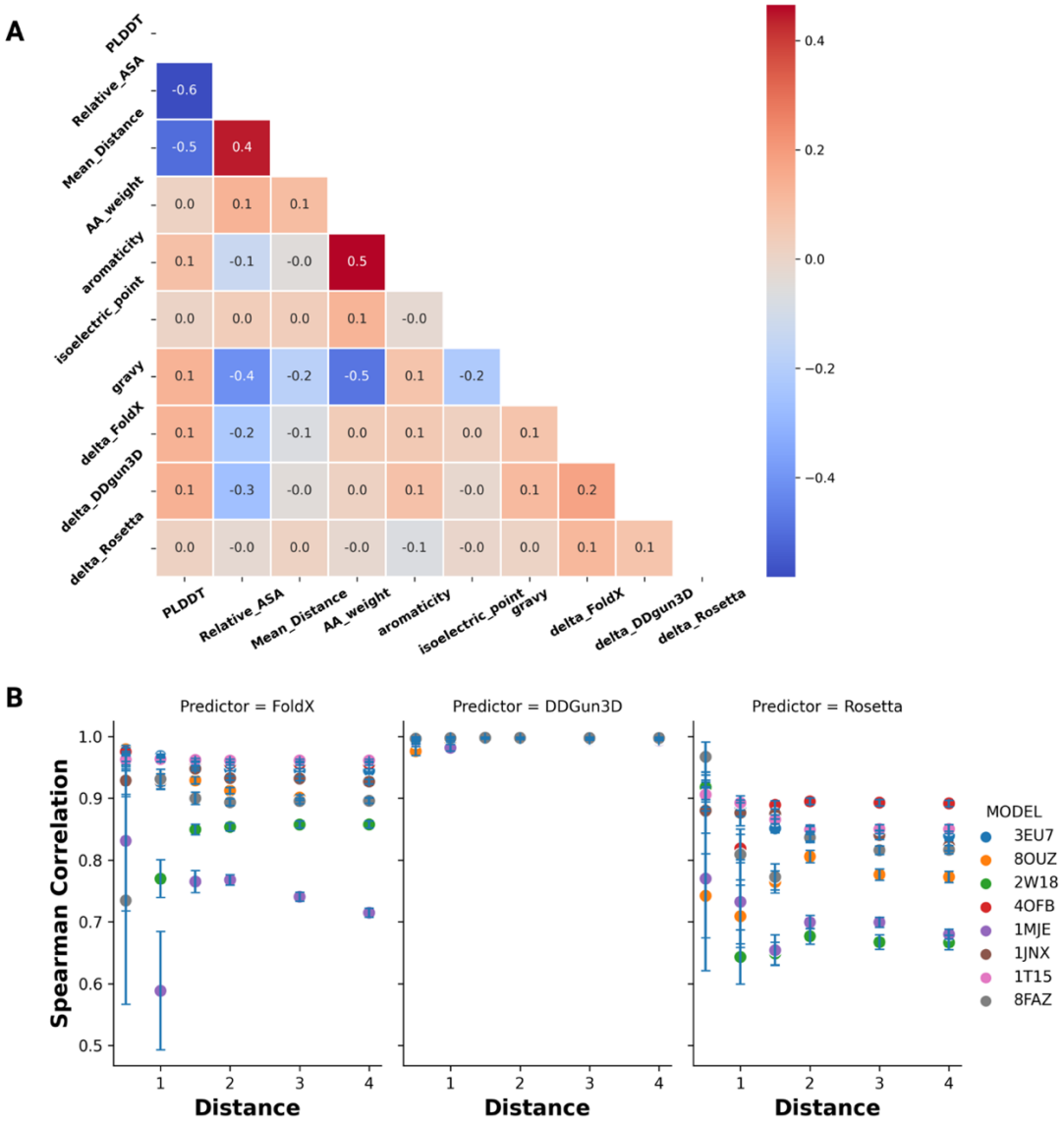

Figure S16: Monotonic association of difference between the |ΔΔG| from AF2 structures vs experimentally-derived structure as the wild-type template analyzed with FoldX, Rosetta and DDGun3D with the features extracted from AlpahFold2 structures and per residue distance between the superimposed AlphaFold2 structure onto the experimentally-derived structure*. A. Spearman rank correlation coefficient denoting the monotonic association of the features derived from AlphaFold2 structures and deltas of the predicted ΔΔG derived from AlphaFold2 structures and experimentally-derived structures analyzed with FoldX, Rosetta and DDGun3D. B Scatterplot denoting the spearman rank correlation of the difference between the predicted ΔΔG from AlphaFold2 structures vs experimentally-derived structure as the wild-type template analyzed by FoldX, DDGun3D and Rosetta stratified by the per residue distance between the superimposed AlphaFold2 structure onto the experimentally-derived structure. The distance of the x-axis is limited to 4Å*

