npj | genomic medicine

Article

Published in partnership with CEGMR, King Abdulaziz University

https://doi.org/10.1038/s41525-024-00416-w

Structure-based network analysis predicts pathogenic variants in human proteins associated with inherited retinal disease

Blake M. Hauser[1], Yuyang Luo[2], Anusha Nathan[3], Ahmad Al-Moujahed[2], Demetrios G. Vavvas[2], Jason Comander[2], Eric A. Pierce[2], Emily M. Place[2], Kinga M. Bujakowska[2], Gaurav D. Gaiha[3,4] & Elizabeth J. Rossin[2]

Advances in gene sequencing technologies have accelerated the identification of genetic variants, but better tools are needed to understand which are causal of disease. This would be particularly useful in fields where gene therapy is a potential therapeutic modality for a disease-causing variant such as inherited retinal disease (IRD). Here, we apply structure-based network analysis (SBNA), which has been successfully utilized to identify variant-constrained amino acid residues in viral proteins, to identify residues that may cause IRD if subject to missense mutation. SBNA is based entirely on structural first principles and is not fit to specific outcome data, which makes it distinct from other contemporary missense prediction tools. In 4 well-studied human disease-associated proteins (BRCA1, HRAS, PTEN, and ERK2) with high-quality structural data, we find that SBNA scores correlate strongly with deep mutagenesis data. When applied to 47 IRD genes with available high-quality crystal structure data, SBNA scores reliably identified disease-causing variants according to phenotype definitions from the ClinVar database. Finally, we applied this approach to 63 patients at Massachusetts Eye and Ear (MEE) with IRD but for whom no genetic cause had been identified. Untrained models built using SBNA scores and BLOSUM62 scores for IRD-associated genes successfully predicted the pathogenicity of novel variants (AUC = 0.851), allowing us to identify likely causative disease variants in 40 IRD patients. Model performance was further augmented by incorporating orthogonal data from EVE scores (AUC = 0.927), which are based on evolutionary multiple sequence alignments. In conclusion, SBNA can used to successfully identify variants as causal of disease in human proteins and may help predict variants causative of IRD in an unbiased fashion.

As genetic sequencing has become increasingly available and less costly, a growing number of patients with clinical presentations of suspected genetic origin are undergoing targeted or whole-exome sequencing. Despite improved accessibility, the genetic basis of disease for a considerable proportion of these patients remains unclear following sequencing[1]. Inherited retinal diseases (IRD), whereby rod and cone photoreceptors degenerate, are a group of Mendelian disorders that represent an important cause of vision loss[2]. With the advancement in the availability of genetic testing[3–6] and the lower cost of exome sequencing, IRD is a field with increasing promise and possibility for gene therapy interventions. However, among patients with an IRD, ~30% do not have a clear genetic basis despite classic retinal changes and a decrease in visual and retinal function[7], making them ineligible candidates for treatment. For these patients, additional tools are needed to better define genetic variants that are not among the group of known causal variants (i.e., variants of uncertain significance—VUS).

[1]Harvard Medical School, Boston, MA, USA. [2]Harvard Medical School, Department of Ophthalmology, Massachusetts Eye and Ear, Boston, MA, USA. [3]Ragon Institute of Mass General, MIT, and Harvard, Cambridge, MA, USA. [4]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA.
e-mail: elizabeth_rossin@meei.harvard.edu

Numerous computational tools that aim to predict the phenotype of genetic variants have been described[8–10], many of which were trained on existing variant classifications or combine multiple metrics that use this type of training data[11–19]. These data were limited by sparsely available annotations[20], and previous studies suggest that some of these algorithms may also have considerable false discovery rates[21,22]. Furthermore, many of these algorithms are limited by circularity, with duplication of variants in the training and test datasets as well as variants from the same protein within the training and test datasets[23]. The clinical applicability of these approaches has been limited as a result[23,24]. Another class of methods uses sequence conservation to estimate the likelihood that a particular variant will have a deleterious phenotypic effect[25,26]. Sequence conservation is an important feature to consider (and provides independent information from protein structure) but can be an imperfect proxy for the functional importance of a particular amino acid position within a protein of interest; this has been previously demonstrated within the context of both model and human proteins[27,28] and also the human immunodeficiency virus (HIV), which has a per-base mutation rate ~10^4 times that of the human genome and thus serves as a model for an accelerated rate of genomic evolution[29–31]. Some of these approaches also leverage recent advances in machine learning, which results in limited post hoc transparency regarding the basis for a particular estimated probability of pathogenicity and is also subject to circularity and overfitting[26,32]. Functional in vitro assays, such as a high throughput assay for the *RHO* gene, can help to characterize individual genetic variants[33]. However, these crucial experiments are not always feasible on a large scale for the entire set of patient-genotype combinations, given the need for appropriate experimental setup, tissue-type, and readout for the particular variant of interest.

To add to these approaches, we developed structure-based network analysis (SBNA) which leverages the application of network theory to protein structure data with the goal of quantifying local residue connectivity, bridging interactions, and ligand proximity in order to identify amino acid residues that are topologically important[29,34]. Using x-ray crystallography and cryogenic electron microscopy (cryo-EM) data, it models proteins as networks of connected amino acids to quantitatively estimate the topological importance of each amino acid as it relates to others in the protein, protein complex, or protein–ligand interaction. This approach is distinct from prior computational tools that use structural information because it is not reliant on pre-defined secondary structure elements; rather, it analyzes the crystallized tertiary structure of the folded protein as a network of weighted inter-residue interactions. Additionally, this approach does not require training on pre-labeled phenotypic data which means that it can provide a metric that is specifically focused on first-principle structural information. SBNA has been previously used to identify highly mutationally constrained amino acid residues and $CD8^+$ T cell epitopes in HIV and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[29,35]. In the case of HIV, these epitopes were found to be preferentially targeted by individuals able to suppress HIV replication in the absence of antiretroviral therapy[29]. Within the context of SARS-CoV-2, highly networked regions have resisted ongoing sequence evolution during the pandemic and thereby may be capable of conferring broad T cell-mediated protection against sarbecovirus infection[35].

In light of these successes with highly mutable viruses, we chose to apply SBNA to human proteins which exist within a complex system of biological interactions. Such an application would potentially aid in better understanding which genetic variants discovered through sequencing are causal of disease. We first verified this approach in well-studied human proteins: breast cancer gene 1 (BRCA1), GTPase HRas (HRAS), phosphatase and tensin homolog (PTEN), and mitogen-activated protein kinase 1 (ERK2). We then sought to investigate whether this approach could generate meaningful results for IRD proteins of interest to estimate the phenotypic impact of missense variants, given that such a sizable number of patients with IRDs harbor VUSs. Therefore, we assessed the performance of SBNA in IRD protein-coding genes and their respective missense variants described in ClinVar, a large public database of reported pathogenic and benign genetic variation. Finally, we addressed a cohort of 63 affected subjects for whom a genetic cause of their condition has yet to be discovered to investigate the use of SBNA in real-world clinical scenarios.

Results

Structure-based network analysis accurately identifies missense variants in human proteins associated with in vitro loss of function and pathogenic clinical phenotypes

To evaluate whether SBNA could be meaningfully applied to human proteins, we first applied the approach to four well-studied human proteins—BRCA1, HRAS, PTEN, and ERK2—which were selected for analysis due to the availability of high-quality structural data. In addition, these proteins had published in vitro saturation mutagenesis experiments, which allowed us to extract the functional consequence of all missense variants and quantify mutational tolerance[36–39]. We next generated network scores for all amino acid residues in the available protein databank (PDB) files (Fig. 1a) and evaluated the correlation with mutational tolerance (Fig. 1b). All four proteins showed a strong inverse correlation between mutational tolerance and network score, which was consistent with previous findings for viral proteins and other model proteins[29]. Of note, only 17% of BRCA1 has been crystallized (the N- and C-terminal ends), but SBNA scores still performed reasonably well (Spearman correlation coefficient −0.514, $p = 3.45e\text{-}22$) despite limited structural data.

