nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19669-x OPEN

Inferring the molecular and phenotypic impact of amino acid variants with MutPred2

Vikas Pejaver[1,2], Jorge Urresti[3], Jose Lugo-Martinez[1,8], Kymberleigh A. Pagel[1,9], Guan Ning Lin[3,10], Hyun-Jun Nam[3], Matthew Mort[4], David N. Cooper[4], Jonathan Sebat[3,5,6], Lilia M. Iakoucheva[3✉], Sean D. Mooney[2✉] & Predrag Radivojac[1,7✉]

Identifying pathogenic variants and underlying functional alterations is challenging. To this end, we introduce MutPred2, a tool that improves the prioritization of pathogenic amino acid substitutions over existing methods, generates molecular mechanisms potentially causative of disease, and returns interpretable pathogenicity score distributions on individual genomes. Whilst its prioritization performance is state-of-the-art, a distinguishing feature of MutPred2 is the probabilistic modeling of variant impact on specific aspects of protein structure and function that can serve to guide experimental studies of phenotype-altering variants. We demonstrate the utility of MutPred2 in the identification of the structural and functional mutational signatures relevant to Mendelian disorders and the prioritization of de novo mutations associated with complex neurodevelopmental disorders. We then experimentally validate the functional impact of several variants identified in patients with such disorders. We argue that mechanism-driven studies of human inherited disease have the potential to significantly accelerate the discovery of clinically actionable variants.

[1] Department of Computer Science, Indiana University, Bloomington, IN, USA. [2] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA. [3] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [4] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK. [5] Beyster Center for Genomics of Psychiatric Diseases, University of California San Diego, La Jolla, CA, USA. [6] Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA. [7] Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [8] Present address: Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. [9] Present address: Institute for Computational Medicine, Whiting School of Engineering, Johns Hopkins University, 220 Hackerman Hall, 3400 N Charles St, Baltimore, MD 21218, USA. [10] Present address: School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, People's Republic of China. ✉email: lilyak@ucsd.edu; sdmooney@uw.edu; predrag@northeastern.edu

The discovery of pathogenic variants, i.e., variants capable of causing disease, generally relies on a combination of family- and population-based sequencing efforts[1]. To assist genetic studies, particularly in characterizing rare variants and dissecting complex disease, machine learning methods have recently been developed to identify the signatures of pathogenicity and to predict the impact of variants of unknown significance[2,3]. Although pathogenicity prediction methods have matured considerably over the past decade and are now routinely integrated into genomic pipelines, they continue to exhibit major shortcomings. Firstly, they remain inadequate to the task in exome-scale applications owing to a less than optimal balance of false-positive and true-positive detection rates[4,5]. Secondly, they do not generate actionable hypotheses regarding the molecular consequences of these variants[6].

The functional impact of variants may lead to a wide range of molecular changes, even within a single protein, including disrupted stability and structure, disrupted macromolecular binding, ablation of posttranslational modification (PTM) sites, among others (Fig. 1a). However, existing approaches generally provide little or no information about the potential mechanisms affected by mutations, or else simply map predicted pathogenic substitutions onto protein feature annotations (which are generally sparse) in public databases. These methods do not therefore explicitly model the type of change in local structure and function, and are fundamentally limited by the incomplete, incorrect, and inevitably biased annotations in major databases[7–9].

To address these challenges, we have extended our existing and widely used machine learning approach, MutPred[10], by developing an improved and statistically rigorous approach, MutPred2. This new algorithm quantifies the pathogenicity of amino acid substitutions, and describes how they affect the phenotype by modeling a broad repertoire of structural and functional alterations from amino acid sequence.

MutPred2 compares favorably with the existing tools recommended in the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) Standards and Guidelines[11] on stringent independent test sets. More importantly, by applying this methodology, we estimate the fraction of pathogenic missense variants in a personal genome, and identify molecular signatures associated with a data set of Mendelian disease variants and a data set of de novo mutations found in individuals diagnosed with neurodevelopmental disorders. Finally, we prioritize several high-scoring variants from this data set and experimentally validate their functional roles. Our results suggest new molecular targets and mechanisms impacted by multiple mutations across neurodevelopmental disorders. More broadly, this study demonstrates the power of the proposed mechanism-driven approach to studying human phenotypes.