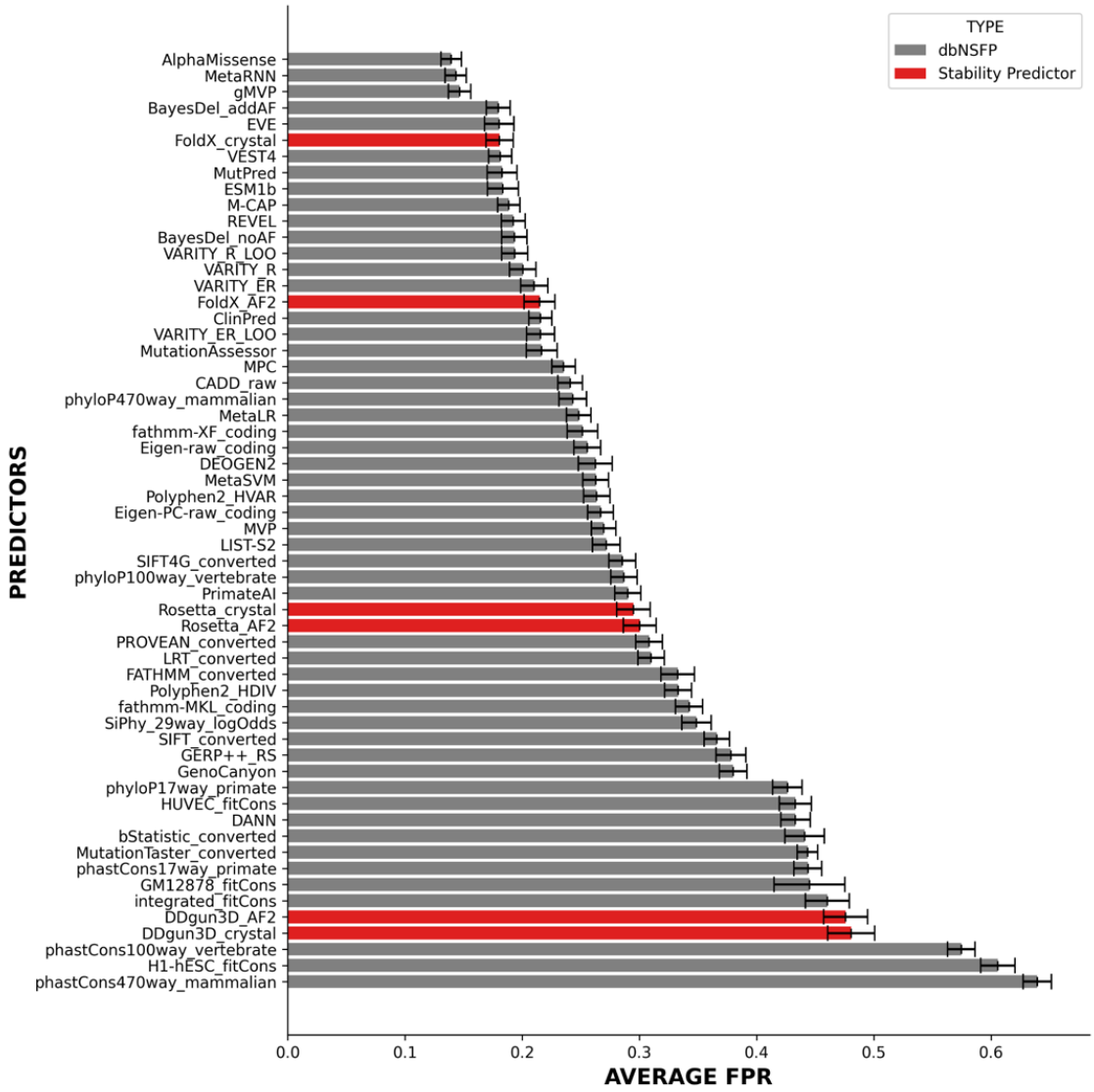

Figure S17*: Average false positive rates, along with 95% confidence intervals describing the for all dbNSFP Insilco missense predictors and stability predictors across the genes BRCA1, BRCA2, PALB2 and RAD512C. The analysis highlights the strong performance of AlphaMissense in predicting loss-of-function, and the performance of stability predictors (in red) to predict loss-of-function in these genes.*

Table S7: In silico predictor performance metrics (AUC, AUC-PR, F1 score) for the top 10 predictors in the dbNSFP database and stability predictor (FoldX, Rosetta and DDGun3D) in predicting LOF activity in BRCA1-BRCT domain

PREDICTOR	AUC (95% CI)	AUC-PR (95% CI)	F1 SCORE (95% CI)	SOURCE

AlphaMissense	0.8991 (0.8971 - 0.9012)	0.8967 (0.8936 - 0.8999)	0.8512 (0.8489 - 0.8534)	dbNSFP
gMVP	0.8944 (0.8924 - 0.8963)	0.9082 (0.9061 - 0.9102)	0.8308 (0.8285 - 0.833)	dbNSFP
VEST4	0.8899 (0.8878 - 0.8921)	0.8892 (0.8866 - 0.8919)	0.8243 (0.8217 - 0.8268)	dbNSFP
VARITY_R	0.8715 (0.8693 - 0.8738)	0.8658 (0.8631 - 0.8684)	0.8016 (0.7989 - 0.8042)	dbNSFP
BayesDel_addAF	0.8698 (0.8676 - 0.872)	0.8757 (0.8731 - 0.8784)	0.8171 (0.8148 - 0.8194)	dbNSFP
REVEL	0.8698 (0.8674 - 0.8722)	0.8662 (0.8632 - 0.8693)	0.8134 (0.8108 - 0.8159)	dbNSFP
ESM1b	0.8677 (0.8655 - 0.8699)	0.8863 (0.8839 - 0.8887)	0.8042 (0.8018 - 0.8066)	dbNSFP
BayesDel_noAF	0.8668 (0.8645 - 0.869)	0.8725 (0.8697 - 0.8752)	0.813 (0.8106 - 0.8153)	dbNSFP
FoldX_AF2_4OFB	0.8615 (0.8582 - 0.8638)	0.8703 (0.8676 - 0.8731)	0.7897 (0.7871 - 0.7924)	Stability Predictor
MetaRNN	0.8612 (0.8587 - 0.8637)	0.8454 (0.8419 - 0.8489)	0.8071 (0.8044 - 0.8098)	dbNSFP
M-CAP	0.8564 (0.8539 - 0.8589)	0.8496 (0.8465 - 0.8528)	0.7852 (0.7824 - 0.788)	dbNSFP
FoldX_crystal_4OFB	0.8478 (0.8444 - 0.8502)	0.8658 (0.8631 - 0.8685)	0.7964 (0.794 - 0.7989)	Stability Predictor
FoldX_AF2_1JNX	0.8365 (0.8332 - 0.8398)	0.8566 (0.8539 - 0.8592)	0.7624 (0.7599 - 0.7649)	Stability Predictor
FoldX_crystal_1T15	0.8351 (0.8325 - 0.8376)	0.8502 (0.8471 - 0.8533)	0.7657 (0.7631 - 0.7684)	Stability Predictor
FoldX_crystal_1JNX	0.7925 (0.7898 - 0.7951)	0.8386 (0.836 - 0.8412)	0.7373 (0.7348 - 0.7399)	Stability Predictor