We next compared network scores to pathogenicity categorizations derived from human data using the ClinVar and gnomAD genetic databases for these same four proteins (Fig. 1c). Missense variants in all subsequent analyses were categorized with respect to human clinical data in line with the American College of Medical Genetics and Genomics/Association of Molecular Pathology (ACMG/AMP)[24] as benign (encompassing "benign" or "likely benign" within ClinVar), VUS or pathogenic (encompassing "pathogenic" or "likely pathogenic" within ClinVar). We restricted our analysis to ClinVar missense variants with at least two-star level evidence, and gnomAD was used to identify additional relatively benign missense variants (variants with at least 250 alternative allele counts across 100,000 individuals). Across all four proteins, network scores assigned to pathogenic variants were significantly greater than those assigned to benign variants (median network scores for benign missense variants −0.936 and pathogenic variants 0.866, $p = 1.836e\text{-}5$, Fig. 1c). Scores assigned to VUSs fell in between those assigned to benign and pathogenic variants. These trends were observed consistently within individual proteins, though smaller sample sizes limit statistical power (Fig. 1d). Overall, network scores correlate with available clinical phenotype data for the four well-studied human proteins (Spearman correlation coefficient 0.228, $p = 2.116e\text{-}23$), suggesting that SBNA can be meaningfully applied to human proteins. Of note, we found that SBNA seems to capture structural properties beyond simple solvent accessibility (i.e., proximity to the core) because relative solvent accessibility (RSA) scores show less correlation with mutational tolerance scores than did SBNA (Supplementary Fig. 1).

SBNA predicts variants in IRD genes associated with pathogenic clinical phenotypes

Having validated SBNA on four canonical, well-studied human proteins, we then applied SBNA to additional human proteins. We analyzed the relationship between network scores and pathogenicity designations from high-quality ClinVar variants and benign gnomAD variants for the 47 human genes associated with IRDs, for which high-quality structural data is available for the encoded protein. This dataset includes both membrane proteins (e.g., ABCA4 and RHO) as well as cytoplasmic proteins (e.g., RPE65 and RPGR) (Supplementary Table 1). We found that pathogenic variants were assigned significantly greater network scores compared to both benign variants and VUSs (median benign −0.841, median VUS −0.188, median pathogenic 0.947 $p = 3.140e\text{-}29$ for benign vs. pathogenic, $p = 1.753e\text{-}60$ for VUS vs. pathogenic; Fig. 2a and Supplementary Fig. 2), which is similar to the pattern observed for the canonical human proteins. Because the number of high-quality ClinVar entries for missense variants in each of the 47 IRD-

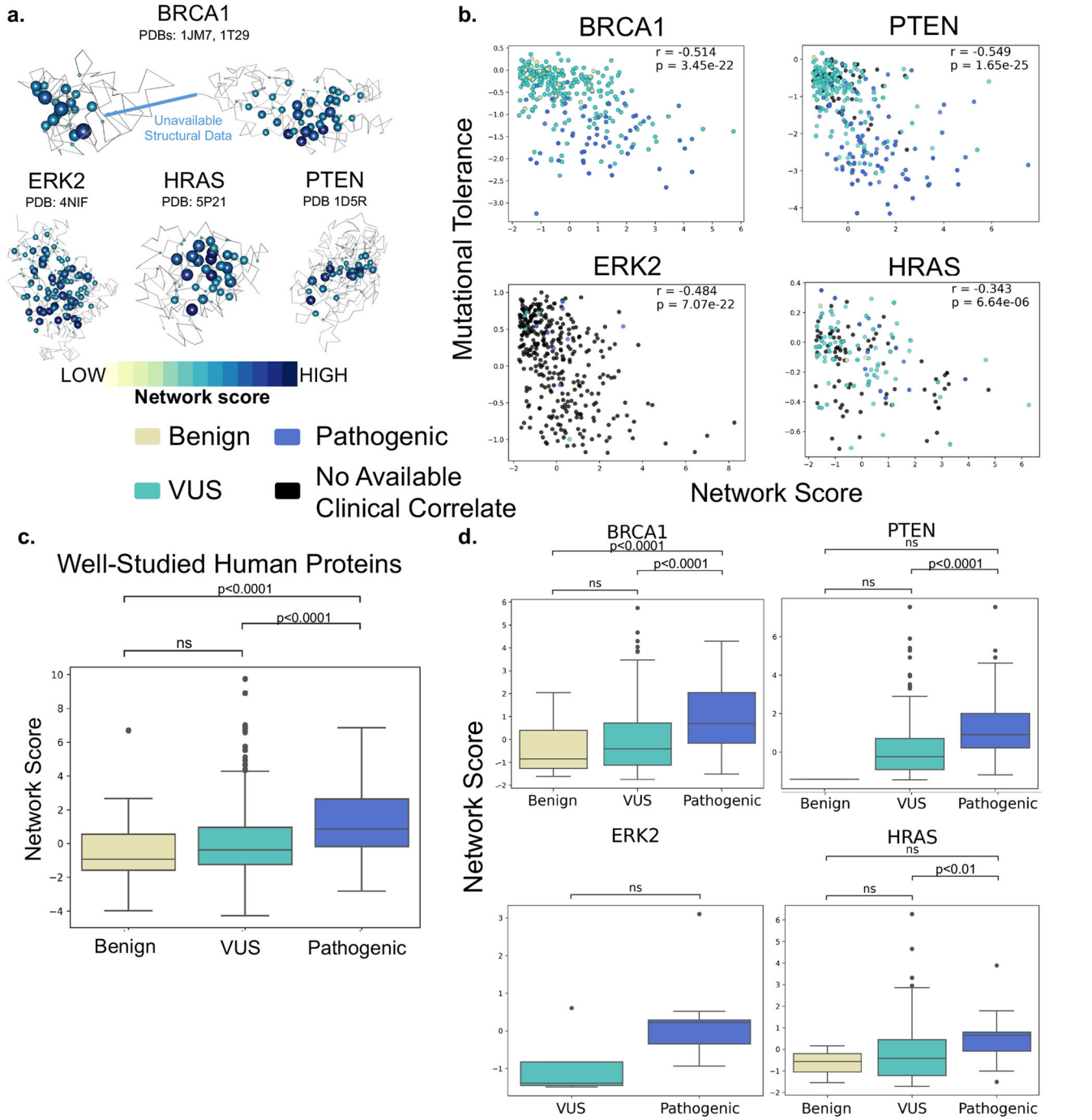


Fig. 1 | Residue network score correlates with mutation intolerance and distinguishes pathogenic variants from benign variants in well-studied human proteins. a Structural representations show network scores at each residue. Sphere radius and color corresponds to network score magnitude at a particular position. **b** Comparison between functional data from saturation mutagenesis experiments and network scores, with Spearman correlation coefficients and p-values displayed for each plot. Points are colored based on available clinical phenotype data. **c** Pooled comparison between network scores for variants with available clinical phenotype data for all four well-studied human proteins. **d** Individual comparisons between network scores for variants with available clinical phenotype data for all four well-studied human proteins.

associated proteins varies considerably (Fig. 2b), we wanted to evaluate whether the difference between median benign and pathogenic network scores remained statistically significant even for proteins with limited variant data available in ClinVar. We grouped proteins by the amount of available high-quality entries in ClinVar and calculated the difference between the median benign and pathogenic network scores across each group of proteins. The magnitude of these differences was robust in the setting of differing levels of available clinical data across genes and was detectable down to 40 high-quality entries per gene (Fig. 2c).

