Assessing Variants of Uncertain Significance Implicated in Hearing Loss Using a Comprehensive Deafness Proteome

Mallory R. Tollefson
The University of Iowa
Rose A. Gogal
The University of Iowa
A. Monique Weaver
The University of Iowa
Amanda M. Schaefer
The University of Iowa
Robert J. Marini
The University of Iowa
Hela Azaiez
The University of Iowa
Diana L. Kolbe
The University of Iowa
Donghong Wang
The University of Iowa
Amy E. Weaver
The University of Iowa
Thomas L. Casavant
The University of Iowa College of Nursing
Terry A. Braun
The University of Iowa
Richard J. H. Smith
The University of Iowa
Michael Schnieders (michael-schnieders@uiowa.edu)
The University of Iowa https://orcid.org/0000-0003-1260-4592

Research Article

Keywords: Deafness, AlphaFold2, Free Energy Simulation, Variant of Uncertain Significance

Posted Date: February 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2508462/v1

Version of Record: A version of this preprint was published at Human Genetics on April 22nd, 2023. See the published version at https://doi.org/10.1007/s00439-023-02559-9.

Abstract

Hearing loss is the leading sensory deficit, affecting ~ 5% of the population. It exhibits remarkable heterogeneity across 223 genes with 6,328 pathogenic missense variants, making deafness-specific expertise a prerequisite for ascribing phenotypic consequences to genetic variants. Deafness-implicated variants are curated in the Deafness Variation Database (DVD) after classification by a genetic hearing loss expert panel and thorough informatics pipeline. However, seventy percent of the 128,167 missense variants in the DVD are “variants of uncertain significance” (VUS) due to insufficient evidence for classification. Here, we use the deep learning protein prediction algorithm, AlphaFold2, to curate structures for all DVD genes. We refine these structures with global optimization and the AMOEBA force field and use DDGun3D to predict folding free energy differences ($\Delta\Delta G_{\text{Fold}}$) for all DVD missense variants. We find that 5,772 VUSs have a large, destabilizing $\Delta\Delta G_{\text{Fold}}$ that is consistent with pathogenic variants. When also filtered for CADD scores (> 25.7), we determine 3,456 VUSs are likely pathogenic at a probability of 99.0%. These VUSs affect 119 patients (~ 3% of cases) sequenced by the OtoSCOPE targeted panel. Approximately half of these patients previously received an inconclusive report, and reclassification of these VUSs as pathogenic provides a new genetic diagnosis for six patients.

Introduction

Hearing loss is the most prevalent sensory deficit, affecting approximately 5% of the world’s population. In its evaluation, following an audiogram, genetic sequencing with a multi-gene panel is recommended as the most informative diagnostic test for infants and children with hearing loss(Alford et al. 2014; Li et al. 2022; Liming et al. 2016; Shearer and Smith 2015). It facilitates identification of an underlying cause in 40% – 56% of patients in an outbred population(Shearer and Smith 2015) and up to 72% in certain ethnicities(Sloan-Heggen et al. 2016). Currently, most panel-based tests screen 23–245 genes for variants that may be implicated in hearing loss(Sloan-Heggen and Smith 2016). OtoSCOPE, the panel we first developed in 2010(Shearer et al. 2010), contains 223 genes in its current iteration (version 9), which in aggregate includes approximately 592,770 nucleotides of coding sequence.

In each patient screened, an average of 545 genetic variants is identified(Shearer et al. 2013). Ascribing a pathogenic consequence to these variants is challenging and requires deafness-specific expertise. To help meet this challenge, we developed the Deafness Variation Database(Azaiez et al. 2018) (DVD). This resource includes 128,167 missense variants, which are classified by a genetic hearing loss expert panel and thorough informatics pipeline into one of five categories: benign (B, n = 1,725), likely benign (LB, n = 27,907), likely pathogenic (LP, n = 2,441), pathogenic (P, n = 6,328), and variant of uncertain significance (VUS, n = 89,766). If a variant is classified as a VUS, a definitive diagnosis cannot be made for patients affected by that variant. For variant reclassification, additional studies are required and can include family segregation analysis, identification of the variant in a family member with hearing loss or an unrelated proband, or specific wet lab based functional evidence(Richards et al. 2015). Given the disproportionate number of VUSs, making genotype-phenotype correlations from such evidence is infeasible. Therefore, we sought to apply deep learning-based protein structure prediction(Jumper et al. 2021), atomic resolution simulation(Tollefson et al.

2019), and thermodynamic analysis(Montanucci et al. 2022; Montanucci et al. 2019) to all DVD missense variants classified as VUSs to determine whether it would be possible to reclassify some VUSs as P.

In 2019, protein structures of deafness-associated genes were known for fewer than 40% of all proteins and missense variants implicated in hearing loss(Tollefson et al. 2019), relegating computational structural variant analysis to only those variants with solved protein structures. The release of the AlphaFold2(Jumper et al. 2021; Tunyasuvunakool et al. 2021) neural network enabled *ab initio* computational prediction of protein structures with an accuracy comparable to experimentally obtained structures. Using AlphaFold2, a comprehensive deafness proteome and *in silico* structural analysis of all deafness-associated variants became possible.

It is well recognized that a protein’s function and its stability are related(Araya et al. 2012; Talley and Alexov 2010). On that basis, computational folding free energy differences ($\Delta\Delta G_{Fold}$) have been used to characterize genes and missense variants implicated in deafness(Buonfiglio et al. 2022) including protein-specific studies (*e.g. FGFR1*(Doss et al. 2012), *TMC1*(Hilgert et al. 2008), *PNPT1*(Bereshneh et al. 2021), *PRPS1*(Agrahari et al. 2018)) by quantifying the degree of protein misfolding caused by a variant. When a missense variant results in protein misfolding, the protein may be targeted for degradation(Balchin et al. 2016; Goldberg 2003; McCafferty and Sergeev 2016; Stein et al. 2019). With AlphaFold2 protein structures, $\Delta\Delta G_{Fold}$ analysis and an accompanying prediction of protein misfolding, abrogated function and possible degradation can be done on a deafness proteome wide basis. However, computing $\Delta\Delta G_{Fold}$ using protein structures from AlphaFold2 as input to rigorous molecular dynamics-based simulation for all 128,167 missense variants listed in the DVD is currently intractable due to computational expense.

As an alternative, we use a high-throughput *in silico* tool to predict $\Delta\Delta G_{Fold}$(Guerois et al. 2002; Montanucci et al. 2019; Parthiban et al. 2006; Rodrigues et al. 2021; Zhou and Zhou 2002) and identify VUSs most likely to induce significant protein misfolding (often $\Delta\Delta G_{Fold}$ >2–3 kcal/mol), potentially allowing these variants to be classified as P. First, we use AlphaFold2 to curate full-length, isoform-specific protein structures for all genes in the DVD (OtoProtein2). We then reduce biophysical inaccuracies (*i.e.*, steric clashes and side-chain errors) in the OtoProtein2 structures by refining them with an amino acid side-chain optimization algorithm(Tollefson et al. 2019) and the AMOEBA(Ponder et al. 2010) polarizable force field. Finally, we use DDGun3D(Montanucci et al. 2019) to predict $\Delta\Delta G_{Fold}$ for all missense variants in the DVD and resolve classifications for VUSs that cause protein instability.

We find that 5,772 VUSs have a $\Delta\Delta G_{Fold}$ consistent with P variants. When filtered for high CADD scores (> 25.7) in addition to large $\Delta\Delta G_{Fold}$, we identify 3,456 destabilizing VUSs that are P at a probability of 99.0%. These priority VUSs affect 119 patients sequenced by OtoSCOPE (~ 3% of cases), half of whom previously received inconclusive reports. Finally, an upgraded classification of P for these priority VUSs results in a definitive genetic diagnosis for six patients.

Materials And Methods

Predicting Deafness Protein Structures with Deep Learning

We used the AlphaFold2(Jumper et al. 2021) deep learning algorithm to predict isoform-specific protein structures for the 218 protein-coding genes in the Deafness Variation Database(Azaiez et al. 2018) (DVD). Trained on experimentally known protein structures from the Protein Data Bank (PDB)(Berman et al. 2000), the AlphaFold2 neural network predicts protein structures from amino acid sequences to an accuracy comparable to experimental results using two modules(Jumper et al. 2021). The first module develops a general hypothesis for the protein's structure in part from relationships between co-evolving amino acids associated with a multiple sequence alignment. The second module predicts the spatial relationships between subsequent amino acids to produce an explicit three-dimensional protein structure. By default, the two modules are generally applied in three iterative cycles to refine the structure prediction; however based on prior work(Mirdita et al. 2022) we applied the modules in 15 cycles to achieve higher quality predictions.

Biophysical Refinement of the AlphaFold2 Deafness Proteome

To improve the biophysics of the AlphaFold2 protein predictions (*i.e.*, reduce atomic clashes, choose favorable amino acid side-chain conformations, *etc.*), we employed both local and global optimization techniques with the AMOEBA(Ponder et al. 2010; Shi et al. 2013) polarizable force field. We first locally minimized all AlphaFold2 protein structures with the limited memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton minimization to relax the backbone and reduce atomic clashes in each protein. After local minimization, we applied a global amino acid side-chain optimization algorithm(Tollefson et al. 2019) to determine energetically favorable side-chain conformations for the amino acids in the AlphaFold2 proteins. We then used the heuristic MolProbity(Chen et al. 2010; Davis et al. 2007) algorithm to evaluate structures before and after optimization to quantify for each protein the improvement in atomic clashes, backbone angles, and side-chain conformations.