Results

Overview of MutPred2. MutPred2 is a machine learning-based method, and software package that integrates genetic and molecular data to reason probabilistically about the pathogenicity of amino acid substitutions. This is achieved by providing (1) a general pathogenicity prediction, and (2) a ranked list of specific molecular alterations potentially affecting the phenotype. MutPred2 is a sequence-based model that utilizes methodology predicated upon recent machine learning advances in training from positive-unlabeled data, and which incorporates estimation of prior and posterior probabilities[12,13]. These estimates not only facilitate the interpretation of pathogenicity and molecular alteration scores, but also provide a framework to rigorously rank the underlying mechanisms[13]. Currently, MutPred2 models a broad range of structural and functional properties, including secondary structure, signal peptide and transmembrane topology, catalytic activity, macromolecular binding, PTMs, metal binding and allostery.

Challenges in variant interpretation tool development. To develop models for the mathematically sound inference of molecular mechanisms of disease, several statistical and computational challenges must be addressed. First, it is necessary to integrate disparate molecular and genetic data to develop models that have similar yet meaningful score interpretations[6]. Second, prediction software tools generally vary not only in terms of their feature representation and prediction algorithms, but also in their implementations, dependencies, and system requirements, which collectively hinder the development of a robust framework that seamlessly incorporates multiple models. Third, structural and functional properties occur with unequal prior probabilities, requiring sophisticated modeling to rank the properties affected by a substitution. Finally, although these property predictors are typically developed independently of one another, they are interrelated; i.e., a single substitution may affect more than one property. This places the burden of interpretation upon the user and can be overwhelming, when multiple properties are considered simultaneously.

To address the first two challenges, we developed sequence-based predictors for over 50 structural and functional protein properties (Supplementary Tables 1–4). All predictors, with minor exceptions, were trained with a common feature set, subjected to the same evaluation protocols, designed to output scores between 0 and 1, and implemented within the positive-unlabeled machine learning framework; i.e., a binary classification setting in which one is given a set of positive data and a set of unlabeled data (a mixture of positive and negative data points) with the goal of performing parameter estimation and learning, in order to discriminate between positive and negative data (Supplementary Methods)[14]. The areas under the ROC curves (AUCs) of these predictors are generally high (Supplementary Table 5). To address the third challenge, we estimated the prior probability for each property[13,15] and used it to transform raw prediction scores to posterior probabilities to facilitate direct comparisons with other properties[13] (Supplementary Table 6). Finally, we constructed a custom ontology of molecular alterations by grouping the properties into broader categories to capture the inherent relationships between the properties (Fig. 1b). This was carried out manually by combining our current understanding of protein structure and function, with the Variation Ontology as a template[16]. Our primary goal was to organize the space of molecular consequences so as to achieve user-friendly interpretation (Fig. 1c).

The MutPred2 pathogenicity model was trained on a set of 53,180 pathogenic and 206,946 unlabeled (putatively neutral) variants obtained from the Human Gene Mutation Database (HGMD)[17], SwissVar[18], dbSNP[19], and interspecies pairwise alignments. The model is a bagged ensemble of feed-forward neural networks[20], each trained on a balanced subset of pathogenic and unlabeled variants. The final prediction score is the average of the scores from all networks and ranges between 0 and 1; higher scores reflect a higher probability of pathogenicity. MutPred2's models for inferring molecular mechanisms were similarly trained from a variety of molecular data sets (Supplementary Methods), thereby ensuring effective integration of genetic and molecular data.

Evaluation of predictor performance. The choice of the training set is critical in machine learning. A common practice in