Incorporating network scores and BLOSUM62 scores successfully predicts variant pathogenicity

To move from aggregate statistics to prediction of pathogenicity using network scores, we constructed a modified score that incorporates not only the SBNA network score but also the degree of chemical and side chain dissimilarity between the reference and mutant amino acid at that position (since missense variants vary in this regard). To capture the latter effect, we subtracted the BLOSUM62 matrix score from the SBNA score (which we will now refer to as the modified SBNA score) to allow for a distinction

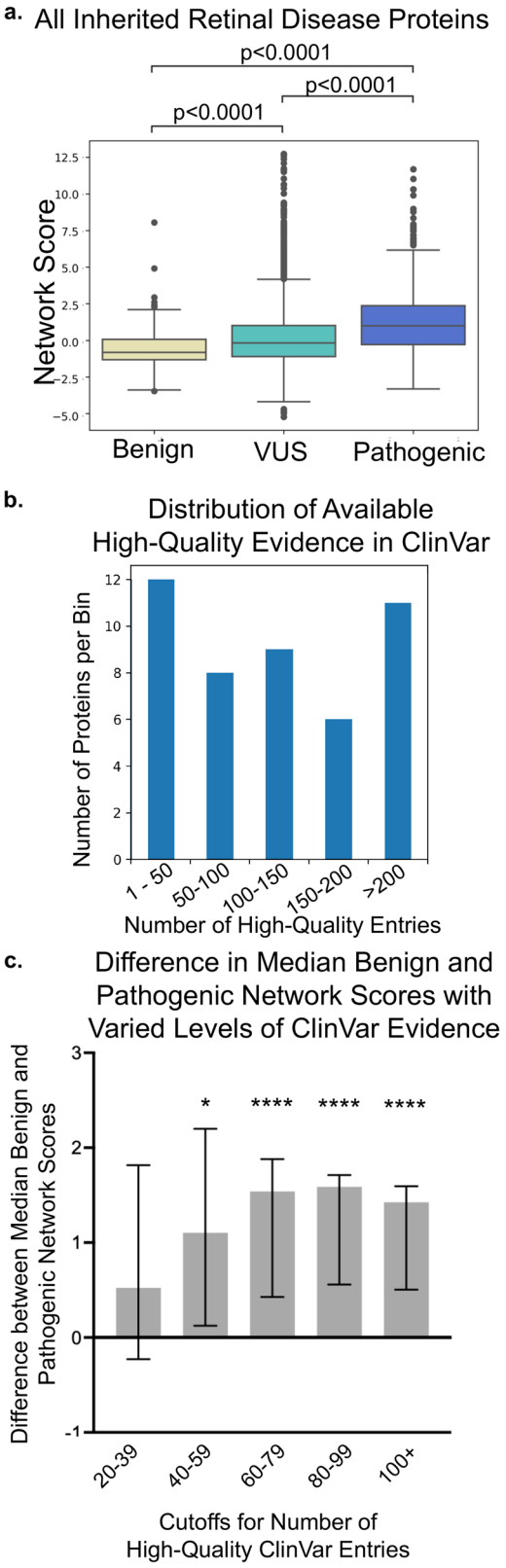


Fig. 2 | Structure-based network analysis identifies pathogenic variants in inherited retinal disease proteins. a Pooled comparison between network scores for variants with available clinical phenotype data for all 47 inherited retinal disease proteins. **b** Distribution of available high-quality evidence in ClinVar across all 47 inherited retinal disease proteins. The bins reflect increasing numbers of high-quality entries in ClinVar, and the height of each bar reflects the number of proteins in each category. **c** Comparison between median benign and pathogenic network scores assigned to variants with available high-quality evidence in ClinVar, grouped by level of available high-quality evidence for each inherited retinal disease protein. Stars above columns indicate statistical significance (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ***$p < 0.0001$). The statistical significance of the difference in network scores between benign and pathogenic variants is lost between 20 and 40 high-quality ClinVar evidence entries.

between non-conservative substitutions (e.g., ILE → TRP) and conservative ones (e.g., ILE → LEU)[40]. BLOSUM62 scores for non-synonymous substitutions range from −4 to 3, while 95% of the raw SBNA scores range from −2.88 to 4.97; thus, the two metrics are on similar scales and can be combined with simple subtraction. We then calculated receiver operating characteristic (ROC) curve statistics for high-quality ClinVar variants and benign gnomAD variants based on the modified SBNA score. Modified SBNA scores predicted variant pathogenicity (AUC 0.851) and outperformed network scores alone, BLOSUM62 scores alone, and RSA scores alone (Fig. 3a, b and Supplementary Fig. 3). We tested multivariable logistic regression modeling with 500 iterations of a 70/30% train-test split as well as a leave-one-out approach using the labels derived from high-quality ClinVar and gnomAD variants and found similar performance to the raw scores (AUC 0.835, Supplementary Figs. 3, 4). Given the advantage that using the raw scores has over a trained approach (which can be subject to poor phenotypic labeling and data circularity[20,23]), all downstream clinical applications described here use the raw modified SBNA score.

Comparison of SBNA to existing methods reveals similar performance without dependence on phenotype labels or evolutionary sequence data

We set out to compare SBNA to three tools that are built from different underlying data: Polyphen2, AlphaMissense, and EVE. Polyphen2 is a widely used computational prediction tool for variant pathogenicity[41]. It is important to note however that PolyPhen2 is used by the ACMG/AMP to assess pathogenicity designations in databases such as ClinVar and HumDiv, so using ClinVar as ground truth might overestimate the performance of PolyPhen2. AlphaMissense[32] incorporates structural data from AlphaFold2[42], protein language modeling, and evolutionary multiple sequence alignments into a machine-learning model using ClinVar labels to train. Similar to PolyPhen2, there is a risk of overfitting and circularity with AlphaMissense. EVE is a variant pathogenicity approach which relies only on evolutionary sequence data rather than clinical labeling for model training and is thus not subject to circularity, similar to SBNA[26]. EVE performs well on two of the well-studied human proteins (Spearman correlation ranging from −0.478 to −0.513, benign versus pathogenic $P < 0.05$ for all; Supplementary Fig. 5); EVE predictions are not available for ERK2. To minimize bias, all algorithms were tested on an independent dataset of 2800 rare variants derived from patients with an IRD who were seen at MEE, though of note, ground truth is still determined using the ClinVar database in accordance with the clinical standard in the field. Thus, methods that train on ClinVar must be interpreted with caution. The modified SBNA scores were compared to results generated using PolyPhen2 trained on two different datasets, HumDiv and HumVar[12,41], as well as EVE scores[26]. All methods showed a significant difference between benign and pathogenic variants, and the modified SBNA scores correlated with the scores from other methods (Supplementary Figs. 6A, B, 7A, B, Spearman correlation coefficient range 0.510–0.571, $p < 5e\text{-}24$ for all). With a threshold of 1.5 for modified SBNA scores, the sensitivity was 0.548, specificity 0.908, positive predictive value 0.963, and negative predictive value 0.312.