Predicting $\Delta\Delta G_{Fold}$ and Prioritizing Missense Variants in the DVD

We predicted $\Delta\Delta G_{Fold}$ for every missense variant in the DVD(Azaiez et al. 2018) using the optimized protein structures and the high throughput *in silico* method DDGun3D(Montanucci et al. 2019). DDGun3D(Montanucci et al. 2019) predicts a $\Delta\Delta G_{Fold}$ by assessing the biochemical features of a variant using its three-dimensional protein structure. We compared the distribution of $\Delta\Delta G_{Fold}$ in variants with P and B DVD(Azaiez et al. 2018) classifications. Using thermodynamic (see supplementary information) observations, we identified a $\Delta\Delta G_{Fold}$ threshold to predict genetic variants that induce significant misfolding, loss of function and possibly protein degradation. We used classified DVD variants to determine the positive predictive value (PPV) of this $\Delta\Delta G_{Fold}$ threshold. We applied this threshold to all P variants to determine which P variants are deleterious due to protein misfolding. We further applied this threshold to all VUSs in the DVD to determine which VUSs most likely impact protein misfolding and are therefore most likely to be P.

Integrating CADD Scores with $\Delta\Delta G_{Fold}$ to Prioritize Variants

We combined the $\Delta\Delta G_{Fold}$ predictions and threshold with CADD(Rentzsch et al. 2018) scores to prioritize VUSs most likely to be deleterious. Because variants with higher CADD scores are predicted to be more damaging(Rentzsch et al. 2018), we anticipated variants with both a large $\Delta\Delta G_{Fold}$ and a high CADD score are more likely to be P. We set the CADD score threshold (25.7) to reflect a 99% PPV for classified DVD variants to be P when both $\Delta\Delta G_{Fold}$ and CADD scores are combined. We then applied both the CADD threshold and the $\Delta\Delta G_{Fold}$ threshold to identify VUSs that are deleterious with 99% certainty.

Curating Variant Features for Further Analysis

In addition to annotating $\Delta\Delta G_{Fold}$ and CADD scores for each DVD variant, we aggregated features from the optimized structures to be used for variant analysis, prioritization, and deep learning. For each variant, we collected AlphaFold2's confidence in the protein structure at that variant's position, which can be used to prioritize analysis of variants in regions where protein structure is predicted with a high degree of confidence. Similarly, because amino acids buried within a protein domain are often intolerant of variation as compared to amino acids on the surface of a protein domain, we computed the percent of solvent accessible surface area (SASA) for each DVD variant. Finally, previous work has shown that minor allele frequency (MAF) can be used to classify common variants as LB in deafness-associated genes(Shearer et al. 2014); therefore, we included the MAF for each variant in the dataset of variant features.

Results

Quality and Characteristics of Deafness Protein Structure Predictions

Using AlphaFold2, we developed complete protein structures for all genes and relevant isoforms in the Deafness Variation Database(Azaiez et al. 2018) (DVD, Fig. 1a, b). Called OtoProtein2, this dataset increases structural coverage of the deafness proteome from approximately 30% by experimental and homology protein structures curated during prior work(Tollefson et al. 2019) (*i.e.*, called OtoProtein) to 100% (Fig. 1c). For each amino acid in a prediction, AlphaFold2 provides a unitless confidence score ranging from 1 to 100, with higher scores corresponding to higher confidence in the prediction. Model confidence is > 70 for 64% of wild-type amino acids and 60% of missense variant locations in the deafness proteome. The remaining amino acids and missense variants fall in regions that are predicted only with low confidence (*i.e.*, confidence < 70).

[Figure 1 Here]

Approximately 41% of missense variants in the deafness proteome belong to a functional protein domain as characterized by InterPro(Apweiler et al. 2001; Blum et al. 2021), while 59% belong to flexible termini, natively disordered regions, or uncharacterized domains (Table 1). InterPro characterized domains are enriched in high confidence protein structures, while natively disordered regions exist in lower confidence regions. Of the 128,167 missense variants in the deafness proteome, 34% belong to both a characterized domain and a high confidence structural region. Although missense variants are evenly distributed across InterPro characterizations (*e.g.*, 41.3% and 41.4% of wild-type amino acids and missense variants are in a characterized domain, respectively), benign and likely benign variants favor lower confidence, uncharacterized regions while pathogenic and likely pathogenic variants favor higher confidence regions with functional protein domains (Tables 2, S1 and S2).

Table 1

Number and percent of Deafness Variation Database missense variants belonging to each AlphaFold2 confidence range based on characterization from InterPro.

	Model Confidence			
InterPro Domain	< 50	50–70	70–90	> 90
Characterized (41.3%)	3371 (2.6%)	5610 (4.4%)	23991 (18.7%)	20028 (15.6%)
Uncharacterized (58.6%)	40230 (31.4%)	8753 (6.8%)	15505 (12.1%)	10679 (8.3%)
Total	43611 (34.0%)	14393 (11.2%)	39574 (30.8%)	30956 (23.9%)

Table 2

Number and percent of Deafness Variation Database missense variants belonging to each AlphaFold2 confidence range based on Deafness Variation Database classification.

	Model Confidence			
DVD Classification	< 50	50–70	70–90	> 90
B (1.4%)	719 (0.6%)	231 (0.2%)	506 (0.4%)	269 (0.2%)
LB (21.8%)	15395 (12.0%)	3153 (2.5%)	5973 (4.7%)	3386 (2.6%)
LP (2.1%)	579 (0.5%)	203 (0.2%)	827 (0.7%)	832 (0.7%)
P (4.9%)	1201 (0.9%)	470 (0.4%)	2084 (1.6%)	2573 (2.0%)
VUS (70.1%)	25707 (20.1%)	10306 (8.0%)	30106 (23.5%)	23647 (18.5%)
Total	43611 (34.1%)	14393 (11.3%)	39574 (30.9%)	30956 (24.0%)

Biophysical Refinement of the Protein Structure Predictions

We applied a global side-chain optimization algorithm(Tollefson et al. 2019) and local minimization with the AMOEBA force field to each of the OtoProtein2 structures, assessing the quality of the structures before and after optimization using the MolProbity algorithm. Compared to the initial deep learning predictions from AlphaFold2, the OtoProtein2 dataset reduced steric clashes per 1000 atoms from 20.75 to 0.11, lowered the percent of rotamers in energetically unfavorable conformations from 4.32–1.12%, decreased the backbone angle outliers from 15.25–1.05%, and increased the favored backbone angles from 76.21–93.50% (Table 3). The refinement procedure improved the dataset’s mean MolProbity score from 2.86 to 0.97 (Fig. 3), making the OtoProtein2 structural quality equivalent to experimental structures at atomic resolution.

Table 3
Average MolProbity refinement statistics for all deafness associated protein models in OtoProtein2 before and after optimization with Force Field X. A lower clash score, a lower percentage of poor rotamers, a higher percentage of favored backbone phi/psi angles, fewer backbone outliers and lower MolProbity score are each better.

Optimization	Clash Score	Poor Rotamers	Favored Backbones	Backbone Outliers	MolProbity Score
AlphaFold2	20.75	4.32%	76.21%	15.25%	2.86
OtoProtein2	0.11	1.12%	93.50%	1.05%	0.97

[Figure 2 Here]

We have incorporated the optimized OtoProtein2 structures with the DVD (www.deafnessvariationdatabase.org) to be visualized in the context of the comprehensive genetic information available therein. With 100% coverage, any DVD missense variant can be selected and visualized on its corresponding protein structure. These structures are also available for download on Github (https://github.com/SchniedersLab/OtoProtein).

Using $\Delta\Delta G_{Fold}$ Predictions to Prioritize Variants of Uncertain Significance

We used DDGun3D(Montanucci et al. 2019) and the optimized OtoProtein2 structures to predict the folding free energy differences ($\Delta\Delta G_{Fold}$) for 128,167 missense variants in the DVD (Fig. 3 and Table S3). In total, 75,072 variants (59%) are destabilizing ($\Delta\Delta G_{Fold}$>0), 34,253 variants (27%) are stabilizing ($\Delta\Delta G_{Fold}$<0), and the remainder are neutral ($\Delta\Delta G_{Fold}$=0). B variants show a mildly destabilizing mean $\Delta\Delta G_{Fold}$ of 0.13 kcal/mol while P variants have a higher destabilizing mean $\Delta\Delta G_{Fold}$ of 0.80 kcal/mol (p-value = $8.54\text{x}10^{-197}$). Within each variant classification (B: p-value = $1.006\text{x}10^{-2}$; P: p-value = $3.68\text{x}10^{-114}$), variants in high confidence regions of a protein structure (*i.e.*, often functional regions) have a higher mean $\Delta\Delta G_{Fold}$ and a wider distribution of $\Delta\Delta G_{Fold}$ than variants that fall within low confidence regions (*i.e.*, often natively disordered protein regions).