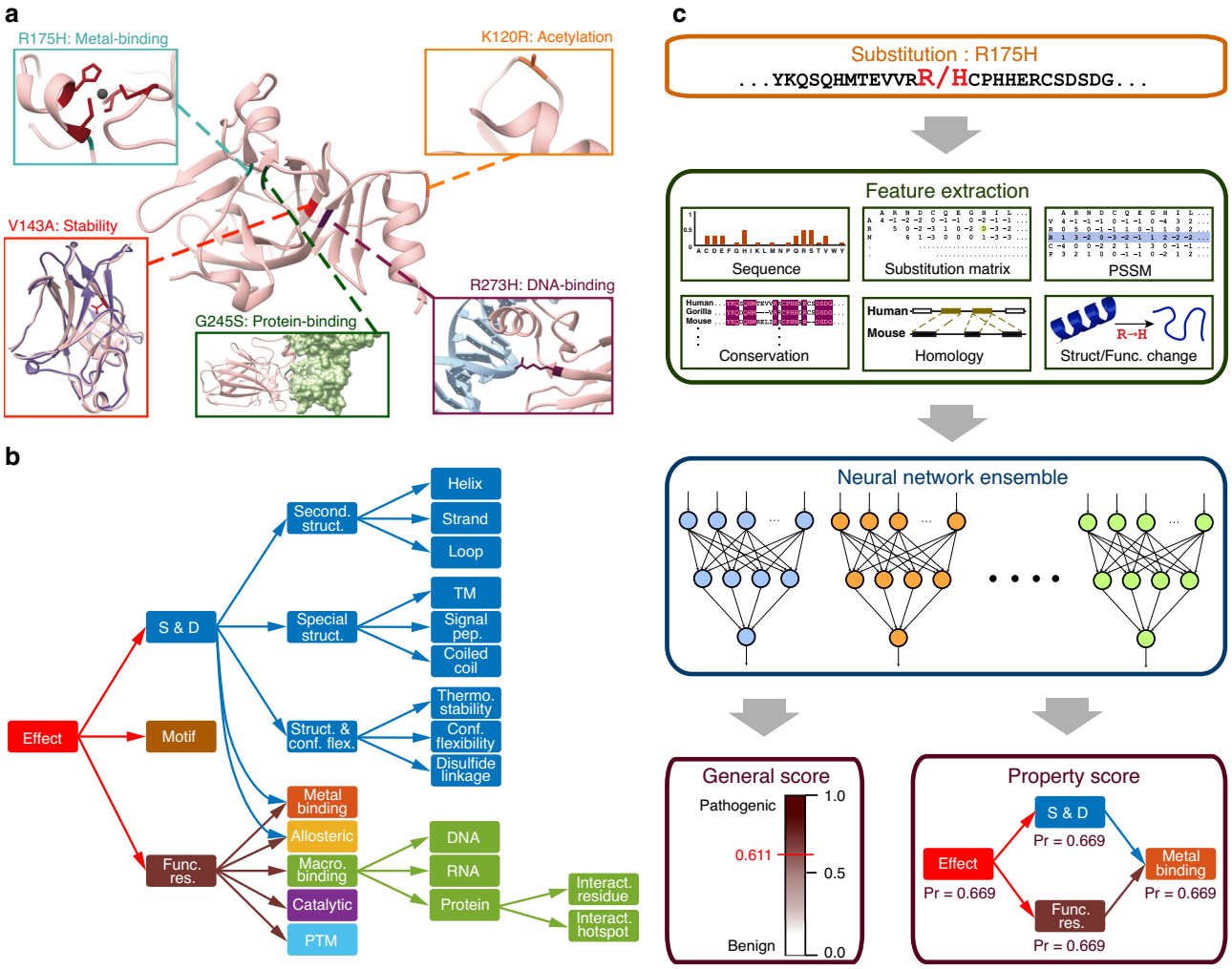

Fig. 1 MutPred2 and the molecular consequences of amino acid substitutions. a The human tumor suppressor p53 as an illustration of the numerous possible effects of amino acid substitutions on protein structure and function. Protein Data Bank IDs for the structures shown are 1TUP, 1YCS, 2J1W, and 2YBG. **b** The ontology constructed in this study to organize the possible structural and functional effects of amino acid substitutions. It is confined to the 53 properties included in MutPred2. **c** The MutPred2 workflow. For a given amino acid sequence and substitution, MutPred2 first extracts six categories of features. Changes in structure and function due to the substitution are also modeled by running the original and mutated sequences through different sequence-based protein property predictors. Two scores are obtained for each property and these are combined to generate two additional scores quantifying the loss and gain of the property in question. All four scores are included as features. Next, all categories of features are presented to an ensemble of 30 neural networks trained to distinguish between pathogenic and benign variants. MutPred2 returns two outputs, the general score and the property score. The general score is obtained from the neural network ensemble and indicates the pathogenicity of the given variant. It ranges between 0 and 1, with a higher score indicating a greater propensity to be pathogenic. The property score is assigned to each of the 53 properties for the given variant and also ranges between 0 and 1. The latter score is the posterior probability of loss or gain (whichever is greater) of the given property due to the substitution. The higher the property score, the more likely that the molecular mechanism of the disease involves the alteration of the property.