ROC curves were generated by testing each of the methods on the dataset of 2800 variants present in MEE patients (AUC range: 0.788 [modified SBNA], 0.829–0.833 [PolyPhen2], 0.819 [AlphaMissense], and 0.809 [EVE]), Supplementary Figs. 6C, 7C). The modified SBNA, PolyPhen2, AlphaMissense, and EVE performed similarly, though PolyPhen2

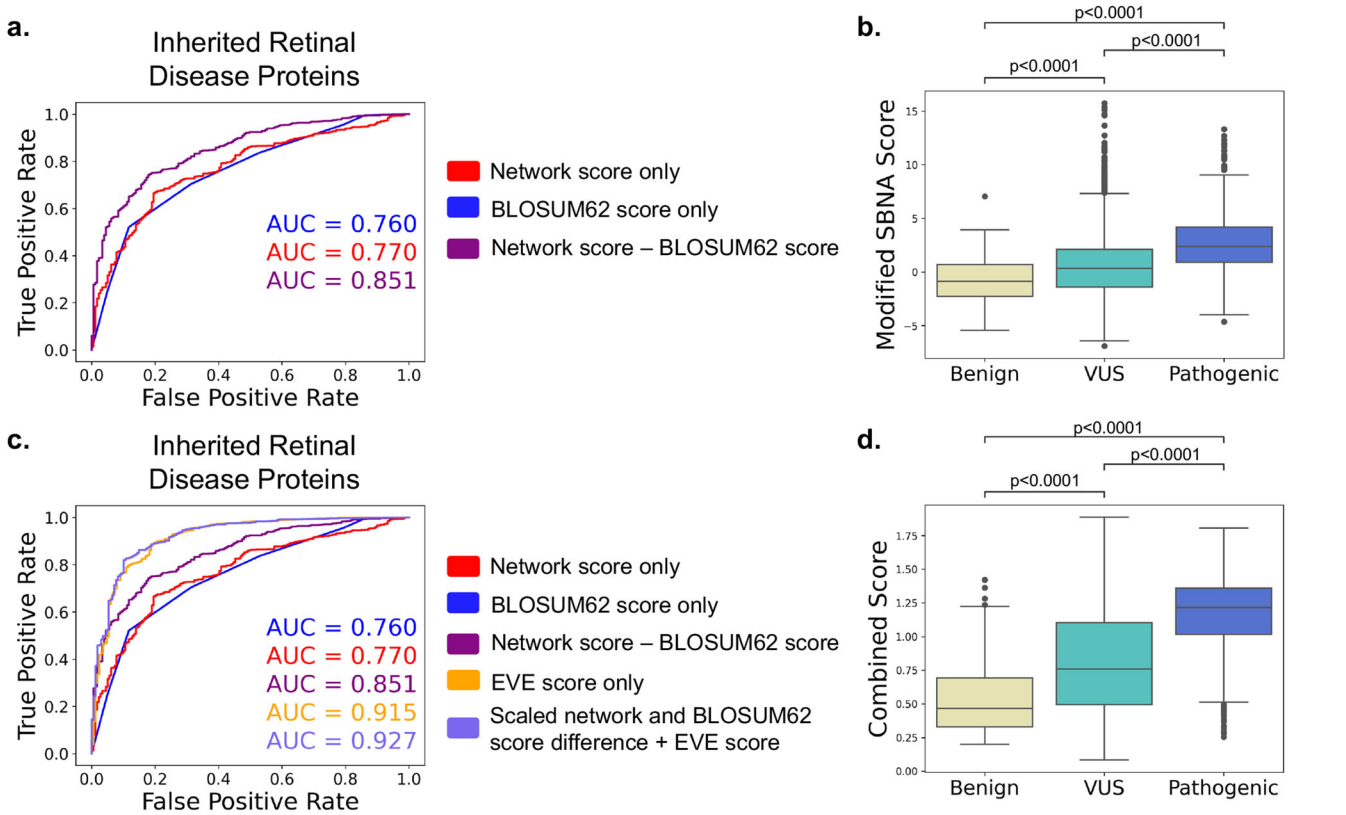


Fig. 3 | Modeling using SBNA and BLOSUM62 is superior to SBNA network scores and BLOSUM62 scores alone. a ROC curves for network scores alone (red), BLOSUM62 scores alone (blue), and the difference between network scores and BLOSUM62 scores (purple). ROC curves were determined using all variants with available clinical phenotype data for all 47 inherited retinal disease proteins. AUC values are shown for each curve. **b** Pooled comparison between the difference between network scores and BLOSUM62 scores for variants with available clinical phenotype data for all 47 inherited retinal disease proteins. **c** ROC as shown in (**a**) with added comparison to EVE scores (orange) and the sum of EVE scores and the scaled difference between network scores and BLOSUM62 scores (light purple). ROC curves were determined using all variants with available clinical phenotype data for all 47 inherited retinal disease proteins. AUC values are shown for each curve. **d** Pooled comparison between the sum of EVE scores and the scaled difference between network scores and BLOSUM62 scores ("combined score") for variants with available clinical phenotype data for all 47 inherited retinal disease proteins.

and AlphaMissense scores may be inflated due to training on ClinVar pathogenicity labels.

Combining SBNA and EVE outperforms all methods individually

The correlation and prediction results suggest that structural information and sequence conservation provide distinctly important insight into pathogenicity, and thus, incorporating orthogonal metrics into a single score may help to improve model correlation with phenotypic benchmarks. We thus sought to use the two unbiased methods (modified SBNA and EVE). The modified network scores were scaled and added to EVE scores to create a combined score with a range of 0 to 2 (as EVE scores fall between 0 and 1, and SBNA scores were scaled based on the maximum and minimum values across all proteins such that they fell between 0 and 1 before adding to the EVE scores). When applied to the 2800 rare variants from MEE patients as well as the 47 IRD genes, this combined score distinguished pathogenicity (Supplementary Fig. 7D, benign vs. pathogenic $p = 3.056e\text{-}17$; Fig. 3d, benign vs. pathogenic $p = 8.839e\text{-}69$) and outperformed all other models with an AUC of 0.859 (Supplementary Fig. 7E) for the 2800 variants and 0.927 (Fig. 3c) for the IRD genes. With a threshold of 1.0 for the combination score, the sensitivity was 0.765, specificity 0.899, positive predictive value of 0.976 and negative predictive value of 0.416. While we note that EVE alone performed quite well (AUC = 0.915), adding the modified SBNA improves performance and, importantly, unlike EVE alone, offers a direct structural explanation for the mechanism through which variation may affect phenotype.

Model incorporating the modified SBNA scores identifies putative disease-causing variants with an unclear genetic basis for clinical disease

A significant percentage of patients with clinical presentations consistent with IRD lack an identified genetic basis for their phenotype, and this is also observed for patients who receive care from the Inherited Retinal Disease Service at MEE. We therefore evaluated genetic data from 3621 probands with a clinical diagnosis of an IRD based on visual acuity, visual field testing, clinical exam, fundus autofluorescence imaging, optical coherence tomography and electroretinogram in individuals who underwent targeted or whole-exome sequencing. Mitochondrial causes of IRD were excluded. Missense variants of interest were defined using variant ranking software[43] and residence in one of the studied 47 IRD genes. There were 2948 unique variants identified and categorized as either "pathogenic", "likely pathogenic", "VUS", or "benign" based on a known variant consequence in the literature using ACMG/AMP criteria[44] and ClinVar designations[20] (Supplementary Fig. 8). Variants were further categorized in the context of individual patients as "likely causal" if they were pathogenic or likely pathogenic, the zygosity was consistent with known modes of inheritance, and the clinical presentation was consistent with the known consequence of the affected gene. Of all the reviewed patients, 455 patients carried variants of interest in the 47 IRD genes. Before applying SBNA, 357 were found to have variants that were "likely causal", while 63 patients harbored one or more VUSs that prohibited a molecular diagnosis. The remaining 35 patients had non-missense variants or variants within a region that lacked available structural data and were therefore excluded.