[Figure 3 Here]

Using thermodynamics principles (see derivation in supplementary information), a $\Delta\Delta G_{Fold}$ of > 1.8 kcal/mol represents a 20-fold decrease in the ratio of folded to unfolded protein. At this threshold, variants with a $\Delta\Delta G_{Fold}$ larger than 1.8 kcal/mol are appreciably destabilizing to a protein fold, likely resulting in loss of function or protein degradation. The 1.8 kcal/mol threshold results in a positive predictive value (PPV) of 97.1% and specificity of 98.2%, with nearly 17% of pathogenic variants (1067 of all P variants in the DVD) falling above 1.8 kcal/mol. Using the $\Delta\Delta G_{Fold}$ with a 1.8 kcal/mol cutoff, 5,772 VUSs are deleterious due to destabilization of the protein fold, loss of native function and possibly protein degradation. The presence of both destabilizing and over stabilizing variants are known to result in disease phentoypes(Stefl et al. 2013; Takano et al. 2012; Witham et al. 2011), and we observed that some pathogenic DVD variants have a largely over stabilizing $\Delta\Delta G_{Fold}$ (<-1.8 kcal/mol). However, using a -1.8 kcal/mol threshold (*i.e.*, a 20-fold increase in the ratio of folded to unfolded protein) to identify over stabilizing variants resulted in a PPV of only 93.0% and applied to only 53 pathogenic variants. Therefore, we focused attention on only destabilizing variants. With nearly 90,000 VUSs in the DVD, DDGun3D provides an efficient means for calculating $\Delta\Delta G_{Fold}$ and identifying deleterious variants.

Integrating CADD Scores with $\Delta\Delta G_{Fold}$ to Prioritize VUSs

CADD scores(Rentzsch et al. 2018) can be used in combination with $\Delta\Delta G_{Fold}$ to prioritize variants most likely to be deleterious. Higher CADD scores are associated with P and LP variants (Fig. 4a). These variants also favor protein regions with high confidence (Fig. 4b) and consist primarily of domains and motifs that are intolerant to variation. Establishing a CADD threshold independently has a reasonable PPV (*e.g.*, a CADD cutoff of 20 results in a PPV of 88.3%). We applied a CADD cutoff of 25.7 and combined this threshold with the $\Delta\Delta G_{Fold}$ threshold, which resulted in a PPV of 99% and a specificity of 99.5%. While these stringent CADD and $\Delta\Delta G_{Fold}$ thresholds limit prioritization to 3,456 destabilizing VUSs (Tables 4 and S4), these VUSs can be classified as LP due to protein misfolding (Fig. 4C).

[Figure 4 Here]

Table 4

Summary of genes with 30 or more prioritized VUSs per 1000 amino acids in length. A comprehensive list of all prioritized VUSs is available in Table S4. These VUSs were prioritized based on having a ΔΔGFold > 1.8 and a CADD score > 25.7.

Gene	Protein Family	Variant Density	Protein Length	# VUSs	Mean $\Delta\Delta G_{Fold}$	Mean CADD
ATP6V1B1	ATPase	33.1	513	17	2.8	27.8
CDC14A	Tyrosine phosphatase	36.9	623	23	2.7	28.3
CLRN1	Clarin	43.1	232	10	2.3	26.6
DCAF17	Not assigned	42.3	520	22	3.2	27.9
DIABLO	Not assigned	58.6	239	14	2.7	28.4
ELMOD3	Not assigned	31.5	381	12	3.1	27.6
GIPC3	GIPC	41.7	312	13	3.1	27.5
GJB2	Connexin	44.2	226	10	3.1	27.6
GJB3	Connexin	40.7	270	11	3.4	27.3
GRXCR1	GRXCR1	34.5	290	10	2.8	27.6
GSDME	Gasdermin	38.3	496	19	3.4	27.4
HARS2	Aminoacyl-tRNA synthetase	33.2	512	17	2.7	28.8
KARS1	Aminoacyl-tRNA synthetase	30.4	625	19	2.7	28.3
LHFPL5	LHFP	32.0	219	7	3.3	28.4
LOXL3	Lysyl oxidase	35.9	753	27	3.0	27.8
MANBA	Glycosyl hydrolase	30.7	879	27	3.1	27.7
MASP1	Peptidase	50.8	728	37	3.1	28.6
MSRB3	Sulfoxide reductase	37.8	185	7	3.4	28.3
MYO3A	Myosin-kinesin ATPase	38.4	1616	62	3.0	28.4
MYO6	Myosin-kinesin ATPase	30.9	1294	40	3.1	27.9
MYO7A	Myosin-kinesin ATPase	39.3	2215	87	3.0	28.0
NARS2	Aminoacyl-tRNA synthetase	46.1	477	22	2.8	28.3
OTOF	Ferlin	30.0	1997	60	3.2	28.4
OTOGL	Otogelin	61.0	2344	143	3.1	28.2
PCDH15	Not assigned	30.7	1790	55	3.1	27.9

POLR1C	RNA polymerase	46.2	346	16	2.9	27.9
RDX	Not assigned	39.7	604	24	3.0	27.5
SEMA3E	Semaphorin	31.0	775	24	3.3	28.1
SLC17A8	Sodium/anion cotransporter	30.6	589	18	2.8	28.9
SLC19A2	Thiamine transporter	70.4	497	35	3.5	28.1
SLC22A4	Cation transporter	38.1	551	21	3.0	28.3
SLC26A4	SLC26A/SulP transporter	57.7	780	45	2.9	28.1
SLC44A4	Choline transporter-like	54.9	710	39	3.0	28.5
SLC52A2	Riboflavin transporter	36.0	445	16	3.2	26.9
SLC52A3	Riboflavin transporter	34.1	469	16	3.2	27.3
TECTA	Not assigned	39.0	2155	84	3.1	28.0
TMC1	TMC	31.6	760	24	3.3	29.0
TSPEAR	Not assigned	31.6	601	19	3.5	28.1
WFS1	Not assigned	51.7	890	46	3.0	27.7

We found that P and LP variants are often in buried residues (*i.e.*, solvent accessible surface area near zero percent) with confident structure regions (Fig. 5ab). The prioritized dataset of 3,456 VUSs are consistently present in buried, confident regions of the OtoProtein2 structures (Fig. 5c). Additionally, $\Delta\Delta G_{Fold}$, CADD scores, solvent accessible surface area, and structure confidence from the OtoProtein2 models for all variants in the DVD can be utilized for deep learning applications or for variant analysis.

[Figure 5 Here]

Discussion

The classification of genetic variation in relationship to a disease phenotype is challenging. For hearing loss, the DVD uses an expert panel and rigorous informatics pipeline to classify changes in deafness-associated genes based on evidence of pathogenicity. This database includes over 128,167 missense variants, the majority of which (> 70%) are classified as VUSs due to insufficient evidence to classify as P or B. A VUS classification is problematic for both the healthcare provider and the patient as a definitive diagnosis cannot be made. Here we show that *in silico* $\Delta\Delta G_{Fold}$ can resolve a portion of VUSs by quantifying the change in protein stability induced by a variant, consequently providing insight as to the variant's mechanism of action (*i.e.*, the variant induces protein misfolding) and its pathogenicity. We used AlphaFold2 and a global optimization algorithm(Tollefson et al. 2019) to develop OtoProtein2, a database of optimized, isoform specific, full-length protein structures for every gene in the DVD. We then used the OtoProtein2 models and the *in silico* tool, DDGun3D(Montanucci et al. 2019), to predict $\Delta\Delta G_{Fold}$ for every missense variant in the DVD. We found that $\Delta\Delta G_{Fold}$ greater than 1.8 kcal/mol are predictive of P variants at a rate of 97.1%. Combining large $\Delta\Delta G_{Fold}$ (> 1.8 kcal/mol) and large CADD scores (> 25.7) results in a positive predictive value (PPV) of 99.0%. Using these $\Delta\Delta G_{Fold}$ and CADD thresholds, we identified 3,456 VUSs that are LP due to protein misfolding.

Of these 3,456 prioritized VUSs, we have observed 79 across 119 patients who underwent comprehensive genetic testing using OtoSCOPE. Over half of these patients (60 patients) previously received an inconclusive genetic diagnosis. In five patients with variants affecting autosomal recessive genes, the proband carried a second LP/P variant in the gene. Segregation analysis (SA) confirmed that the second LP/P variant occurs on the opposite allele in three of five patients; in the remaining two patients, SA was not available. One patient carried a variant affecting an autosomal dominant gene. The work here delivers a definitive genetic diagnosis for these six patients and directly impacts their subsequent healthcare (Table 5). For example, patient six carried a known P variant in *TMPRSS3* in trans with a novel missense variant predicted to cause protein destabilization by this work (Fig. 6). The phenotype of the patient's hearing loss is highly specific for *TMPRSS3*-related hearing loss (DFNB8/10). Reclassification of patient six's novel missense variant from VUS to LP results in a definitive genetic diagnosis, ultimately directing subsequent medical care and recurrence risk calculations for offspring. Current guidelines established by the American College of Medical Genetics and Genomics (ACMG) for hearing loss do not incorporate *in silico* $\Delta\Delta G_{Fold}$ calculations, however, our work demonstrates the utility of protein modeling for hearing loss diagnostics. Further work is indicated to guide incorporation of protein modeling into ACMG guidelines for hearing loss and deafness.

[Figure 6 Here]

Table 5
Patients with definitive diagnoses from upgraded classification of priority VUSs. Segregation analysis confirms that the second variant occurs on the opposite allele in three probands. Table cells with NA are not available.