Rosetta_AF2_1T15	0.7154 (0.712 - 0.7187)	0.7472 (0.7433 - 0.7512)	0.7053 (0.7023 - 0.7083)	Stability Predictor
Rosetta_crystal_4OFB	0.7005 (0.6972 - 0.7039)	0.6891 (0.6844 - 0.6939)	0.6283 (0.6253 - 0.6313)	Stability Predictor
Rosetta_AF2_1JNX	0.6997 (0.696 - 0.7033)	0.6977 (0.6929 - 0.7024)	0.6384 (0.6349 - 0.642)	Stability Predictor
Rosetta_AF2_4OFB	0.6949 (0.6914 - 0.6984)	0.683 (0.6782 - 0.6878)	0.6279 (0.6246 - 0.6312)	Stability Predictor
Rosetta_crystal_1T15	0.6906 (0.6872 - 0.6939)	0.7315 (0.7276 - 0.7353)	0.687 (0.684 - 0.69)	Stability Predictor
Rosetta_crystal_1JNX	0.6839 (0.6803 - 0.6874)	0.6893 (0.6846 - 0.6941)	0.6348 (0.6317 - 0.6378)	Stability Predictor
DDgun3D_crystal_4OFB	0.3293 (0.3258 - 0.3329)	0.4526 (0.4488 - 0.4564)	0.3818 (0.3785 - 0.3852)	Stability Predictor
DDgun3D_crystal_1T15	0.3287 (0.3255 - 0.3319)	0.4448 (0.4409 - 0.4487)	0.3781 (0.3748 - 0.3814)	Stability Predictor
DDgun3D_AF2_4OFB	0.3287 (0.3251 - 0.3322)	0.4508 (0.447 - 0.4546)	0.3798 (0.3764 - 0.3831)	Stability Predictor
DDgun3D_AF2_1T15	0.3286 (0.3255 - 0.3318)	0.4463 (0.4424 - 0.4501)	0.3772 (0.374 - 0.3804)	Stability Predictor
DDgun3D_AF2_1JNX	0.328 (0.3244 - 0.3315)	0.4531 (0.4492 - 0.4569)	0.378 (0.3747 - 0.3814)	Stability Predictor
DDgun3D_crystal_1JNX	0.3272 (0.3236 - 0.3308)	0.4469 (0.4431 - 0.4507)	0.3821 (0.3787 - 0.3856)	Stability Predictor

Table S8: In silico predictor performance metrics (AUC, AUC-PR, F1 score) for the top 10 predictors in the dbNSFP database and stability predictor (FoldX, Rosetta and DDGun3D) in predicting LOF activity in BRCA1-RING domain

PREDICTOR	**AUC (95% CI)**	**AUC-PR (95% CI)**	**F1 SCORE (95% CI)**	**SOURCE**

AlphaMissense	0.8969 (0.8946 - 0.9003)	0.9051 (0.9014 - 0.9088)	0.8336 (0.8297 - 0.8375)	dbNSFP
BayesDel_addAF	0.8909 (0.8873 - 0.8944)	0.8998 (0.8961 - 0.9034)	0.8232 (0.819 - 0.8274)	dbNSFP
gMVP	0.8903 (0.8869 - 0.8937)	0.9166 (0.9137 - 0.9194)	0.8349 (0.8311 - 0.8387)	dbNSFP
BayesDel_noAF	0.89 (0.8864 - 0.8935)	0.9013 (0.8978 - 0.9049)	0.8199 (0.8157 - 0.824)	dbNSFP
PROVEAN	0.8849 (0.8813 - 0.8885)	0.8868 (0.8826 - 0.8909)	0.8177 (0.8135 - 0.8219)	dbNSFP
VEST4	0.8769 (0.8734 - 0.8804)	0.896 (0.8925 - 0.8995)	0.8317 (0.8278 - 0.8355)	dbNSFP
REVEL	0.8694 (0.8655 - 0.8733)	0.8862 (0.8824 - 0.8901)	0.8012 (0.7967 - 0.8057)	dbNSFP
MetaRNN	0.8661 (0.8623 - 0.8699)	0.8819 (0.8783 - 0.8856)	0.786 (0.7817 - 0.7904)	dbNSFP
VARITY_R	0.8647 (0.8611 - 0.8684)	0.8885 (0.885 - 0.892)	0.7886 (0.7843 - 0.7928)	dbNSFP
MVP	0.8612 (0.8575 - 0.8649)	0.8726 (0.8691 - 0.8762)	0.7876 (0.7832 - 0.792)	dbNSFP
FoldX_crystal_7LYB	0.835 (0.831 - 0.84)	0.867 (0.8633 - 0.8707)	0.7718 (0.7674 - 0.7762)	Stability Predictor
FoldX_AF2_7LYB	0.7416 (0.7357 - 0.7485)	0.7905 (0.7849 - 0.7961)	0.7183 (0.7133 - 0.7233)	Stability Predictor
Rosetta_AF2_7LYB	0.6424 (0.6367 - 0.6481)	0.6473 (0.6394 - 0.6552)	0.6227 (0.6179 - 0.6275)	Stability Predictor
Rosetta_crystal_7LYB	0.5272 (0.5207 - 0.5337)	0.5828 (0.5747 - 0.5909)	0.5502 (0.5447 - 0.5556)	Stability Predictor
DDgun3D_crystal_7LYB	0.1869 (0.1823 - 0.1915)	0.3681 (0.3634 - 0.3728)	0.2887 (0.2838 - 0.2936)	Stability Predictor

DDgun3D_AF2_7LYB	0.1811 (0.1766 - 0.1855)	0.3655 (0.3609 - 0.3701)	0.2849 (0.28 - 0.2897)	Stability Predictor

Table S9:In silico predictor performance metrics (AUC, AUC-PR, F1 score) for the top 10 predictors in the dbNSFP database and stability predictor (FoldX, Rosetta and DDGun3D) in predicting LOF activity in BRCA2 DBD domain