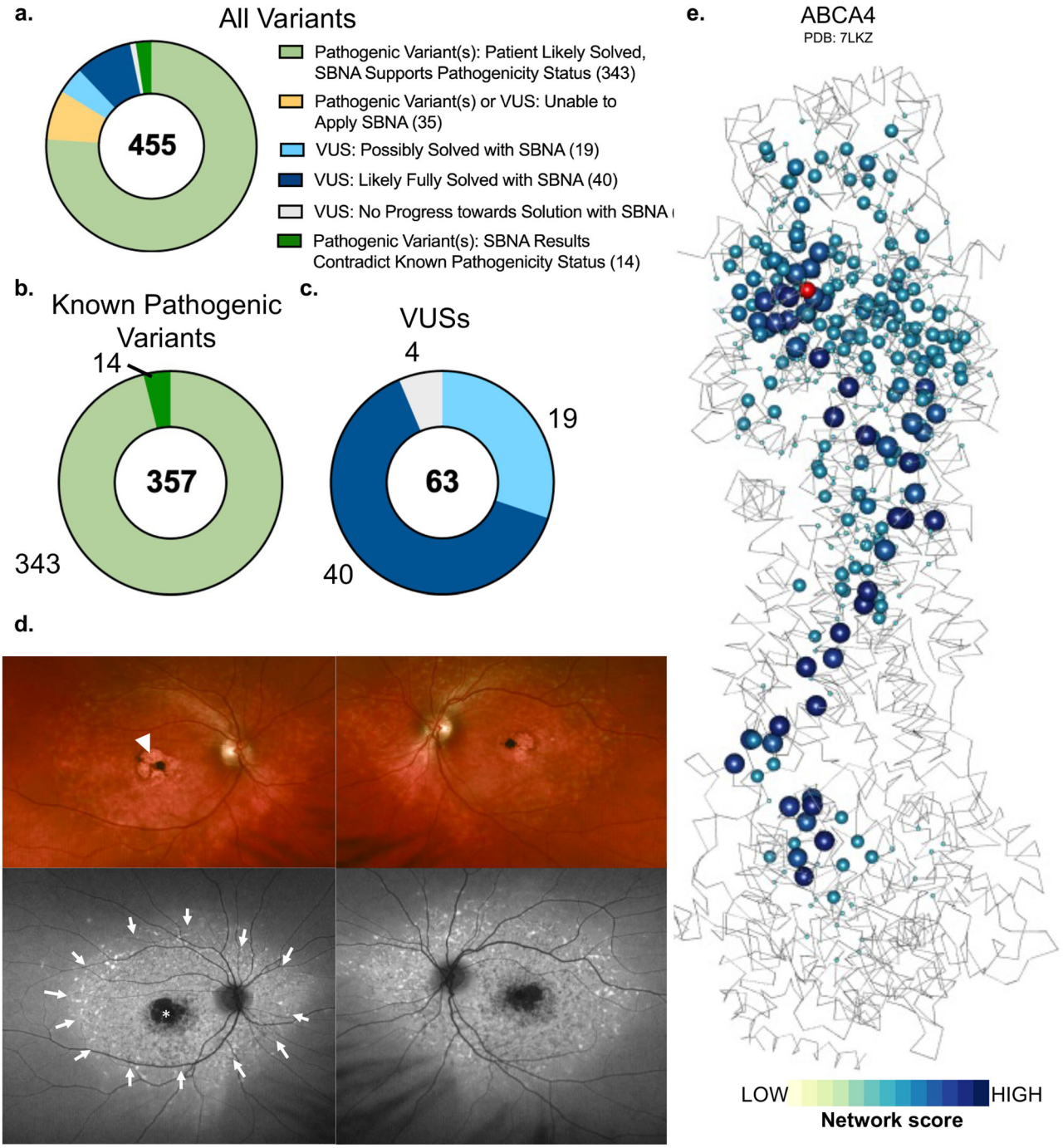


Fig. 4 | SBNA helps identify pathogenic variants in patients with inherited retinal disease. a Categorization of results from the application of modified structure-based network analysis (SBNA) to a dataset of possibly solving patient variants. Results were further subdivided into those from patients with known putative genetic causes of disease (**b**) and those from patients with only VUSs in known inherited retinal disease-associated genes (**c**). **d** Representation of network scores for a sample structure with putative solving genetic variants. Sphere radius corresponds to network score magnitude at a particular position. A patient with clinical evidence of ABCA4 disease (**d**) as evidenced by bilateral foveal pigmentary changes (arrowhead) on color fundus photo and bilateral RPE atrophy (star) and hyper autofluorescent flecks throughout the posterior pole on fundus autofluorescence imaging (arrows) but with no complete genetic explanation was fully solved using SBNA which highlighted two variants (Pro1380Leu and Arg1097Ser) that score highly in the ABCA4 protein structure (**e**). Arg1097Ser was a VUS and is indicated in red within the structure.

We generated pathogenicity predictions for the 2,948 IRD gene variants discovered in the probands (Fig. 4a). Variants receiving a modified SBNA score of >1.5 (calibrated using probability estimates from the regression modeling) were deemed pathogenic. Scores were considered in the context of the identified genetic variants in known IRD genes for each patient with a clinical presentation consistent with IRD. Variants were matched with any phenotypic data available in ClinVar to roughly benchmark the pathogenicity scores. Similar to the ClinVar analysis of IRD genes, there were observable differences between the modified SBNA scores assigned to known pathogenic variants as compared to benign variants and VUS/variants without any available clinical data (benign median −1.478, VUS median −0.656, pathogenic median 2.148; benign vs. pathogenic $p = 1.713e\text{-}30$).

For the 357 patients who harbored known pathogenic variants sufficient to cause disease, the modified SBNA scores were concordant with these pathogenicity categorizations in 96.0% of cases (Fig. 4b). For the 63 patients with VUSs as categorized by ACMG/AMP standards[44] and/or ClinVar, the modified SBNA scores offered support for a genetic cause of disease for 40 patients (23 unique variants, Fig. 4c). By contrast, EVE scores offered support for a genetic cause of disease in 25 patients (20 in common with SBNA), PolyPhen2 scores trained on HumVar offered support in 33 patients (27 in common with SBNA), and AlphaMissense offered support in 40 patients (34 in common with SBNA). Combined EVE and modified SBNA scores offered support in 23 patients (20 in common with SBNA), but this analysis was limited because EVE does not provide scores for 15 patients. In the 15 patients where modified SBNA scores suggested a putative causative variant but EVE scores did not, one patient had a variant for which no EVE score was provided in the database, and the remainder had at least one candidate variant with an EVE score below the predicted pathogenicity threshold. Modes of inheritance included autosomal recessive (in combination with a known pathogenic variant or a second VUS with a high estimated probability of pathogenicity; $n = 15$), autosomal dominant ($n = 2$), and X-linked recessive ($n = 5$). For example, a patient with autosomal recessive *ABCA4*-related disease was found to have variants p.(Pro1380Leu) (known pathogenic) and p.(Arg1097Ser) (VUS). The p.(Arg1097Ser) variant scored highly (SBNA score 3.672, BLOSUM62 score -1, score difference 4.672), suggesting it is likely pathogenic and thus completing the genetic solution for this patient. Similarly, the VUS p.(Cys302Tyr) in *RPGR* was found in a hemizygous patient with phenotypic findings consistent with X-linked IRD and also scored highly (SBNA score 2.154, BLOSUM62 score -2, score difference 4.154) (Supplementary Fig. 9). For 19 patients, the modified SBNA scores contributed towards identifying a possible but not completely solved genetic cause, such as only one heterozygous VUS receiving a strong score. Finally, for the four remaining unsolved patients, SBNA was not possible due to lack of crystal structure data for those portions of the IRD-associated proteins or due to the presence of non-missense variants.

AlphaFold2 can improve structural coverage for SBNA

Despite evidence of strong performance when applied to IRD-associated proteins, SBNA remains broadly limited by the availability of high-quality structural data for proteins of interest. This structural coverage must also overlap with the availability of high-quality phenotypic data from ClinVar, limiting the scope of analysis (Fig. 5a). Applying AlphaFold2 may provide a path toward overcoming this limitation. For example, NR2E3 was excluded from the set of 47 IRD genes because the only available structural data is from a chimera formed between one domain of NR2E3 and an unrelated bacterial protein (PDB: 4LOG). SBNA performs poorly on this non-physiologic structure with relatively poor coverage of NR2E3 (47%) (Fig. 5b). However, when SBNA is applied to the full AlphaFold2-generated human NR2E3 structure, performance improves considerably (Fig. 5c).