Patient ID	Gene	Inheritance	Priority VUS	Second Variant (Classification)	Segregation Analysis
1	*CDH23*	AR	NP_071407.4:p.Tyr2883Ser	Arg2795Ter (P)	NA
2	*GRXCR1*	AR	NP_001073945.1:p.Tyr142Cys	Gln283Ter (P)	Yes
3	*HARS2*	AR	NP_036340.1:p.Tyr364Cys	Arg150Cys (LP)	Yes
4	*MYO6*	AD	NP_001355794.1:p.Cys1236Arg	None	NA
5	*PDZD7*	AR	NP_001182192.1:p.Ile269Ser	Arg56ProfsTer24 (P)	NA
6	*TMPRSS3*	AR	NP_076927.1:p.Met384Lys	His70ThrfsTer19 (P)	Yes

The number of prioritized VUSs and impacted patients is greatly affected by adjustments to the $\Delta\Delta G_{Fold}$ and CADD thresholds. We used a $\Delta\Delta G_{Fold}$ threshold of 1.8 kcal/mol and a CADD threshold of 25.7 to reach a PPV of 99.0% (false positive rate < 0.5%), but by increasing the CADD threshold to 30.0, the PPV approaches 100%. These stringent thresholds leave negligible room for a false positive diagnosis but provide a prioritized dataset of only 419 VUSs that are LP. Seven of these 419 VUSs impact 18 OtoSCOPE patients. Alternatively, a more lenient PPV of 95% is reached by disregarding CADD scores and dropping the $\Delta\Delta G_{\text{Fold}}$ cutoff to 1.0 kcal/mol. These parameters provide a substantially larger dataset of 12,585 VUSs that are LP, albeit with a 5.6% false positive rate and impact 775 OtoSCOPE patients.

Though we applied the $\Delta\Delta G_{\text{Fold}}$ and CADD thresholds on a deafness-proteome-wide scale, these cutoffs can be tuned to better fit a protein, domain, or amino acid specific level. Biochemical, environmental, and structural differences contribute to a protein's ability to tolerate changes to its structure. For example, *ACTG1* encodes gamma actin, a highly conserved cytoskeletal protein; even small $\Delta\Delta G_{\text{Fold}}$ significantly disrupt gamma actin's structure and function. While no P *ACTG1* variants from the DVD surpass both the $\Delta\Delta G_{\text{Fold}}$ and CADD cutoffs for the proteome-wide scale, a smaller $\Delta\Delta G_{\text{Fold}}$ threshold may detect subtle structural changes that will affect gamma actin's highly conserved structure. Similarly, different domains within an individual protein can benefit from domain-specific $\Delta\Delta G_{\text{Fold}}$ analysis. Cochlin, the protein product of the *COCH* gene, has one Limulus factor C (LCCL) domain and two Von Willebrand factor A (VWFA) domains. P variants in *COCH* are known to localize in the LCCL and second VWFA domains(Gallant et al. 2013). Known P variants aggregating in just one of cochlin's two VWFA domains demonstrate the need for domain-specific analysis to identify which domains are more sensitive to amino acid variation and are intolerant of misfolding. Even individual amino acid characteristics such as the structural confidence of the wild-type amino acid, SASA, or number of hydrogen bonds can affect an amino acid's ability to tolerate a missense variant that disrupts the protein's structure. As approaches for $\Delta\Delta G_{\text{Fold}}$ predictions are improved, context-dependent thresholds will be significant for variant interpretation.

The $\Delta\Delta G_{\text{Fold}}$ and CADD thresholds used to identify VUSs that induce substantial protein destabilization can also provide an estimate of the number of deafness-causing genetic variants yet to be classified as P. Because $\Delta\Delta G_{\text{Fold}}$ quantifies the disruption to protein folding induced by variants, $\Delta\Delta G_{\text{Fold}}$ resolves only those VUSs that are P due to protein misfolding. Applying these thresholds to listed P and LP variants in the DVD allows us to identify that subset of missense variants that destabilize protein structure. Of the 6,328 known P variants, 793 (12.5%) exceed the $\Delta\Delta G_{\text{Fold}}$ and CADD thresholds and fall into this category, while the remaining P variants (5,535 variants, or 87.5%) are P for reasons unrelated to protein misfolding. Consequently, if the 3,456 VUSs we identified as LP due to misfolding represent ~ 12.5% of the remaining deleterious variants to be found, we estimate that approximately 24,192 VUSs are P for reasons unrelated to protein misfolding.

There are two important limitations to this work: 1) the accuracy of $\Delta\Delta G_{\text{Fold}}$ predictions and 2) the inherent ability of $\Delta\Delta G_{\text{Fold}}$ to quantify only protein misfolding. With respect to the former, DDGun3D predictions of $\Delta\Delta G_{\text{Fold}}$ are expected to be within ~ 1.5 kcal/mol of an experimentally known $\Delta\Delta G_{\text{Fold}}$(Montanucci et al. 2019), and the leading molecular dynamics software (FEP+) for calculating $\Delta\Delta G_{\text{Fold}}$ is within ~ 1.1 kcal/mol of the experimentally known values(Duan et al. 2020). While this degree of accuracy is sufficient to identify VUSs that are LP (*i.e.*, impact protein folding), more refinement may be needed for validating and discriminating amongst highly similar variants. There is, however, a trade-off in time. DDGun3D $\Delta\Delta G_{\text{Fold}}$ requires only minutes of compute time, while an equivalent $\Delta\Delta G_{\text{Fold}}$ calculation(Duan et al. 2020) with the Nanoscale Molecular Dynamics (NAMD) software package(Chen et al. 2020) requires on the order of one month of simulation time using a Graphical Processing Unit (GPU). This time increase also makes calculating $\Delta\Delta G_{Fold}$ with FEP+(Duan et al. 2020) or NAMD too computationally expensive for a dataset of 128,167 variants. However, these simulations may be suitable for systematically improving $\Delta\Delta G_{Fold}$ results of the most noteworthy prioritized VUSs or for validation prior to wet-lab experiments.

With respect to the second limitation, $\Delta\Delta G_{Fold}$ quantifies only the change in protein stability induced by a variant, and is therefore limited to testing the hypothesis that a missense variant disrupts protein folding(Stefl et al. 2013). Although $\Delta\Delta G_{Fold}$ provides a biochemical hypothesis for one mechanism by which a variant can affect protein function (*i.e.*, protein misfolding), $\Delta\Delta G_{Fold}$ does not test for possible pathogenicity due to reasons unrelated to protein misfolding such as interrupting an active site(Zhang et al. 2011; Zhang et al. 2010) or altering protein-protein interactions(Teng et al. 2009).

Future directions for this work include computing binding free energy differences (*i.e.*, $\Delta\Delta G_{Bind}$) and expanding our analysis beyond missense variants. In contrast to $\Delta\Delta G_{Fold}$, $\Delta\Delta G_{Bind}$ quantifies the difference in binding caused by a missense variant and tests the hypothesis that a variant alters a protein-protein interaction. Accurate structures of protein complexes and sufficient knowledge of interactions are a prerequisite for computing meaningful $\Delta\Delta G_{Bind}$, and while progress is being made in this direction (methods such as AlphaFold2-Multimer(Bryant et al. 2022), ColabFold(Mirdita et al. 2022), and AF2Complex(Gao et al. 2022) can predict protein complexes), only ~ 20% of complex predictions are considered high accuracy according to criteria established by the Critical Assessment of Predicted Interactions(Yin et al. 2022). Further, finite hardware memory combined with the memory requirements for deep learning-based protein model predictions often require that monomeric proteins be predicted in segments. This memory limitation is only exacerbated by the prediction of protein complexes where memory limits are more easily reached. Nevertheless, attaining a comprehensive model of the deafness interactome and subsequent analysis of $\Delta\Delta G_{Bind}$ will be the subject of future studies. The analysis of indels, non-coding variants, and other variants, are beyond the scope of our current work, however, prioritization and characterization of these variants should be considered in context with the VUSs prioritized herein. Regardless of the work remaining, the deafness proteome and $\Delta\Delta G_{Fold}$ analysis we present has revealed trends for P variants and provides insight on VUSs that are LP due to protein misfolding.

In summary, by using *ab initio* protein structure prediction, optimization, and thermodynamic analysis, with 99% confidence, we have identified 3,456 VUSs that are LP in patients with hearing loss due to protein misfolding. The deafness protein structures developed here have been incorporated with the DVD to inform deafness-associated variant analysis. As atomic resolution protein structures and *in silico* variant analysis techniques progress, continued and refined analysis of free energy differences for deafness-associated variants will inform pathogenicity classifications and lead to enhanced patient diagnoses. All data accumulated during this project are available on Github (https://github.com/SchniedersLab/OtoProtein).

Declarations

Acknowledgements

All computations were performed on The University of Iowa Argon cluster with support and guidance from Glenn Johnson.

Funding

Author MRT was supported by the NSF (National Science Foundation) Graduate Research Fellowship under Grant No. 000390183 and an American Association of University Women American Dissertation Fellowship. Author RAG was supported by NIH R01DC012049, NIH T32 Grant No. 2T32GM008365-26A 1, the Center for Biocatalysis and Bioprocessing (CBB), and the Graduate College Iowa Recruitment Fellowship. Author MJS was supported by NIH R01DC012049 and NSF CHE-1751688. Author RJHS was supported by NIDCDs R01's DC002842, DC012049, and DC017955.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Conceptualization of the study was completed by Mallory Tollefson, Rose Gogal, Richard Smith, and Michael Schnieders. Data curation was completed by Mallory Tollefson, Rose Gogal, Monique Weaver, Amanda Schaefer, Robert Marini, Hela Azaiez, Diana Kolbe, Donghon Wang, and Amy Weaver. Data analysis was performed by Mallory Tollefson, Rose Gogal, Monique Weaver, Amanda Schaefer, Richard Smith, and Michael Schnieders. Software was written by Mallory Tollefson, Rose Gogal, Robert Marini, Tom Casavant, Terry Braun and Michael Schnieders. Supervision of the project was done by Tom Casavant, Terry Braun, Richard Smith and Michael Schnieders. Writing of the original manuscript draft was done by Mallory Tollefson, Richard Smith, and Michael Schnieders. Review and editing of the manuscript was completed by Mallory Tollefson, Rose Gogal, Monique Weaver, Amanda Schaefer, Hela Azaiez, Tom Casavant, Terry Braun, Richard Smith and Michael Schnieders. All authors read and approved the final manuscript.