PREDICTOR	**AUC (95% CI)**	**AUC-PR (95% CI)**	**F1 SCORE (95% CI)**	**SOURCE**
AlphaMissense	0.9193 (0.9169 - 0.9216)	0.9122 (0.9089 - 0.9156)	0.8751 (0.8722 - 0.878)	dbNSFP
MetaRNN	0.9103 (0.9075 - 0.913)	0.9105 (0.9075 - 0.9136)	0.8402 (0.8369 - 0.8435)	dbNSFP
BayesDel_addAF	0.9031 (0.9002 - 0.906)	0.8964 (0.8926 - 0.9002)	0.8342 (0.8306 - 0.8378)	dbNSFP
VEST4	0.9008 (0.8979 - 0.9036)	0.9035 (0.9003 - 0.9066)	0.8408 (0.8376 - 0.8441)	dbNSFP
gMVP	0.9007 (0.8979 - 0.9035)	0.8901 (0.8856 - 0.8946)	0.8494 (0.8462 - 0.8525)	dbNSFP
ClinPred	0.8906 (0.8877 - 0.8936)	0.8779 (0.8739 - 0.882)	0.8148 (0.8113 - 0.8183)	dbNSFP
BayesDel_noAF	0.8867 (0.8836 - 0.8897)	0.8812 (0.8774 - 0.8851)	0.8253 (0.8218 - 0.8288)	dbNSFP
MPC	0.8838 (0.8806 - 0.8871)	0.868 (0.8635 - 0.8725)	0.8236 (0.8202 - 0.8271)	dbNSFP
MVP	0.8645 (0.8609 - 0.8681)	0.853 (0.8484 - 0.8576)	0.8061 (0.8021 - 0.81)	dbNSFP
CADD_raw	0.8633 (0.8596 - 0.867)	0.8278 (0.8218 - 0.8338)	0.8103 (0.8067 - 0.814)	dbNSFP
FoldX_AF2_1MJE	0.8364 (0.8322 - 0.8407)	0.8819 (0.8761 - 0.8878)	0.7797 (0.7743 - 0.7851)	Stability Predictor

FoldX_crystal_1MJE	0.8299 (0.8213 - 0.8320)	0.8912 (0.8861 - 0.8962)	0.7871 (0.7822 - 0.7919)	Stability Predictor
Rosetta_AF2_1MJE	0.6813 (0.675 - 0.6866)	0.7966 (0.7905 - 0.8028)	0.6622 (0.6577 - 0.6668)	Stability Predictor
Rosetta_crystal_1MJE	0.6649 (0.6587 - 0.6722)	0.795 (0.7889 - 0.8011)	0.6576 (0.6524 - 0.6627)	Stability Predictor
DDgun3D_AF2_1MJE	0.26 (0.2553 - 0.2647)	0.5805 (0.5745 - 0.5865)	0.3326 (0.3279 - 0.3374)	Stability Predictor
DDgun3D_crystal_1MJE	0.2553 (0.2506 - 0.26)	0.5795 (0.5735 - 0.5854)	0.3216 (0.3169 - 0.3262)	Stability Predictor

Table S10: In silico predictor performance metrics (AUC, AUC-PR, F1 score) for the top 10 predictors in the dbNSFP database and stability predictor (FoldX, Rosetta and DDGun3D) in predicting LOF activity in PALB2

PREDICTOR	**AUC (95% CI)**	**AUC-PR (95% CI)**	**F1 SCORE (95% CI)**	**SOURCE**
MetaRNN	0.9015 (0.8925 - 0.9204)	0.9263 (0.9182 - 0.9344)	0.8753 (0.8651 - 0.8855)	dbNSFP
MutationTaster	0.8638 (0.8525 - 0.8751)	0.8693 (0.8556 - 0.8829)	0.8408 (0.8299 - 0.8516)	dbNSFP
M-CAP	0.8582 (0.847 - 0.8694)	0.9021 (0.8935 - 0.9106)	0.8334 (0.8222 - 0.8446)	dbNSFP
MPC	0.8275 (0.8102 - 0.8448)	0.7816 (0.759 - 0.8042)	0.8221 (0.807 - 0.8371)	dbNSFP
MutPred	0.8251 (0.8102 - 0.84)	0.895 (0.8848 - 0.9051)	0.8144 (0.8014 - 0.8275)	dbNSFP
EVE	0.8241 (0.801 - 0.8472)	0.7639 (0.7336 - 0.7942)	0.8362 (0.8161 - 0.8562)	dbNSFP
BayesDel_addAF	0.8102 (0.7955 - 0.8249)	0.8243 (0.8102 - 0.8385)	0.7876 (0.7733 - 0.8019)	dbNSFP

phyloP470way_mammalian	0.8073 (0.7914 - 0.8232)	0.7801 (0.7607 - 0.7995)	0.8071 (0.7928 - 0.8214)	dbNSFP
PrimateAI	0.8026 (0.785 - 0.8202)	0.695 (0.6684 - 0.7216)	0.8047 (0.7895 - 0.8199)	dbNSFP
ClinPred	0.8005 (0.7834 - 0.8175)	0.7785 (0.7602 - 0.7967)	0.8107 (0.7963 - 0.8252)	dbNSFP
FoldX_crystal_3EU7	0.7443 (0.7173 - 0.7713)	0.783 (0.7578 - 0.8082)	0.801 (0.7802 - 0.8219)	Stability Predictor
FoldX_AF2_3EU7	0.7215 (0.7108 - 0.7471)	0.7571 (0.7301 - 0.7841)	0.8021 (0.7809 - 0.8232)	Stability Predictor
FoldX_AF2_2W18	0.7012 (0.6738 - 0.7287)	0.6864 (0.6583 - 0.7146)	0.7623 (0.7409 - 0.7837)	Stability Predictor
FoldX_crystal_2W18	0.6886 (0.6612 - 0.716)	0.6622 (0.6321 - 0.6924)	0.7397 (0.7184 - 0.7609)	Stability Predictor
Rosetta_crystal_2W18	0.4055 (0.3796 - 0.435)	0.4183 (0.3964 - 0.4402)	0.537 (0.5122 - 0.5618)	Stability Predictor
Rosetta_crystal_3EU7	0.3767 (0.3512 - 0.4021)	0.4082 (0.3866 - 0.4299)	0.502 (0.4773 - 0.5267)	Stability Predictor
Rosetta_AF2_3EU7	0.372 (0.3462 - 0.3979)	0.4232 (0.401 - 0.4453)	0.4901 (0.4654 - 0.5148)	Stability Predictor
Rosetta_AF2_2W18	0.349 (0.3296 - 0.3883)	0.4183 (0.3961 - 0.4406)	0.4786 (0.4533 - 0.5038)	Stability Predictor
DDgun3D_AF2_3EU7	0.1571 (0.1352 - 0.1799)	0.3334 (0.3142 - 0.3526)	0.3019 (0.2761 - 0.3277)	Stability Predictor
DDgun3D_AF2_2W18	0.1596 (0.136 - 0.1812)	0.3334 (0.3142 - 0.3526)	0.3019 (0.2761 - 0.3277)	Stability Predictor
DDgun3D_crystal_2W18	0.1278 (0.1111 - 0.1446)	0.3151 (0.2972 - 0.3331)	0.2705 (0.2465 - 0.2944)	Stability Predictor
DDgun3D_crystal_3EU7	0.1236 (0.1065 - 0.1407)	0.3746 (0.355 - 0.3943)	0.2511 (0.225 - 0.2772)	Stability Predictor