To further establish that AlphaFold2 can help to overcome the limited availability of high-quality structural data for SBNA, we selected ten IRD-associated genes without available structural data that had a considerable amount of data in ClinVar (Fig. 5d). We performed SBNA on the AlphaFold2-generated structures for these genes and found a significant difference between the SBNA scores assigned to benign and pathogenic variants ($p = 1.201e\text{-}7$; benign median −0.539, VUS median −0.491, pathogenic median 0.076) (Fig. 5e). However, the magnitude of difference between the median benign and pathogenic network scores was decreased compared to SBNA performed on the structural data from the 47 IRD-associated genes. Still, these results suggest that AlphaFold2 could potentially be useful in expanding the applicability of SBNA, although it is not clear that the quality of this analysis would be superior to that performed on experimentally generated structural data.

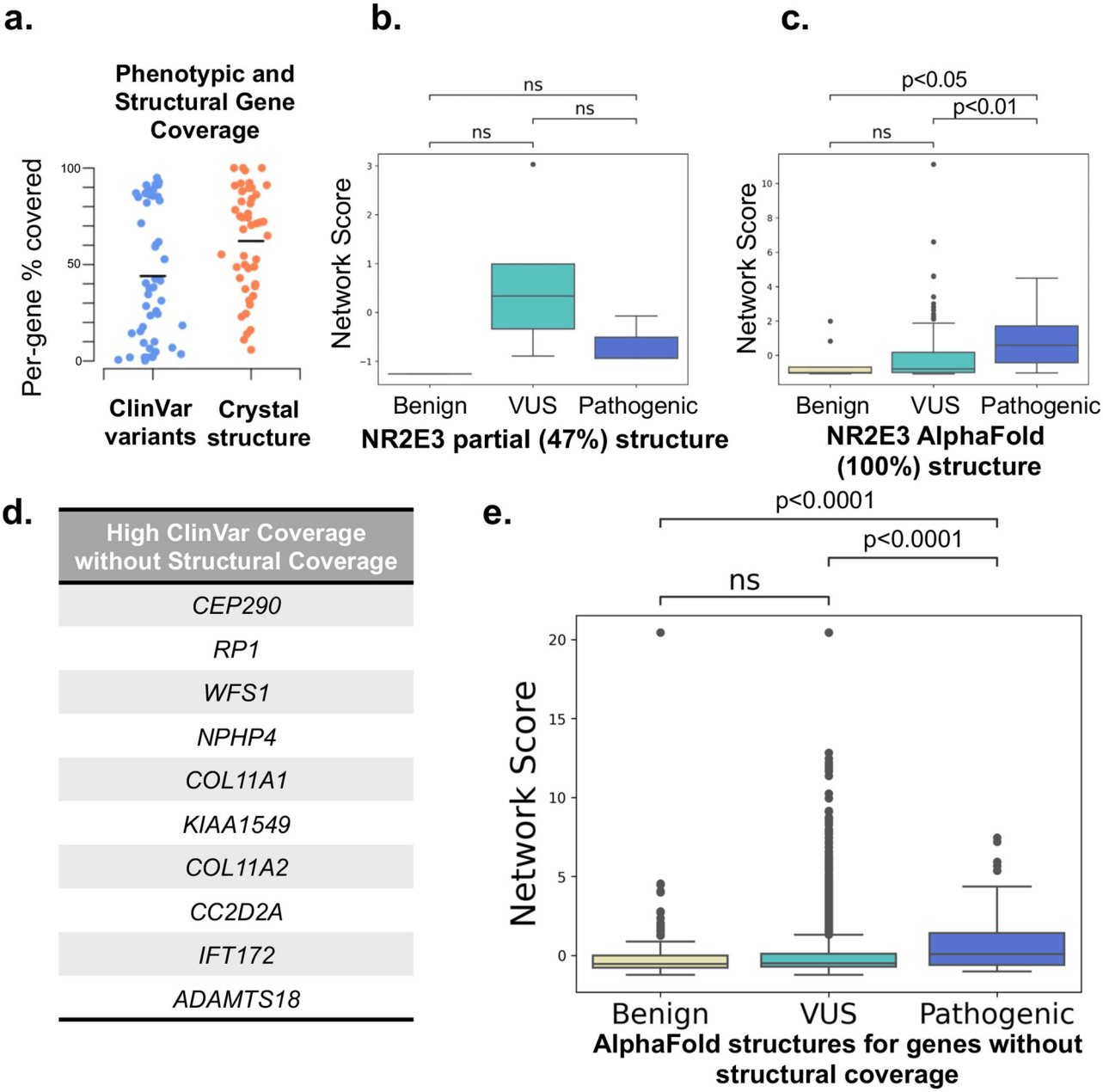
Discussion

In this study, we applied SBNA to human proteins and demonstrated that the resulting residue network scores, especially when augmented with BLOSUM62 substitution scores, correlate strongly with missense variants that cause functional deficit or clinical disease. We leveraged those scores, in combination with BLOSUM62 substitution scores, to generate estimates of the likelihood that a particular missense variant was pathogenic. Using these estimates, we were able to nominate putative genetic solutions for 40 patients with clinical evidence of IRD. These results suggest that SBNA provides meaningful insights for patients with an unclear genetic basis for their clinical symptoms, particularly for patients with inherited retinal diseases. Importantly, we note that this method is based purely on structural principles rather than training on labeled outcome data, which means that it not only provides an unbiased prediction of pathogenicity but also the mechanism of pathogenicity (i.e., structural change) is proposed.

Numerous computational tools for the prediction of variant pathogenicity have been developed[15,32,45–56]. Virtually all of these tools use sequence conservation and protein secondary structure or domain information. These two categories of features can work well, especially when used in conjunction with one another. However, SBNA is distinct in that it captures the structural topology of individual amino acid residues in the context of the protein architecture and does not rely on pre-existing annotations. Importantly, unlike all existing tools, the performance of SBNA does not seem to rely on a training step involving thousands of phenotype-genotype combinations or multiple sequence alignment. As shown, the raw SBNA score combined with BLOSUM62 (without training) achieves an AUC of 0.851 on the dataset of 47 IRD proteins. Furthermore, the biophysical basis for SBNA scores can be analyzed on an atomic level. This transparency facilitates downstream applications, such as the consideration of possible gene therapy targets, and is not available within models that heavily leverage machine-learning approaches[26,32]. The predictive value of SBNA can be further augmented by incorporating EVE scores, which are ultimately based on evolutionary multiple sequence alignments[26], to achieve an AUC of 0.927. This suggests that incorporating multiple orthogonal metrics may strengthen predictive models.

Applying SBNA may provide additional insight into patient candidacy for both approved and developmental gene therapy treatments. By identifying candidate pathogenic variants with high confidence, SBNA could be used to nominate a limited subset of variants for expedited in vivo validation to fast-track delivery of appropriate therapies to patients. Broadly, the patients that may benefit clinically from SBNA fall into three categories. The first are patients who may be a candidate for gene-specific, variant-agnostic therapies, such as the FDA-approved *RPE65*-targeted gene therapy voretigene neparvovec[57]. This also applies to any gene with variant-agnostic therapies in ongoing clinical trials. A second category of patients could potentially benefit from variant-specific gene editing therapies, such as those facilitated by CRISPR-Cas9-mediated non-homologous end joining, which are currently under development for *CEP290*[58]. The third category of patients with IRD who may benefit from SBNA are not yet candidates for any existing clinical therapies. However, using SBNA on a large scale to identify candidate disease-causing variants may be informative for nominating new genes for variant-agnostic therapy development and new variants for gene editing therapy development in order to maximize potential patient benefit.