Data and Code Availability

The datasets generated during this study are available at https://github.com/SchniedersLab/OtoProtein.

OtoProtein2 models and folding free energy differences: https://github.com/SchniedersLab/OtoProtein

Force Field X software for protein model optimization:

https://ffx.biochem.uiowa.edu

References

1. Agrahari AK, Sneha P, Doss CGP, Siva R, Zayed H (2018) A profound computational study to prioritize the disease-causing mutations in PRPS1 gene. Metabolic Brain Disease 33: 589-600. doi: 10.1007/s11011-017-0121-2

2. Alford RL, Arnos KS, Fox M, Lin JW, Palmer CG, Pandya A, Rehm HL, Robin NH, Scott DA, Yoshinaga-Itano C, Loss AWGoUoGEGftEDoCH, Professional P, Guidelines C (2014) American College of Medical Genetics and Genomics guideline for the clinical evaluation and etiologic diagnosis of hearing loss. Genet Med 16: 347-55. doi: 10.1038/gim.2014.2
3. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti T, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29: 37-40. doi: DOI 10.1093/nar/29.1.37
4. Araya CL, Fowler DM, Chen WT, Muniez I, Kelly JW, Fields S (2012) A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proceedings of the National Academy of Sciences of the United States of America 109: 16858-16863. doi: 10.1073/pnas.1209751109
5. Azaiez H, Booth KT, Ephraim SS, Crone B, Black-Ziegelbein EA, Marini RJ, Shearer AE, Sloan-Heggen CM, Kolbe D, Casavant T, Schnieders MJ, Nishimura C, Braun T, Smith RJH (2018) Genomic landscape and mutational signatures of deafness-associated genes. Am J Hum Genet 103: 484-497. doi: 10.1016/j.ajhg.2018.08.006
6. Balchin D, Hayer-Hartl M, Hartl FU (2016) In vivo aspects of protein folding and quality control. Science 353. doi: ARTN aac4354 10.1126/science.aac4354
7. Bereshneh AH, Rezaei Z, Jafarinia E, Rajabi F, Ashrafi MR, Tavasoli AR, Garshasbi M (2021) Crystallographic modeling of the PNPT1:c.1453A > G variant as a cause of mitochondrial dysfunction and autosomal recessive deafness; expanding the neuroimaging and clinical features. Mitochondrion 59: 1-7. doi: 10.1016/j.mito.2021.03.012
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28: 235-42. doi: 10.1093/nar/28.1.235
9. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi HY, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD (2021) The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49: D344-D354. doi: 10.1093/nar/gkaa977
10. Bryant P, Pozzati G, Elofsson A (2022) Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13. doi: 10.1038/s41467-022-28865-w
11. Buonfiglio PI, Bruque CD, Lotersztein V, Luce L, Giliberto F, Menazzi S, Francipane L, Paoli B, Goldschmidt E, Elgoyhen AB, Dalamon V (2022) Predicting pathogenicity for novel hearing loss mutations based on genetic and protein structure approaches. Scientific Reports 12. doi: 10.1038/s41598-021-04081-2
12. Chen HC, Maia JDC, Radak BK, Hardy DJ, Cai WS, Chipot C, Tajkhorshid E (2020) Boosting free-energy perturbation calculations with GPU-accelerated NAMD. J Chem Inf Model 60: 5301-5307. doi: 10.1021/acs.jcim.0c00745
13. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta

Crystallographica Section D-Biological Crystallography 66: 12-21. doi: 10.1107/S0907444909042073

14. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 35: W375-W383. doi: 10.1093/nar/gkm216
15. Doss CGP, Rajith B, Garwasis N, Mathew PR, Raju AS, Apoorva K, William D, Sadhana NR, Himani T, Dike IP (2012) Screening of mutations affecting protein stability and dynamics of FGFR1-A simulation analysis. Applied and Translational Genomics 1: 37-43. doi: 10.1016/j.atg.2012.06.002
16. Duan JX, Lupyan D, Wang LL (2020) Improving the accuracy of protein thermostability predictions for single point mutations. Biophys J 119: 115-127. doi: 10.1016/j.bpj.2020.05.020
17. Gallant E, Francey L, Fetting H, Kaur M, Hakonarson H, Clark D, Devoto M, Krantz ID (2013) Novel COCH mutation in a family with autosomal dominant late onset sensorineural hearing impairment and tinnitus. Am J Otolaryngol 34: 230-5. doi: 10.1016/j.amjoto.2012.11.002
18. Gao M, An DN, Parks JM, Skolnick J (2022) AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. Nat Commun 13. doi: 10.1038/s41467-022-29394-2
19. Goldberg AL (2003) Protein degradation and protection against misfolded or damaged proteins. Nature 426: 895-899. doi: 10.1038/nature02263
20. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology 320: 369-387. doi: 10.1016/S0022-2836(02)00442-4
21. Hilgert N, Alasti F, Dieltjens N, Pawlik B, Wollnik B, Uyguner O, Delmaghani S, Weil D, Petit C, Danis E, Yang T, Pandelia E, Petersen MB, Goossens D, Favero JD, Sanati MH, Smith RJH, Van Camp G (2008) Mutation analysis of TMC1 identifies four new mutations and suggests an additional deafness gene at loci DFNA36 and DFNB7/11. Clinical Genetics 74: 223-232. doi: 10.1111/j.1399-0004.2008.01053.x
22. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596: 583-589. doi: 10.1038/s41586-021-03819-2
23. Li MM, Abou Tayoun A, DiStefano M, Pandya A, Rehm HL, Robin NH, Schaefer AM, Yoshinaga-Itano C, Guideli APP (2022) Clinical evaluation and etiologic diagnosis of hearing loss: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). Genet Med 24: 1392-1406. doi: 10.1016/j.gim.2022.03.018
24. Liming BJ, Carter J, Cheng A, Choo D, Curotta J, Carvalho D, Germiller JA, Hone S, Kenna MA, Loundon N, Preciado D, Schilder A, Reilly BJ, Roman S, Strychowsky J, Triglia JM, Young N, Smith RJ (2016) International Pediatric Otolaryngology Group (IPOG) consensus recommendations: hearing loss in the pediatric patient. Int J Pediatr Otorhinolaryngol 90: 251-258. doi: 10.1016/j.ijporl.2016.09.016
25. McCafferty CL, Sergeev YV (2016) In silico mapping of protein unfolding mutations for inherited disease. Scientific Reports 6. doi: ARTN 37298 10.1038/srep37298
26. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022) ColabFold - Making protein folding accessible to all. bioRxiv: 2021.08.15.456425. doi: 10.1101/2021.08.15.456425

27. Montanucci L, Capriotti E, Birolo G, Benevenuta S, Pancotti C, Lal D, Fariselli P (2022) DDGun: an untrained predictor of protein stability changes upon amino acid variants. Nucleic Acids Res 50: W222-W227. doi: 10.1093/nar/gkac325
28. Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P (2019) DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinform 20: 335. doi: 10.1186/s12859-019-2923-1
29. Parthiban V, Gromiha MM, Schomburg D (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34: W239-W242. doi: 10.1093/nar/gkl190
30. Ponder JW, Wu C, Ren P, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RA, Jr., Head-Gordon M, Clark GN, Johnson ME, Head-Gordon T (2010) Current status of the AMOEBA polarizable force field. J Phys Chem B 114: 2549-64. doi: 10.1021/jp910674d
31. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2018) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. doi: 10.1093/nar/gky1016
32. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, Committee ALQA (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17: 405-24. doi: 10.1038/gim.2015.30
33. Rodrigues CHM, Pires DEV, Ascher DB (2021) DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Science 30: 60-69. doi: 10.1002/pro.3942
34. Shearer AE, Black-Ziegelbein EA, Hildebrand MS, Eppsteiner RW, Ravi H, Joshi S, Guiffre AC, Sloan CM, Happe S, Howard SD, Novak B, DeLuca AP, Taylor KR, Scheetz TE, Braun TA, Casavant TL, Kimberling WJ, LeProust EM, Smith RJH (2013) Advancing genetic testing for deafness with genomic technology. Journal of Medical Genetics 50: 627-634. doi: 10.1136/jmedgenet-2013-101749
35. Shearer AE, DeLuca AP, Hildebrand MS, Taylor KR, Gurrola J, Scherer S, Scheetz TE, Smith RJH (2010) Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America 107: 21104-21109. doi: 10.1073/pnas.1012989107
36. Shearer AE, Eppsteiner RW, Booth KT, Ephraim SS, Gurrola J, 2nd, Simpson A, Black-Ziegelbein EA, Joshi S, Ravi H, Giuffre AC, Happe S, Hildebrand MS, Azaiez H, Bayazit YA, Erdal ME, Lopez-Escamez JA, Gazquez I, Tamayo ML, Gelvez NY, Leal GL, Jalas C, Ekstein J, Yang T, Usami S, Kahrizi K, Bazazzadegan N, Najmabadi H, Scheetz TE, Braun TA, Casavant TL, LeProust EM, Smith RJ (2014) Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am J Hum Genet 95: 445-53. doi: 10.1016/j.ajhg.2014.09.001
37. Shearer AE, Smith RJ (2015) Massively parallel sequencing for genetic diagnosis of hearing loss: the new standard of care. Otolaryngol Head Neck Surg 153: 175-82. doi: 10.1177/0194599815591156
38. Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, Ren P (2013) The polarizable atomic multipole-based AMOEBA force field for proteins. J Chem Theory Comput 9: 4046-4063. doi: 10.1021/ct4003702
39. Sloan-Heggen CM, Bierer AO, Shearer AE, Kolbe DL, Nishimura CJ, Frees KL, Ephraim SS, Shibata SB, Booth KT, Campbell CA, Ranum PT, Weaver AE, Black-Ziegelbein EA, Wang D, Azaiez H, Smith RJH (2016)