pathogenicity prediction involves the exclusion of rare variants from the unlabeled set to minimize biases arising from potentially undiscovered pathogenic variants; additional filtering based on specific types of data source may also be performed. To investigate the effects of various filtering criteria in training sets on classification performance and to select the most appropriate training set, different combinations of training and test sets were evaluated in an all-against-all performance assessment (Supplementary Table 7). We found that filtering the training set is beneficial only when the test set is also filtered, using the same criteria. Furthermore, using the entire unfiltered training set resulted in comparable or better performance in most cases, consistent with recent theoretical results justifying training from positive-unlabeled data[12,13]. Therefore, we chose not to perform any filtering in subsequent steps with the reasoning that bias introduced through different filtering schemes is more detrimental than random noise[12,13].

Using a per-protein tenfold cross-validation, the AUC was estimated at 87.7% (Fig. 2a). The training data, however, contains class-label noise; i.e., the set of disease variants may contain mutations incorrectly labeled as pathogenic and the set of unlabeled variants is by definition also a mixture of pathogenic and benign variants. We estimate the proportion of noisy positive variants to be 2.8% and the proportion of unlabeled variants that are pathogenic to be 5.8% (Supplementary Methods). These results allow us to provide a corrected estimate of the AUC of 91.3% (ref. [21]).

Consistent with previous studies, conservation-based features were the most discriminative[4,22,23] (Supplementary Tables 8 and 9). MutPred2 relies on precomputed databases of multiple sequence alignments and conservation scores to calculate these features. In cases where the input substitutions come from novel protein sequences or alternate isoforms, these data may often be