This approach is limited by the availability of high-quality structural data, a requirement for SBNA[29], though we note that any tool that proposes to use structural data will be reliant on this. Numerous proteins, including the majority that correspond to known genetic variants associated with IRD, lack available x-ray crystallography or cryo-EM structures altogether. In some cases, structures are available but are not of sufficient resolution to facilitate downstream network analysis. Furthermore, this approach will only be applicable to variants that result in a negative structural change. Other types of variants – such as splice site variants – will not be captured with this approach. However, given that this tool only requires protein structure data, we expect that the performance of SBNA will continue to improve. Software that leverages artificial intelligence to predict protein



d.

High ClinVar Coverage without Structural Coverage
CEP290
RP1
WFS1
NPHP4
COL11A1
KIAA1549
COL11A2
CC2D2A
IFT172
ADAMTS18

Fig. 5 | AlphaFold2 improves coverage for SBNA. a The percentages of high confidence (≥ 2 stars) ClinVar variants for each of 47 IRD genes captured with SBNA are widely distributed (blue); the per-protein percentages of solved structure are widely distributed (orange). **b** NR2E3 only has structural data available for a single domain as part of a chimera. SBNA scores for NR2E3 correlate poorly with pathogenicity because the structure is partial and non-physiologic, but (**c**) the performance improves when using AlphaFold2 to model the full human protein. **d** Ten IRD-associated genes without available structural data were selected, and **e** SBNA scores were calculated for the full AlphaFold2 structures.

structure, such as AlphaFold2 or RoseTTAFold, have already emerged to help compensate for the lack of available structural data[42,59]. We have demonstrated that SBNA can be meaningfully applied to some of these in silico-generated structures, which may provide a path towards overcoming the limitation of high-quality structural data availability going forward.

In conclusion, SBNA is a new tool to estimate the likely extent of the phenotypic impact of missense variants by assessing the topological important of affected residues to protein structure. We demonstrated that this technique could be meaningfully applied to human proteins and showcased the use of SBNA in IRD patients who lack a clear genetic diagnosis. These types of insights could contribute to the design of novel gene therapies targeted at implicated genetic variants[57,60,61].

Methods

Structural data

All structural data were downloaded from the Protein Data Bank[34]. Individual accession numbers for the well-studied human proteins and inherited retinal disease proteins are listed in Supplementary Table 1. For the non-human benchmark proteins, the same files were used as previously described[29]. In cases where multiple human structures were available, the highest oligomeric state of the protein with a resolution of ~3 angstroms or better was used. In cases where no human structure was available, the structure of one or more homologs was used. If multiple chains were present within the PDB file due to crystal packing rather than true oligomerization, only one of these chains was used for SBNA. Solvent and water molecules

were removed from all PDB files prior to SBNA, but ligands and protein binding partners were included in the analysis. Only the protein of interest was designated to have a network score calculated by SBNA.

Structure-based Network Analysis

Structure-based network analysis was used to calculate network scores as previously described[29,35]. The details of this method have been described previously. As before, in cases where multiple conformations of a structure were used, network scores were averaged for each amino acid position. In cases where no human structure was available, the structures of one or more non-human homologs as listed in Supplementary Table 1 were used as templates in Modeler to generate a predicated human structure. Relative solvent accessibility (RSA) was calculated using DSSP files[62] and previously published maximum solvent accessible surface area values[63]. Code to perform structure-based network analysis is publicly available via Zenodo (https://doi.org/10.5281/zenodo.2597484).

Data analysis and visualization

Data analysis was performed using Python (version 3.8.2), with visualizations generated using the "matplotlib" package. Logistic regressions were performed using the "glm" package in R (version 4.0.4). Intercepts were set to zero for all logistic regression models. SBNA and logistic regression results for all tested variants are available via the Harvard Dataverse (https://doi.org/10.7910/DVN/YEEPDY). Network score visualizations were generated in R (version 4.0.4) using the "rgl" package to implement OpenGL. The backbone centroid (centroid of nitrogen, alpha carbon, and carbon) positions were plotted along x, y, and z axes, and nodes were colored and given sphere radii corresponding to network scores. The protein structure backbone was then plotted, connecting the alpha carbons. Plotted structures were manually rotated, and two-dimensional views of interest were downloaded for inclusion.

Phenotype data

Clinical phenotype data was downloaded from ClinVar. Clinical evidence with a two-gold star level designation (meaning that "two or more submitters with assertion criteria and evidence [or a public contact] provided the same interpretation") or better was included in the published analyses[20]. Pathogenicity designations from ClinVar were binned into Benign ("Benign", "Benign/Likely benign", "Likely benign"), VUS ("Uncertain significance", "not provided", "Conflicting interpretations of pathogenicity"), and Pathogenic ("Pathogenic", "Pathogenic/Likely pathogenic", "Likely pathogenic") categories for analysis.

Additionally, allelic variation data from gnomAD[64] was considered for each gene. Loci with at least 250 available allelic variants in gnomAD were considered benign variants. Results from ClinVar and gnomAD for each gene were considered in combination for these analyses.

Functional data

Functional data for the four well-studied human proteins—BRCA1[36], ERK2[39], PTEN[37], and HRAS[38]—that had been previously published was used for this analysis. In cases where functional scores were assigned to multiple amino acid variants at the same position (e.g., if different functional scores were calculated for Ala101Pro and Ala101Gln), the arithmetic mean of all functional scores at that position was used.

Patient data

Patient data was gathered from among those presenting to the Inherited Retinal Disorders Service at Massachusetts Eye and Ear between the 1980s and 2020s. Appropriate written informed consent was obtained from all included patients, and approval was granted by the Mass General Brigham/Massachusetts Eye and Ear Institutional Review Board Protocol Number 2019P001098. This study was approved by the local institutional review board and adhered to the Declaration of Helsinki. Informed consent was obtained from all individuals on whom clinical data and genetic testing were performed.

Genetic analyses

Pre-existing genetic solutions were available for all of the 3621 IRD cases, where genetic diagnoses for 3018 cases was obtained by targeted next-generation sequencing approaches[43] or whole genome sequencing, and the remaining solutions were available from prior single-strand conformation polymorphism (SSCP) analysis and Sanger sequencing. Sequence data was aligned to the hg38 genome build, and the subsequent variant calling, annotation, and analyses were performed as described[43].

Model Comparisons

SBNA was compared to results from PolyPhen2[41] (trained on both HumDiv and HumVar and EVE[26], with scores from both models generated as described in their respective publications. EVE scores were downloaded from https://evemodel.org; PolyPhen2 scores were generated using the "batch query" function at http://genetics.bwh.harvard.edu/pph2/. To analyze the likelihood of pathogenicity, SBNA scores, EVE scores, and PolyPhen2 scores (trained on HumVar) were calculated for a set of 2800 variants. Cutoffs for each model were as follows: modified SBNA score >1.5 (corresponding to an approximate pathogenicity probability of 75% in the logistic regression model), EVE score ≥0.65, and PolyPhen2 designation of "probably damaging". The "combined score" was generated by scaling modified SBNA scores to fall between 0 and 1 based on the minimum and maximum values across all proteins and adding them to EVE scores. A threshold of 1.0 was selected as this corresponds to the sum of the minimum pathogenic EVE score (0.65) and the scaled version of the minimum pathogenic modified SBNA score (1.5, which scales to 0.37) with two significant figures.

Statistical analysis

Statistical analysis, including the generation of ROC curves, was performed using the "scipy.stats" package in Python as well as GraphPad Prism (version 9). Comparisons between three or more categories were made using the non-parametric Kruskal-Wallis test with Dunn's post hoc analysis corrected for multiple comparisons with a Bonferroni correction. Comparisons between the two categories were performed using the non-parametric Mann–Whitney *U*-test. The correlation between the two datasets was calculated using non-parametric Spearman correlation coefficients. Spearman correlation coefficients between network scores and ClinVar pathogenicity designations was calculated by assigning 0 to benign variants, 1 to VUS, and 2 to pathogenic variants.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

ClinVar, gnomAD, and Protein Data Bank data were publicly available and can be accessed using the gene names and unique identifiers in Supplementary Table 1. Patient data from Massachusetts Eye and Ear cannot be shared publicly as specified by the Institutional Review Board due to concerns regarding possible identifiability of patients with relatively rare genetic conditions. Qualified researchers may be able to access the data via collaboration and data usage agreement by contacting the corresponding author.