Table S11:In silico predictor performance metrics (AUC, AUC-PR, F1 score) for the top 10 predictors in the dbNSFP database and stability predictor (FoldX, Rosetta and DDGun3D) in predicting LOF activity in RAD51C

PREDICTOR	**AUC (95% CI)**	**AUC-PR (95% CI)**	**F1 SCORE (95% CI)**	**SOURCE**
gMVP	0.9595 (0.9558 - 0.9633)	0.9511 (0.9458 - 0.9564)	0.9342 (0.9294 - 0.939)	dbNSFP
VARITY_ER	0.9565 (0.9531 - 0.9599)	0.9547 (0.9507 - 0.9588)	0.9088 (0.9037 - 0.9139)	dbNSFP
VARITY_ER_LOO	0.9531 (0.9496 - 0.9565)	0.9523 (0.9483 - 0.9564)	0.9042 (0.8992 - 0.9092)	dbNSFP
VARITY_R	0.9482 (0.9444 - 0.9519)	0.9437 (0.9385 - 0.9488)	0.9 (0.895 - 0.905)	dbNSFP
VARITY_R_LOO	0.9476 (0.9439 - 0.9514)	0.9436 (0.9385 - 0.9487)	0.9014 (0.8964 - 0.9063)	dbNSFP
REVEL	0.9421 (0.9375 - 0.9467)	0.9296 (0.9215 - 0.9376)	0.9077 (0.9023 - 0.913)	dbNSFP
MetaRNN	0.9393 (0.9353 - 0.9434)	0.9465 (0.9424 - 0.9506)	0.8912 (0.8859 - 0.8965)	dbNSFP
EVE	0.9387 (0.9338 - 0.9437)	0.9348 (0.9286 - 0.941)	0.9093 (0.9033 - 0.9154)	dbNSFP
AlphaMissense	0.9364 (0.9294 - 0.9405)	0.9178 (0.9096 - 0.926)	0.8986 (0.8935 - 0.9037)	dbNSFP
MutPred	0.918 (0.9127 - 0.9232)	0.9329 (0.9271 - 0.9386)	0.8648 (0.8587 - 0.8709)	dbNSFP
Rosetta_crystal_8OUZ	0.7749 (0.7657 - 0.7841)	0.7959 (0.7857 - 0.8061)	0.7253 (0.7168 - 0.7337)	Stability Predictor
Rosetta_AF2_8FAZ	0.7827 (0.7738 - 0.7915)	0.7915 (0.7814 - 0.8016)	0.7301 (0.7216 - 0.7386)	Stability Predictor

Rosetta_crystal_8FAZ	0.7602 (0.7513 - 0.769)	0.7747 (0.7646 - 0.7849)	0.7202 (0.7122 - 0.7282)	Stability Predictor
FoldX_crystal_8FAZ	0.7598 (0.7483 - 0.7713)	0.8369 (0.8294 - 0.8443)	0.7795 (0.7726 - 0.7865)	Stability Predictor
Rosetta_AF2_8OUZ	0.7393 (0.7301 - 0.7484)	0.7718 (0.7616 - 0.782)	0.7197 (0.7118 - 0.7276)	Stability Predictor
FoldX_AF2_8FAZ	0.7289 (0.7199 - 0.738)	0.7785 (0.7685 - 0.7886)	0.7168 (0.709 - 0.7245)	Stability Predictor
FoldX_crystal_8OUZ	0.7131 (0.7028 - 0.7235)	0.7853 (0.7758 - 0.7948)	0.7566 (0.7483 - 0.7648)	Stability Predictor
FoldX_AF2_8OUZ	0.7008 (0.6908 - 0.7109)	0.7704 (0.7596 - 0.7811)	0.7009 (0.6917 - 0.7101)	Stability Predictor
DDgun3D_crystal_8FAZ	0.4651 (0.4543 - 0.4759)	0.571 (0.5593 - 0.5827)	0.5206 (0.5109 - 0.5303)	Stability Predictor
DDgun3D_AF2_8FAZ	0.4646 (0.4537 - 0.4754)	0.578 (0.5662 - 0.5898)	0.5153 (0.5059 - 0.5247)	Stability Predictor
DDgun3D_crystal_8OUZ	0.452 (0.4411 - 0.4628)	0.5509 (0.539 - 0.5628)	0.5092 (0.4991 - 0.5192)	Stability Predictor
DDgun3D_AF2_8OUZ	0.4481 (0.4372 - 0.459)	0.5656 (0.554 - 0.5772)	0.5118 (0.5017 - 0.522)	Stability Predictor

Supplementary material for chapter 3

Supplementary Methods

Gene-split training and testing

To evaluate the potential for enhanced predictive performance, we retrained the XGBoost model by incorporating AlphaMissense, ESM1b, and EVE predictions as additional features in separate experiments. Each external predictor was integrated individually with our baseline feature set to assess its specific contribution to variant classification accuracy. To prevent data leakage from variants within the same gene, the dataset was partitioned using a gene-based splitting strategy implemented through the Optuna optimization framework, ensuring that all variants from a given gene were assigned exclusively to either the training or test set, but never both. This approach prevents the model from learning gene-specific patterns during training that could artificially inflate performance on test variants from the same genes. The gene-based partitioning resulted in 1,852 variants for training and 252 variants for testing. HPO employed the Optuna framework across 150 optimization trials, with 10-fold stratified cross-validation performed exclusively on the training set while maintaining gene-based separation throughout the cross-validation process to prevent information leakage during model selection. The objective function maximized PR-AUC during cross-validation. Following optimization, final models were evaluated on the held-out test set, with performance confidence intervals calculated through bootstrap resampling over 1,000 iterations to ensure robust statistical assessment.