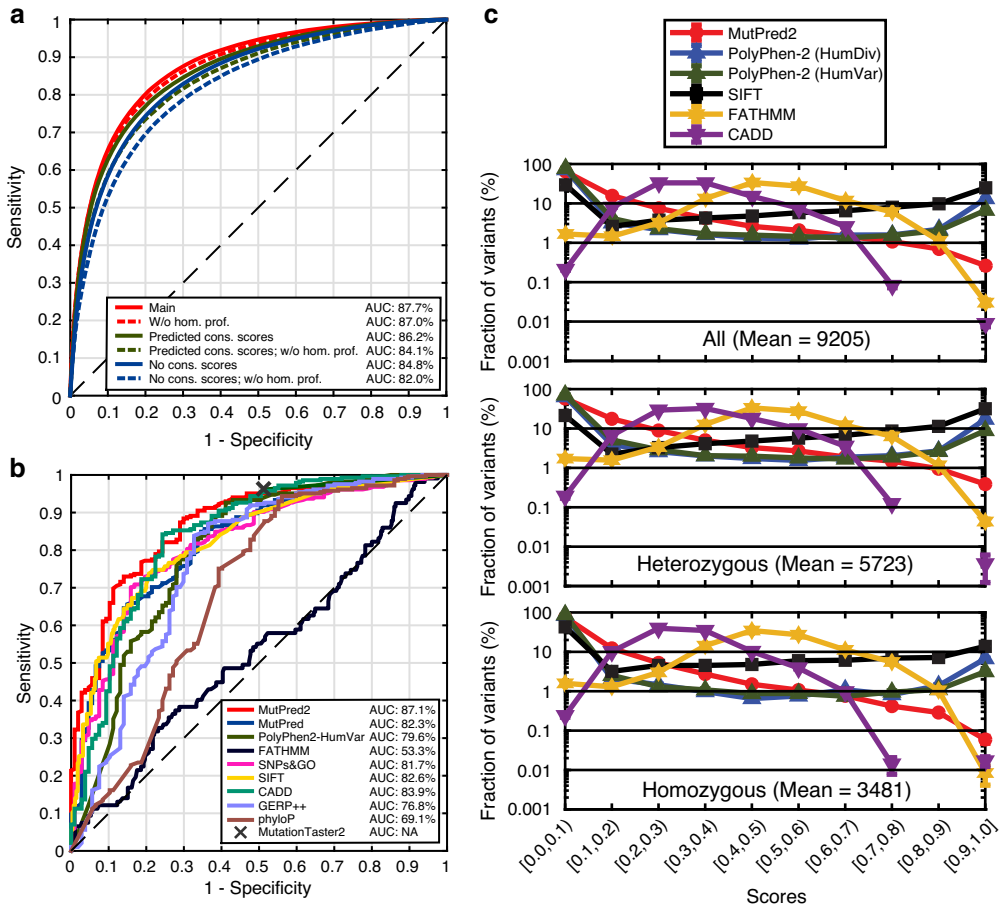

Fig. 2 Performance and interpretability of MutPred2. a ROC curves obtained through tenfold cross-validation on the MutPred2 training set. The main model represents MutPred2 in the default setting (with real conservation scores and homolog count profiles). All lines are paired with the solid line representing the model with homolog count profiles and the dashed line, representing the model without the profiles. **b** ROC curves on an independent test set, obtained from ClinVar and SwissVar by letting the data accumulate in these databases for 3 years. MutationTaster2 only returns a value of zero or one and therefore its performance is plotted as a single point (X). Since some tools could not assign scores to all variants, results from the subset of the variants (285 pathogenic and 107 benign) that are covered by all methods are shown. Detailed performance measures on this subset and a less stringent set (filtered at 80% sequence identity) are shown in Supplementary Tables 10 and 11. **c** Mean score distributions for MutPred2, PolyPhen-2 HumDiv, PolyPhen-2 HumVar, SIFT, FATHMM, and CADD applied to ten randomly selected exomes from the 1000 Genomes Project. Error bars represent the standard errors of the means, estimated by dividing the standard deviation in each bin by the square root of 10. All heterozygous and homozygous variants were plotted in separate panels. The mean in each panel represents the average number of variants found in an individual for the given category.

unavailable, prompting tools to avoid assigning predictions. To ensure that every input variant has a prediction and maximize data coverage, MutPred2 provides an option of predicting conservation-based features from sequence and PSI-BLAST position-specific scoring matrices (PSSMs) in cases when these features are unavailable. Although predicted conservation features only moderately correlate with actual values (Supplementary Fig. 1), models that included them performed two percentage points better than those that did not use any conservation features (Fig. 2a).

Previous studies reported conflicting results on the association between duplicated genes and their involvement in disease[24,25]. To investigate this, we created features that, for a given protein, enumerate homologous proteins from human and mouse at various levels of sequence similarity (50% or greater; Supplementary Methods). We observed that excluding these homology profiles did not drastically affect performance, when true conservation features were available, but resulted in a decrease of two percentage points in all other cases (Fig. 2a). This outcome supports the evidence for compensatory mechanisms in a variety of gene families, and a modest negative correlation between gene family size and variant pathogenicity[25].

We then evaluated MutPred2 against MutPred on its original training set under the same evaluation protocol. We found that MutPred2 had a similar AUC value as before (88.0%), outperforming the original MutPred approach by about five percentage points (Supplementary Fig. 2). Additional experiments are described in Supplementary Notes.

Evaluation against external tools. We next compared the performance of the default MutPred2 model with several state-of-the-art methods recommended in the ACMG/AMP Standards and Guidelines on the interpretation of sequence variants[11]; specifically, CADD[26], FATHMM[27], GERP++ (ref. [28]), MutationTaster2 (ref. [29]), MutPred, PhyloP[30], PolyPhen-2 (ref. [23]), SIFT[22], and SNPs&GO[31]. This analysis was carried out on an independent data set compiled from ClinVar[32] and SwissVar. We note that such comparisons are limited by differences in motivating goals, problem formulations, training data, and information used to make predictions, and are best addressed through

community-wide challenges, such as the Critical Assessment of Genome Interpretation; CAGI (http://genomeinterpretation.org). For a fair and rigorous comparison, we minimized potential biases by including only variants that were not in the training sets (where known) of the methods listed above. The analyses were confined to a nonredundant set of variants by comparing 25-residue fragments centered at the variant position to all such fragments in these training sets, and filtering out those that shared >50% sequence identity. This resulted in a highly stringent data set of 343 pathogenic mutations and 137 nonpathogenic mutations.