Code availability

Code to perform structure-based network analysis is publicly available via Zenodo (https://doi.org/10.5281/zenodo.2597484).

Received: 15 November 2023; Accepted: 2 May 2024;
Published online: 27 May 2024

References

1. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).

2. Berger, W., Kloeckener-Gruissem, B. & Neidhardt, J. The molecular basis of human retinal and vitreoretinal diseases. *Prog. Retin. Eye Res.* **29**, 335–375 (2010).
3. Consugar, M. B. et al. Panel-based genetic diagnostic testing for inherited eye diseases is highly accurate and reproducible, and more sensitive for variant detection, than exome sequencing. *Genet. Med.* **17**, 253–261 (2015).
4. Wang, F. et al. Next generation sequencing-based molecular diagnosis of retinitis pigmentosa: identification of a novel genotype-phenotype correlation and clinical refinements. *Hum. Genet.* **133**, 331–345 (2014).
5. Ge, Z. et al. NGS-based molecular diagnosis of 105 eyeGENE((R)) probands with retinitis pigmentosa. *Sci. Rep.* **5**, 18287 (2015).
6. Hafler, B. P. Clinical progress in inherited retinal degenerations: gene therapy clinical trials and advances in genetic sequencing. *Retina* **37**, 417–423 (2017).
7. Ben-Yosef, T. Inherited retinal diseases. *Int. J. Mol. Sci*. https://doi.org/10.3390/ijms232113467 (2022).
8. Peterson, T. A., Doughty, E. & Kann, M. G. Towards precision medicine: advances in computational approaches for the analysis of human variants. *J. Mol. Biol.* **425**, 4047–4063 (2013).
9. Niroula, A. & Vihinen, M. Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.* **37**, 579–597 (2016).
10. Rost, B., Radivojac, P. & Bromberg, Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* **590**, 2327–2341 (2016).
11. Pejaver, V. et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).
12. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
13. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
14. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
15. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
16. Feng, B. J. PERCH: a unified framework for disease gene prioritization. *Hum. Mutat.* **38**, 243–251 (2017).
17. Raimondi, D. et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201–W206 (2017).
18. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet*. **48**, 214–220 (2016).
19. Dong, C. et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
20. Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
21. Wang, T. et al. Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. *Nat. Commun.* **9**, 441 (2018).
22. Miosge, L. A. et al. Comparison of predicted and actual consequences of missense mutations. *Proc. Natl Acad. Sci. USA* **112**, E5189–E5198 (2015).
23. Grimm, D. G. et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
24. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* **17**, 405–424 (2015).
25. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
26. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
27. Mishra, P., Flynn, J. M., Starr, T. N. & Bolon, D. N. A. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell Rep.* **15**, 588–598 (2016).
28. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
29. Gaiha, G. D. et al. Structural topology defines protective CD8(+) T cell epitopes in the HIV proteome. *Science* **364**, 480–484 (2019).
30. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
31. Cuevas, J. M., Geller, R., Garijo, R., Lopez-Aldeguer, J. & Sanjuan, R. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol.* **13**, e1002251 (2015).
32. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
33. Wan, A., Place, E., Pierce, E. A. & Comander, J. Characterizing variants of unknown significance in rhodopsin: a functional genomics approach. *Hum. Mutat.* **40**, 1127–1144 (2019).
34. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
35. Nathan, A. et al. Structure-guided T cell vaccine design for SARS-CoV-2 variants and sarbecoviruses. *Cell* **184**, 4401–4413 e4410 (2021).
36. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
37. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
38. Hidalgo, F. et al. A saturation-mutagenesis analysis of the interplay between stability and activation in Ras. *Elife* https://doi.org/10.7554/eLife.76595 (2022).
39. Brenan, L. et al. Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell Rep.* **17**, 1171–1183 (2016).
40. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* **22**, 1035–1036 (2004).
41. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* https://doi.org/10.1002/0471142905.hg0720s76 (2013).
42. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
43. Zampaglione, E. et al. The importance of automation in genetic diagnosis: lessons from analyzing an inherited retinal degeneration cohort with the Mendelian Analysis Toolkit (MATK). *Genet. Med.* **24**, 332–343 (2022).
44. Kleinberger, J., Maloney, K. A., Pollin, T. I. & Jeng, L. J. An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. *Genet. Med.* **18**, 1165 (2016).
45. Bao, L., Zhou, M. & Cui, Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* **33**, W480–482, (2005).
46. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
47. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
48. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
49. Tang, H. & Thomas, P. D. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics* **32**, 2230–2232 (2016).

50. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
51. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **315**, 771–786 (2002).
52. Li, B. et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744–2750 (2009).
53. Saunders, C. T. & Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891–901 (2002).
54. Seifi, M. & Walter, M. A. Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms. *PLoS ONE* **13**, e0195971 (2018).
55. Shihab, H. A. et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
56. Tavtigian, S. V. et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43**, 295–305 (2006).
57. Russell, S. et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet* **390**, 849–860 (2017).
58. Burnight, E. R. et al. Using CRISPR-Cas9 to generate gene-corrected autologous iPSCs for the treatment of inherited retinal degeneration. *Mol. Ther.* **25**, 1999–2013 (2017).
59. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
60. Nuzbrokh, Y., Ragi, S. D. & Tsang, S. H. Gene therapy for inherited retinal diseases. *Ann. Transl. Med.* **9**, 1278 (2021).
61. Fenner, B. J. et al. Gene-based therapeutics for inherited retinal diseases. *Front. Genet.* **12**, 794805 (2021).
62. Joosten, R. P. et al. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411–419, (2011).
63. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* **8**, e80635 (2013).
64. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

Acknowledgements

B.M.H. is supported by award number T32GM144273 from the National Institute of General Medical Sciences and award number F30AI160908 from the National Institute of Allergy and Infectious Diseases. G.D.G. is supported by award numbers DP2AI154421, R01AI176533, and DP1DA058476, the Bill and Melinda Gates Foundation, the Burroughs Wellcome Career Award for Medical Scientists, and the Howard Goodman Fellowship. E.J.R. is supported by award number K12EY016335 from the National Eye Institute and the Massachusetts Lions Eye Research Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceptualization: B.M.H., G.D.G., and E.J.R.; Methodology: B.M.H., Y.L., A.N., G.D.G., and E.J.R.; Software: B.M.H., Y.L., and E.J.R.; Validation: B.M.H., Y.L., A.N., J.C., E.A.P., E.M.P., K.M.B., G.D.G., and E.J.R.; Formal analysis: B.M.H., Y.L., E.M.P., K.M.B., and E.J.R.; Resources: E.A.P., E.M.P., K.M.B., G.D.G., and E.J.R.; Data curation: B.M.H., J.C., E.A.P., E.M.P., K.M.B., and E.J.R.; Writing—original draft: B.M.H. and E.J.R.; Writing—review and editing: B.M.H., Y.L., A.N., D.G.V., J.C., E.A.P., E.M.P., K.M.B., G.D.G., and E.J.R.; Visualization: B.M.H. and E.J.R.; Supervision: D.G.V., J.C., E.A.P., K.M.B., G.D.G., and E.J.R.; Project administration and funding acquisition: E.J.R.

Competing interests

G.D.G. discloses research funding from Merck. E.J.R. and G.D.G. have filed two patent applications related to the use of SBNA in HIV and SARS-CoV-2.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-024-00416-w.

Correspondence and requests for materials should be addressed to Elizabeth J. Rossin.