Supplementary Results

Gene-split training and testing evaluation

Based on the variant distribution and bias in the classification of variants by genes, there was imbalance in the proportion of deleterious variants in the training compared to the testing. The train class distribution was 1612 neutral variants and 240 deleterious variants from 232 genes, while the test class distribution had neutral 176 variants and 76 deleterious variants from 58 genes. Following retraining with enhanced models incorporating AlphaMissense, ESM1b, and EVE, we observed improved PR-AUC performance across all three scenarios when evaluated against the test set compared to the standalone model. The AlphaMissense enhanced model demonstrated modest improvement, with PR-AUC increasing from 0.953 to 0.958 relative to the standalone approach. This marginal gain suggests that performance may be approaching an optimal threshold for this test set. More substantial gains were observed with the ESM1b enhanced model, which achieved a notable improvement from PR-AUC 0.7618 to 0.857. The EVE enhanced model demonstrated intermediate improvement, with PR-AUC increasing from 0.8651 to 0.907.

To assess the practical impact of model enhancement, we conducted mutation level analysis of prediction accuracy on the held-out test set. The AlphaMissense-enhanced model correctly classified 6 variants that were misclassified by standalone AlphaMissense, comprising 4 deleterious and 2 neutral variants. The EVE enhanced model demonstrated superior improvement, accurately predicting 16 variants missed by standalone EVE (13 deleterious and 3 neutral variants). The ESM1b enhanced model correctly classified 11

variants that standalone ESM1b failed to predict accurately (7 deleterious and 4 neutral variants). These findings indicate that model enhancement provides particular benefit for deleterious variant identification, with EVE showing the most substantial improvement in capturing previously missed pathogenic mutations (Table S13).

Supplementary Figures

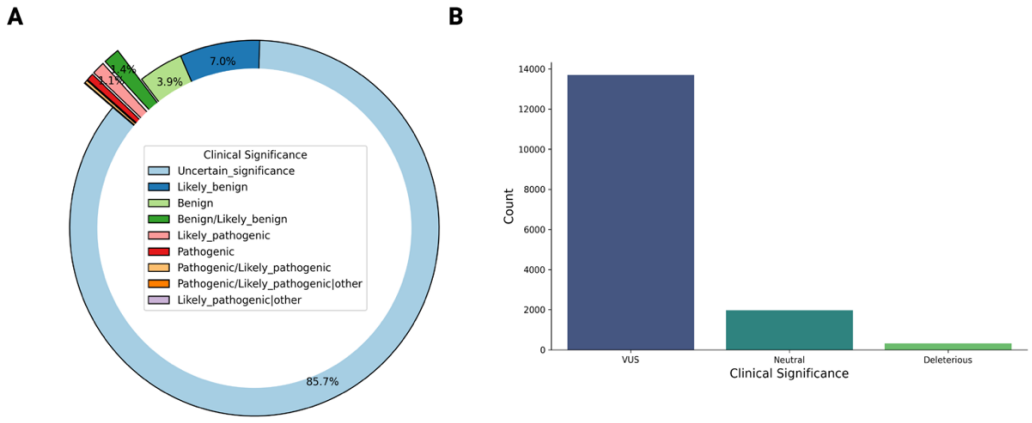

Figure S18: Variant proportions and counts in predicted intrinsic disordered regions (IDRs) in ClinVar*. A.) Proportion of variants found in predicted IDRs stratified by clinical significance listed in the ClinVar database. B.) Counts of ClinVar variants after re-grouping the Clinical significance in three functional groups (Deleterious, Neutral and VUS)*