Comprehensive genetic testing in the clinical evaluation of 1119 patients with hearing loss. Hum Genet 135: 441-450. doi: 10.1007/s00439-016-1648-8

40. Sloan-Heggen CM, Smith RJH (2016) Navigating genetic diagnostics in patients with hearing loss. Curr Opin Pediatr 28: 705-712. doi: 10.1097/Mop.0000000000000410
41. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E (2013) Molecular mechanisms of disease-causing missense mutations. Journal of Molecular Biology 425: 3919-3936. doi: DOI 10.1016/j.jmb.2013.07.014
42. Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K (2019) Biophysical and mechanistic models for disease-causing protein variants. Trends in Biochemical Sciences 44: 575-588. doi: 10.1016/j.tibs.2019.01.003
43. Takano K, Liu D, Tarpey P, Gallant E, Lam A, Witham S, Alexov E, Chaubey A, Stevenson RE, Schwartz CE, Board PG, Dulhunty AF (2012) An X-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity. Hum Mol Genet 21: 4497-4507. doi: 10.1093/hmg/dds292
44. Talley K, Alexov E (2010) On the pH-optimum of activity and stability of proteins. Proteins 78: 2699-2706. doi: 10.1002/prot.22786
45. Teng SL, Madej T, Panchenko A, Alexov E (2009) Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. Biophys J 96: 2178-2188. doi: 10.1016/j.bpj.2008.12.3904
46. Tollefson MR, Litman JM, Qi G, O'Connell CE, Wipfler MJ, Marini RJ, Bernabe HV, Tollefson WTA, Braun TA, Casavant TL, Smith RJH, Schnieders MJ (2019) Structural insights into hearing loss genetics from polarizable protein repacking. Biophys J 117: 602-612. doi: 10.1016/j.bpj.2019.06.030
47. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D (2021) Highly accurate protein structure prediction for the human proteome. Nature 596: 590-596. doi: 10.1038/s41586-021-03828-1
48. Witham S, Takano K, Schwartz C, Alexov E (2011) A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. Proteins 79: 2444-2454. doi: 10.1002/prot.23065
49. Yin R, Feng BY, Varshney A, Pierce BG (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Science 31. doi: 10.1002/pro.4379
50. Zhang Z, Norris J, Schwartz C, Alexov E (2011) In silico and in vitro investigations of the mutability of disease-causing missense mutation sites in spermine synthase. Plos One 6. doi: 10.1371/journal.pone.0020373
51. Zhang Z, Teng SL, Wang LJ, Schwartz CE, Alexov E (2010) Computational analysis of missense mutations causing Snyder-Robinson syndrome. Hum Mutat 31: 1043-1049. doi: 10.1002/humu.21310
52. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714-26. doi: 10.1110/ps.0217002

Figures

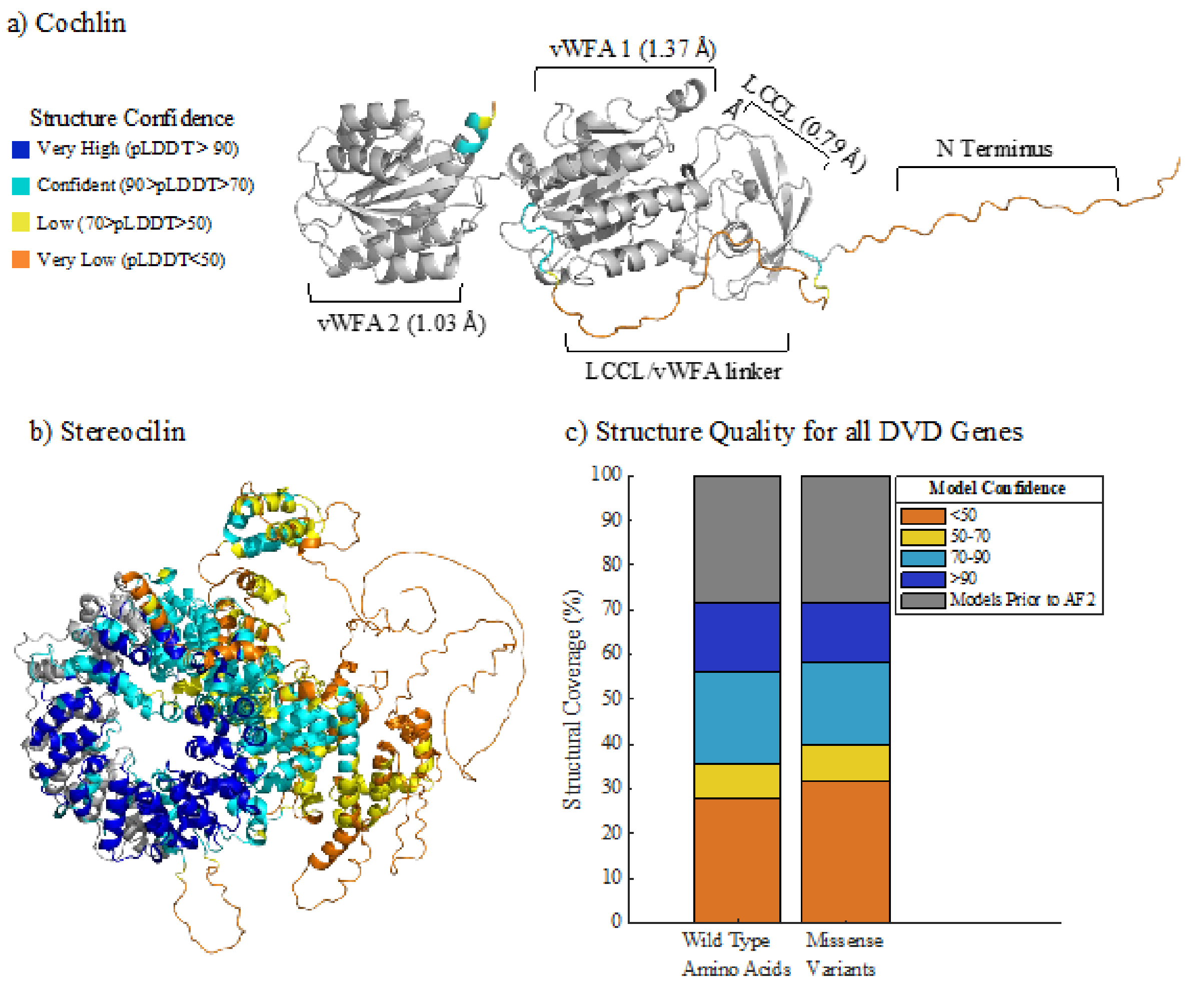

Figure 1

Structures and quality of proteins implicated in deafness. AlphaFold2's novel predicted protein regions are color coded by confidence in the prediction. Gray domains represent homology or experimental structures curated in prior work for a) cochlin and b) stereocilin. a) The root-mean-square deviation (RMSD) of the LCCL and vWFA domains of cochlin (*COCH*) from AlphaFold2's domain predictions to the previous models are shown in parentheses. b) This work increased protein structural coverage of stereocilin (*STRC*) from 12% to 100%. c) Structural model coverage of wild-type amino acids and missense variants for the entire deafness proteome shows that this work increased coverage from <30% (gray, prior work) to 100% coverage. The stacked bars are color coded based on confidence in the protein structure. The wild-type amino acids and missense variants in the deafness proteome are present in similar proportions across all structural confidence ranges, indicating that specific confidence regions are not enriched for the presence of missense variants

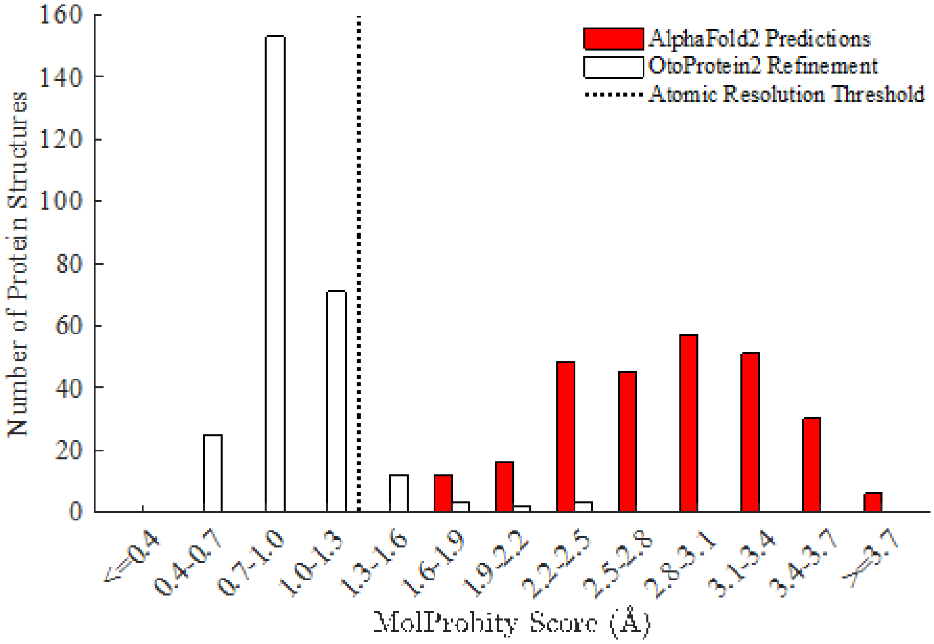

Figure 2

MolProbity score histogram for the OtoProtein2 database. Before optimization (red), the mean MolProbity score of the models is 2.86 and after optimization (blue) the structures are consistent with atomic resolution at a mean MolProbity score of 0.97. MolProbity scores are calibrated to reflect the expected crystallographic resolution of the diffraction dataset employed to create a protein structural model (*i.e.*, a MolProbity score of 1.0 indicates that the structure is consistent with 1.0 Å resolution X-ray diffraction data)