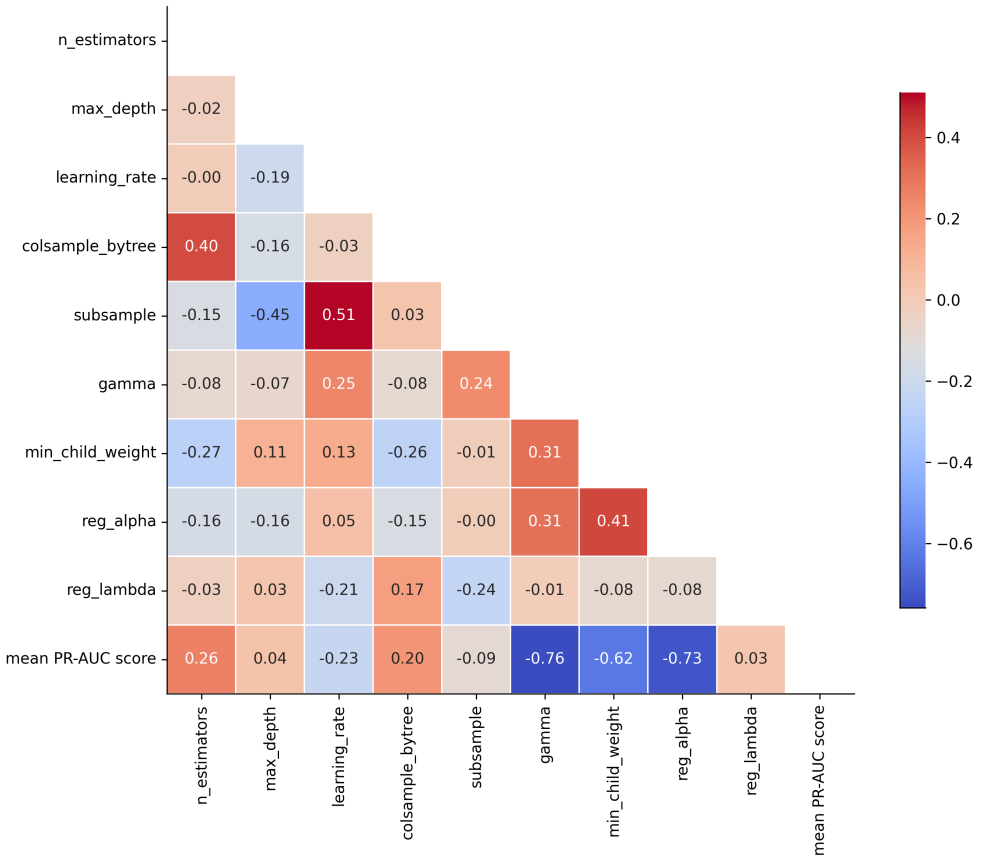

Figure S19: *The pairwise-correlation denoting the influence of hyperparameters settings to the evaluation metric mean PR-AUC and other hyperparameters from XGBoost model in predicting protein function from variants in IDRs. The analysis highlights the importance of the hyperparameter gamma and min_child_weight towards the improvement of mean PR-AUC scores.*

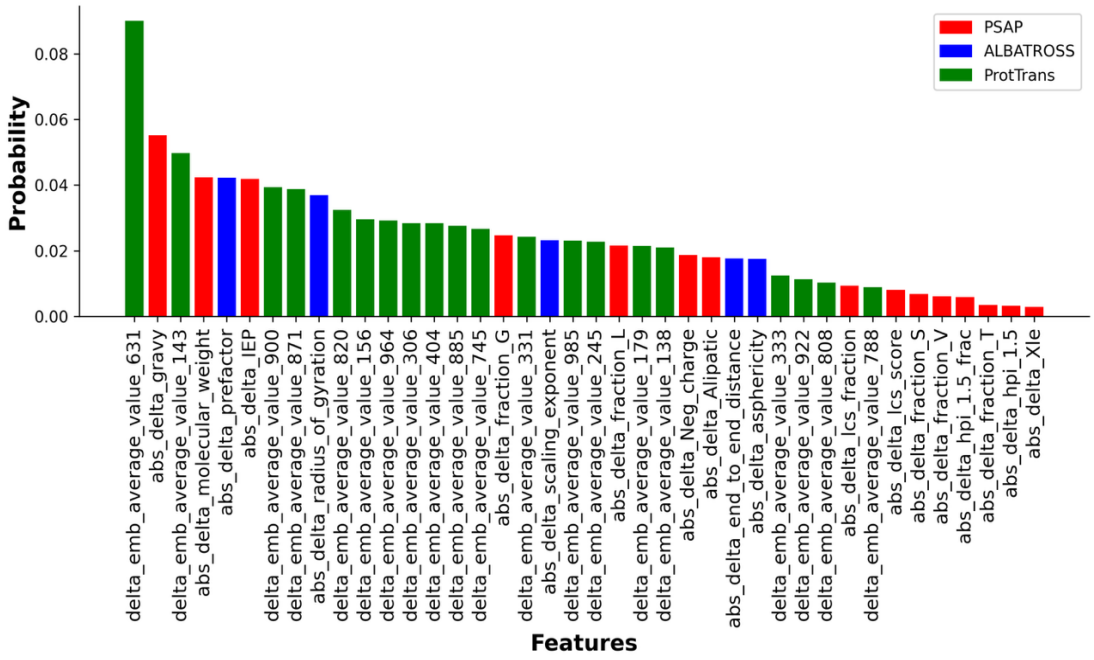

Figure S20: *The top 40 most important feature from the optimized XGBoost model trained on features predicting absolute change in phase separation, absolute change in global conformation and average embedding combination in predicting protein function from a missense variant. SHAP probabilities highlights the importance of the features towards model prediction. The colors (red, blue and green) indicate the source of the features.*

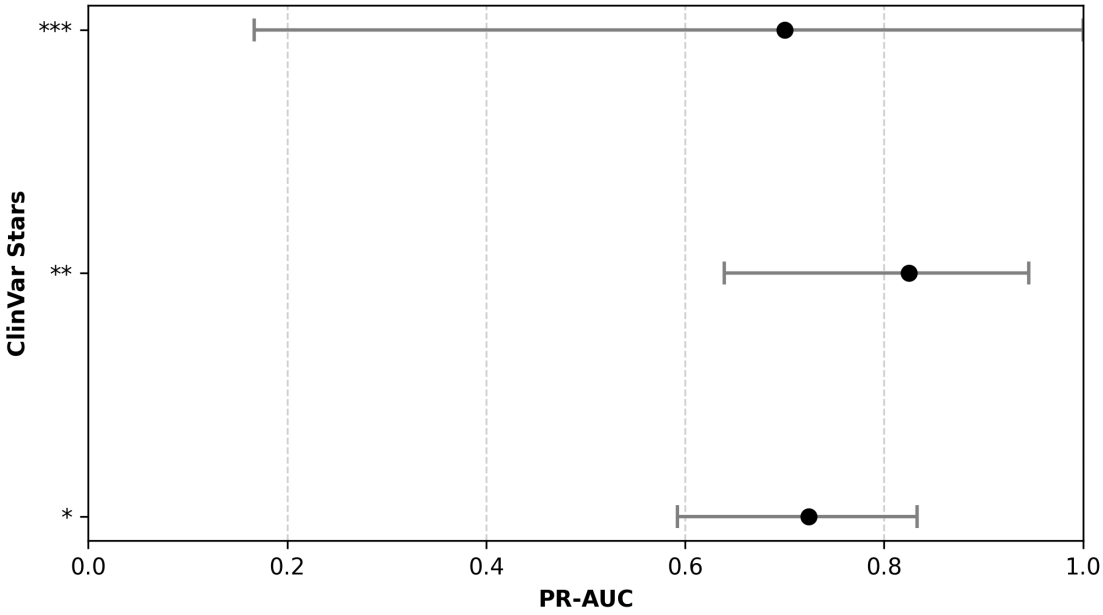

Figure S21: *AUC with 95% confidence intervals for the predictive performance of the proposed model on hold-out test variants, stratified by ClinVar review status. Variants are grouped into categories based on*

*ClinVar's star ratings: * for "criteria provided, single submitter", ** for "criteria provided, multiple submitters, no conflicts", and *** for "reviewed by expert panel"*

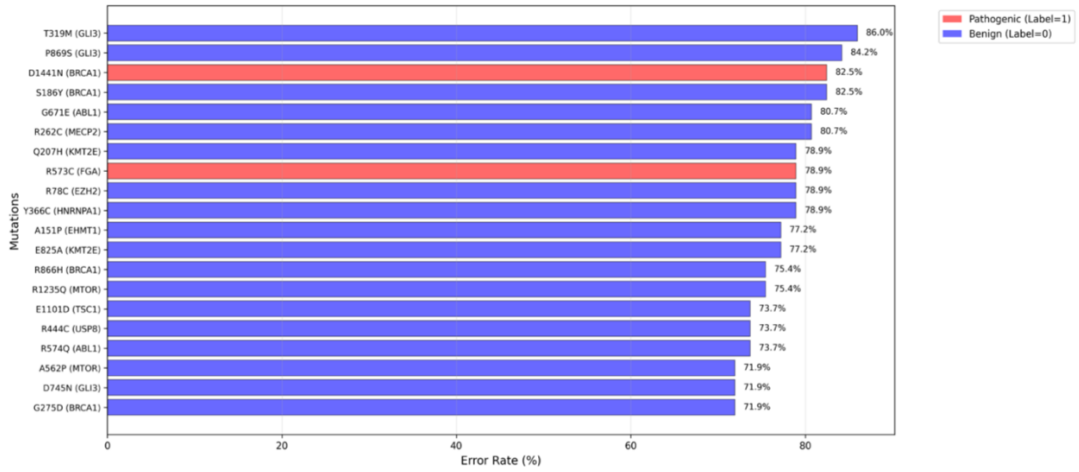

***Figure S22**: Comparative error rates for challenging missense variants across in silico predictors. The horizontal bar chart displays error rates (%) for the 20 most difficult-to-classify variants from the 421variant hold-out test dataset, showing prediction accuracy challenges for both pathogenic (red bars, Label=1) and benign (blue bars, Label=0) classifications. Variants are labeled with their amino acid change and associated gene symbol.*

Supplementary Tables

Table S12: In Silico missense predictor performance of all mutations in IDR regions predicted by AlphaFold-RSA that have ClinVar classifications

dbNSFP predictor	AUC (95% CI)	PR-AUC (95% CI)	F1 (95% CI)	Model Category
PHACTboost	0.955 (0.954-0.957)	0.954 (0.953-0.956)	0.903 (0.901-0.905)	Supervised
MetaRNN	0.949 (0.948-0.950)	0.953 (0.952-0.954)	0.881 (0.879-0.883)	Supervised
ClinPred	0.949 (0.947-0.50)	0.947 (0.945-0.949)	0.893 (0.891-0.895)	Supervised
BayesDel_addAF	0.922 (0.920-0.923)	0.924 (0.922-0.926)	0.844 (0.842-0.847)	Supervised
VARITY_R	0.92 (0.919-0.922)	0.927 (0.925-0.929)	0.85 (0.848-0.852)	Supervised
AlphaMissense	0.907 (0.906-0.909)	0.914 (0.912-0.916)	0.84 (0.838-0.842)	Unsupervised
VEST4	0.901 (0.899-0.903)	0.898 (0.895-0.900)	0.831 (0.829-0.833)	Supervised

REVEL	0.876 (0.874-0.879)	0.89 (0.888-0.892)	0.788 (0.785-0.790)	Supervised
BayesDel_noAF	0.874 (0.872-0.877)	0.882 (0.879-0.884)	0.802 (0.800-0.805)	Supervised
ESM1b	0.846 (0.844-0.848)	0.862 (0.859-0.866)	0.805 (0.803-0.808)	Unsupervised
EVE	0.714 (0.711-0.717)	0.769 (0.766-0.772)	0.696 (0.694-699)	Unsupervised

*Table S*13*: Performance Comparison of Enhanced Models vs. Standalone Predictors on Specific Missense Variants*

Mutation	***Gene***	***ClinVar***	***Improved AlphaMissense***	***Improved ESM1b***	***Improved EVE***
G730D	*CTNNB1*	Likely pathogenic	NO	NO	YES
D149H	*EMD*	Benign/Like Benign	YES	NO	NO
S509W	*GEN1*	Likely benign	NO	YES	NO
H287Y	*KLF4*	Likely benign	YES	NO	NO
R44C	*MCM2*	Pathogenic (Flagged)	NO	NO	YES
L741R	*NLGN3*	Likely pathogenic	NO	NO	YES
S352Y	*NR3C2*	Likely benign	NO	YES	YES
S419L	*NR3C2*	Likely benign	NO	NO	YES
S567L	*NR3C2*	Likely benign	NO	NO	YES
E728Q	*PPP1R12A*	Likely benign	YES	NO	NO
P837S	*PPP1R12A*	Likely benign	NO	YES	NO
S423I	*SMAD3*	Likely pathogenic	NO	NO	YES
S425C	*SMAD3*	Pathogenic	NO	NO	YES
D190G	*TNNI3*	Pathogenic	NO	YES	NO

D196N	*TNNI3*	Pathogenic/Likely Pathogenic	NO	YES	YES
E182K	*TNNI3*	Pathogenic/Likely Pathogenic	NO	YES	YES
L198V	*TNNI3*	Pathogenic/Likely Pathogenic	NO	NO	YES
N185K	*TNNI3*	Pathogenic	NO	NO	YES
N185S	*TNNI3*	Likely pathogenic/Conflicting	YES	YES	YES
N194K	*TNNI3*	Likely pathogenic	NO	YES	NO
R186Q	*TNNI3*	Pathogenic/Likely Pathogenic	YES	YES	YES
S44A	*TNNI3*	Pathogenic	YES	NO	NO
P383L	*USP8*	Likely benign	NO	YES	NO
P720R	*USP8*	Pathogenic	NO	NO	YES
S718C	*USP8*	Pathogenic	NO	NO	YES