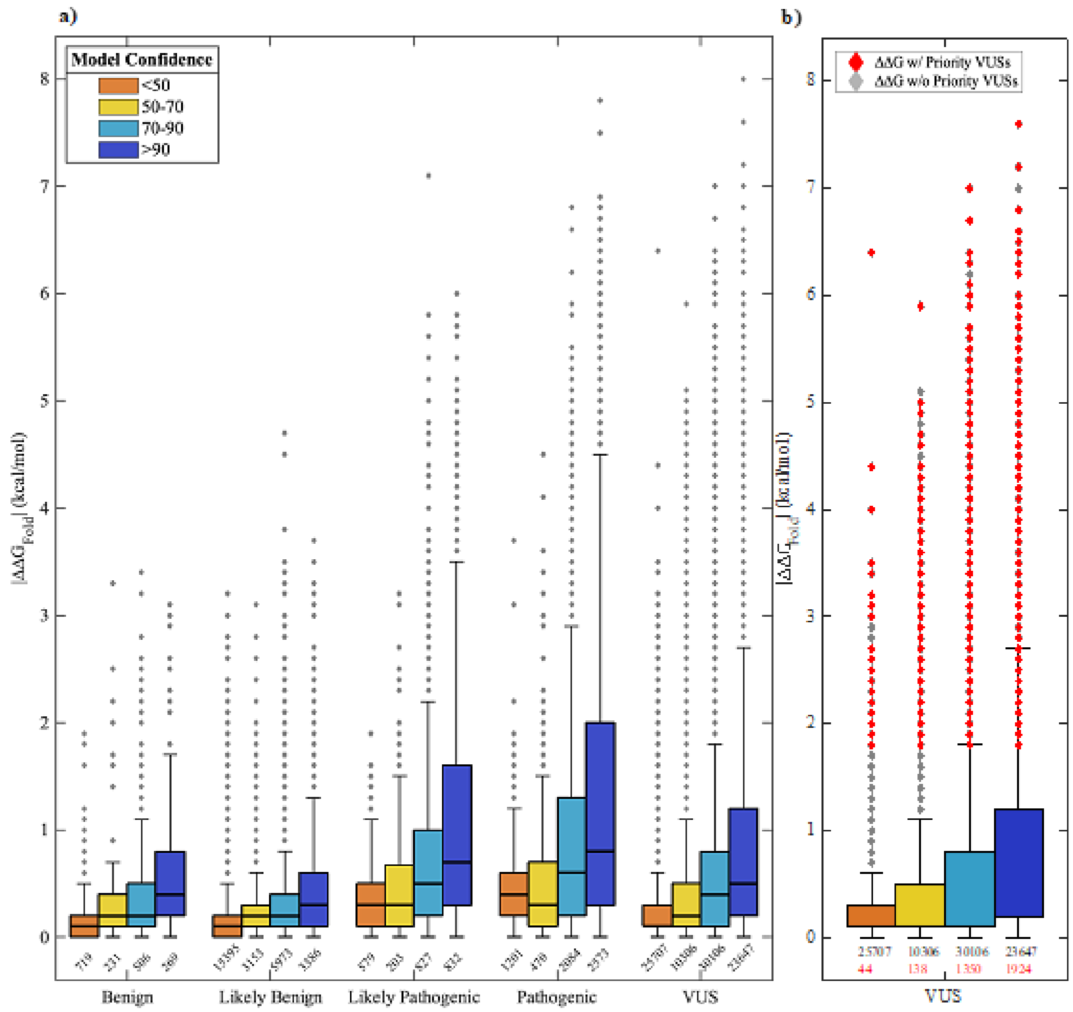

Figure 3

The range of $\Delta\Delta G_{\mathrm{Fold}}$ predictions for missense variants in the Deafness Variation Database (DVD). a) Box plots are grouped based on DVD pathogenicity classification and bars are colored based on the structure confidence at the variant's amino acid position. Pathogenic variants and variants in confident portions of protein models have a larger distribution of $\Delta\Delta G_{\mathrm{Fold}}$ than the benign and low confidence (*e.g.*, usually solvent exposed) counterparts. The number of observations belonging to each box is printed below the box. b) A box plot for all VUSs in the DVD. Each outlier in the boxplot can represent multiple VUSs due to overlap in $\Delta\Delta G_{\mathrm{Fold}}$. Outliers colored in red are prioritized VUSs that have a large $\Delta\Delta G_{\mathrm{Fold}}$ ($\geq$1.8 kcal/mol) and a high CADD score (>25.7). Unprioritized VUSs do not have a high CADD score. The number of prioritized VUSs belonging to each box is printed in red below the total number of observations belonging to the box

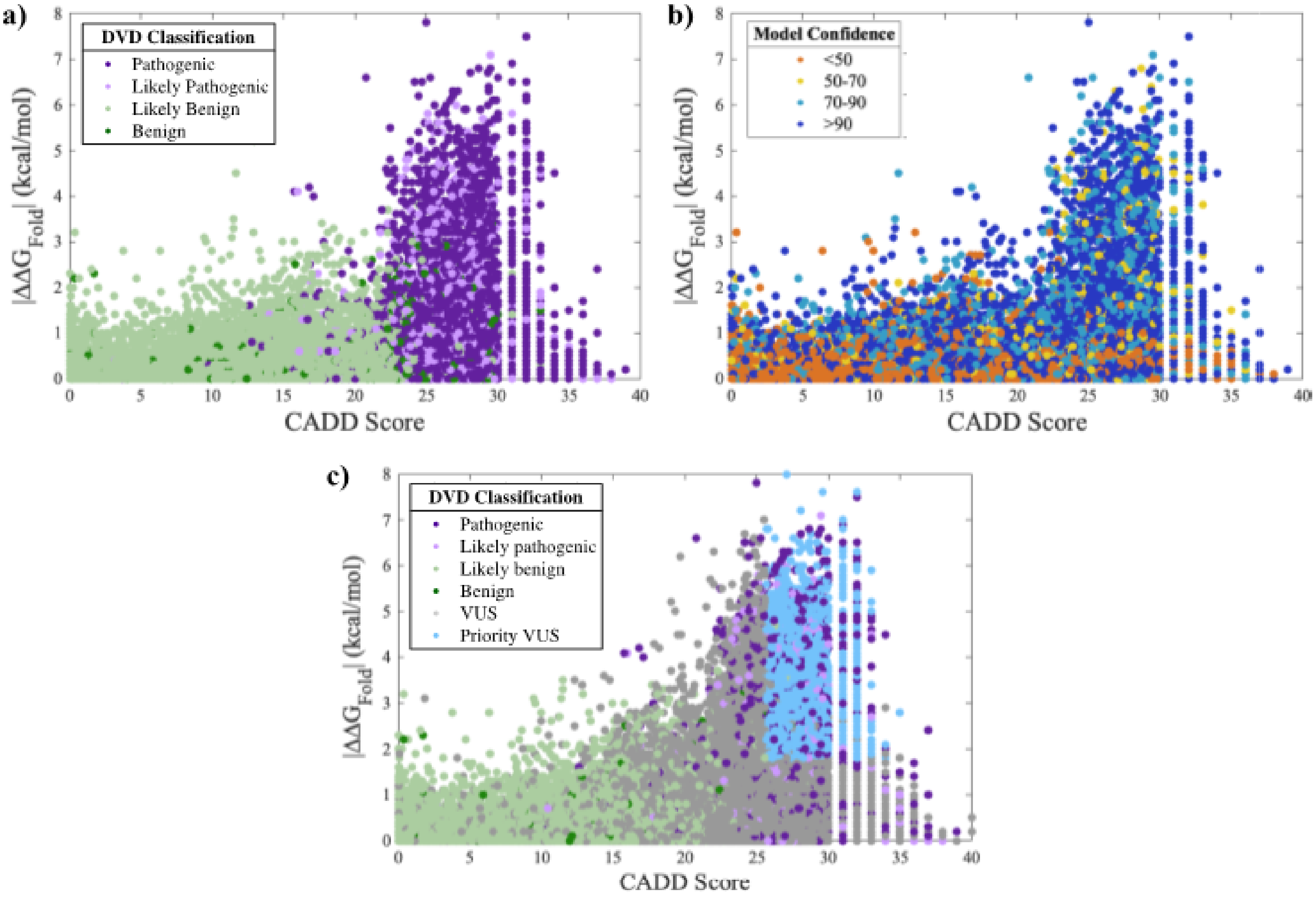

Figure 4

Prioritizing variants of uncertain significance (VUSs) from thermodynamic data and CADD scores. Folding free energy differences ($\Delta\Delta G_{Fold}$) versus CADD score for all missense variants observed in the Deafness Variation Database (DVD) with points colored according to DVD classification (panels a and c) or model confidence at the variant's amino acid position (panel b). CADD score and $\Delta\Delta G_{Fold}$ show a positive correlation. A high $\Delta\Delta G_{Fold}$ and high CADD score in confident regions of a protein model favor pathogenic variants; low $\Delta\Delta G_{Fold}$ and low CADD score favor benign variants and exhibit greater variety in model confidence. Prioritized VUSs have both high $\Delta\Delta G_{Fold}$ and high CADD scores

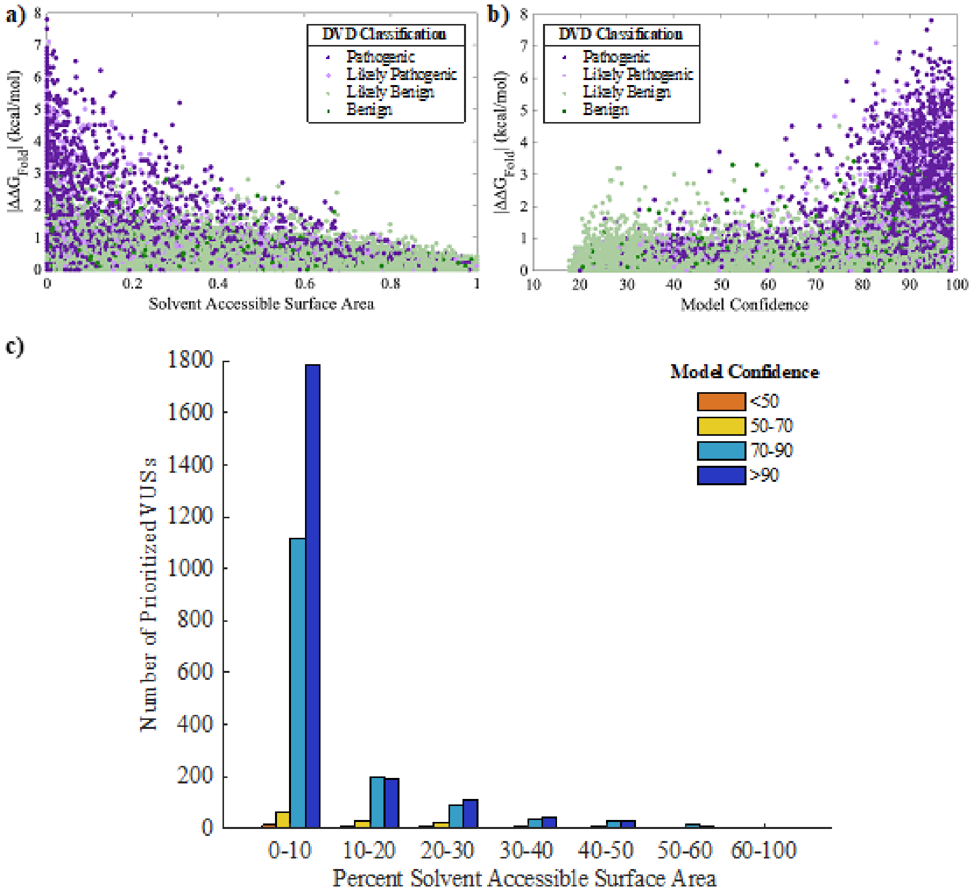

Figure 5

Protein features for prioritizing VUSs. Folding free energy differences ($\Delta\Delta G_{Fold}$) versus a) percent of solvent accessible surface area (SASA) at a variant's amino acid position, and b) model confidence at the variant position for all classified missense variants in the Deafness Variation Database (DVD). Pathogenic and likely pathogenic variants favor buried, high confidence protein regions. c) A histogram of the percent SASA for all prioritized VUSs. Similar to known P and LP, the prioritized VUSs are mostly in buried, high confidence protein regions

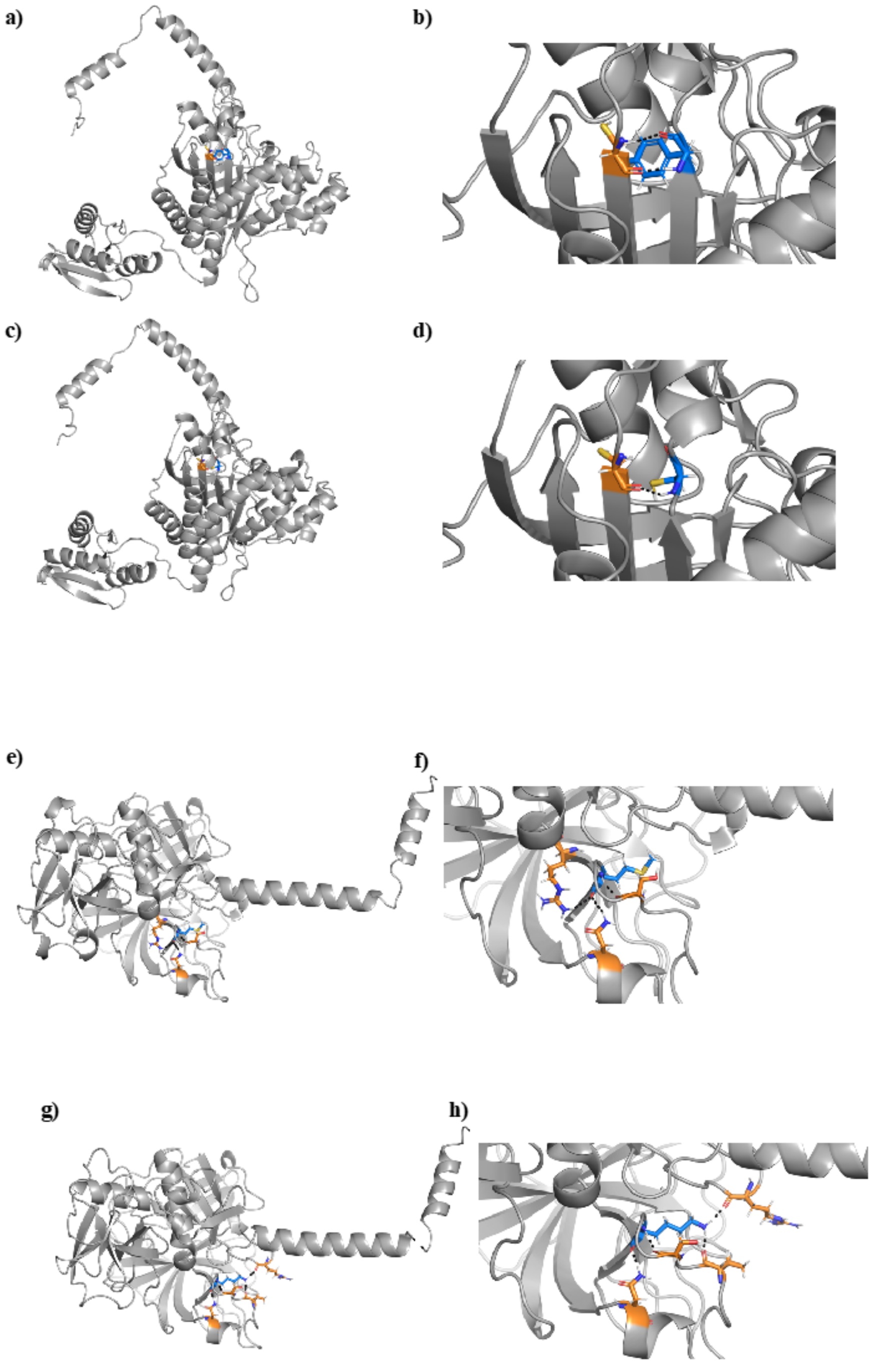

Figure 6

The protein structure of *HARS2* variant NP_036340.1:p.Tyr364Cys (Panels A-D) and *TMPRSS3* variant NP_076927.1:p.Met384Lys (Panels E-H). A) The wildtype *HARS2* protein contains a tyrosine (blue) at position 364, which interacts with a neighboring cysteine amino acid (orange) B) Augmentation of the boxed region in Panel A shows two hydrogen bonds between the tyrosine and cysteine. C) The NP_036340.1:p.Tyr364Cys variant introduces a new cysteine (blue) in place of tyrosine. D) Enlargement of the boxed region from Panel C shows that the variant cysteine (blue) interacts with the original neighboring cysteine (orange), disrupting the two hydrogen bonds to form a single hydrogen bond or a disulfide bond. E) The wildtype *TMPRSS3* protein shows a methionine (blue) at position 384, which interacts with three neighboring amino acids (orange). F) Magnification of Panel E shows three hydrogen bonds between the methionine and neighboring amino acids. G) The NP_076927.1:p.Met384Lys variant introduces a lysine (blue) in place of methionine, which interacts with four neighboring amino acids, only one of which remains the same as the wildtype interacting neighbors. H) Enlargement of the boxed region from Panel G shows four hydrogen bonds between the lysine and neighboring amino acids. While one hydrogen bond remains the same between the wildtype and variant structures, the NP_076927.1:p.Met384Lys variant results in significant misfolding

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInfo.docx
- TableS3.xlsx
- TableS4.xlsx