We found that MutPred2 performed substantially better than FATHMM, GERP++, MutationTaster2, PhyloP, and PolyPhen-2 in terms of AUC (Fig. 2b and Supplementary Table 10). The remaining methods resulted in AUCs of at least 80% and ROC curves that grouped together. However, MutPred2 still had the highest AUC (87.1%), largely attributable to its high sensitivities at lower false-positive rates (FPRs). This is especially relevant when considering the second-best performing tool, CADD. Despite the possibility that some of these variants may have been present in its training set and the fact that its AUC value was very close to that of MutPred2, CADD was more sensitive only when the FPR was high (between 20 and 30%). Interestingly, contrary to results from previous studies[33,34], the next best performing method on this independent data set was SIFT.

Given that most methods considered here do not return predictions for at least a proportion of the independent data set, the small data set size limits the interpretability of these findings. To mitigate this, we relaxed the fragment sequence identity threshold to 80% (as opposed to 50%), and expanded the independent data set to include 700 pathogenic mutations and 282 nonpathogenic mutations. Although individual performance values changed, the general trends remained unaffected, with the exception of SIFT's reduced performance (Supplementary Table 11). Finally, we have additionally undertaken a comprehensive prospective evaluation of these methods with a considerably larger independent data set comprising pathogenic variants obtained from the newest version of HGMD and unlabeled variants obtained from gnomAD. On this data set, while performance of all methods was lower, MutPred2 emerged as the most accurate on the full data set, as well as a subset comprising only those proteins that contained both types of variants (Supplementary Fig. 3, Supplmentary Table 12 and Supplementary Notes).

Interpretability of prediction scores. The interpretability of prediction scores is a generally underappreciated aspect of pathogenicity prediction. From this perspective, it is desirable that predicted pathogenic variants show a sufficient dispersion of scores so that they can be further grouped into meaningful bins for human interpretation. It is also desirable that the scaling of scores is linear. For example, for two variants with scores of 0.9 and 0.7, respectively, one should be able to infer not only that both are pathogenic, but also that the variant scored 0.9 is more likely to be pathogenic. However, for variants with scores of 0.82 and 0.80, such an interpretation would be problematic. Intuitively, one should also be able to interpret the difference between 0.9 and 0.7 in a manner similar to that between 0.7 and 0.5; i.e., quantitative differences should reflect qualitative differences.

To visualize this, we applied MutPred2, PolyPhen-2, SIFT, FATHMM, and CADD to missense variants from ten randomly selected presumably healthy individuals from the 1000 Genomes Project[35], and plotted the resulting score distributions (Fig. 2c). We found that the predictions for PolyPhen-2 and SIFT tended to peak at the extremes of the distribution, with ~10% of the variants in a healthy genome having scores over 0.9. The predictions for FATHMM and CADD peaked toward the middle of the score range with nearly half of their predictions in the region of unknown significance. On the other hand, MutPred2 outputs a generally decreasing score distribution (0.3% of variants with scores >0.9), that we believe is better suited for the interpretation of personalized genome-scale data, as it allows for the treatment of scores approximately as probabilities.

We also investigated whether these score distributions were impacted by minor allele frequencies (MAFs). We found that MutPred2 scores showed a stronger negative correlation with MAFs (Supplementary Fig. 4), further suggesting that MutPred2-scores align better with theoretical expectations of the allele frequencies of slightly deleterious variants.

Estimating proportions of pathogenic missense variants. The set of unlabeled substitutions in the MutPred2 training data comprises both benign and pathogenic variants that have not yet been characterized or annotated as such. This is also the case for the set of substitutions labeled as pathogenic, as a consequence of possible errors and misannotations in our positive set. To quantify these proportions in our training set, we generated MutPred2 score distributions on these sets and applied the AlphaMax algorithm[13]. On our training set, we estimate that 5.8% of the unlabeled variants may indeed be pathogenic and that 2.8% of the pathogenic variants may actually not be associated with disease (Supplementary Fig. 5). Using these class prior estimates as proxies for those occurring in nature, we then estimated the average proportion of pathogenic missense variants over the genomes of the ten individuals from the 1000 Genomes Project. We estimated that, on average, up to 1.3% of all missense variants in a (presumably) healthy genome are pathogenic. This fraction was greater for heterozygous variants (1.7%) than for homozygous variants (0.6%).

Enriched molecular mechanisms in Mendelian diseases. We sought to identify preferentially disrupted mechanisms in the MutPred2 training set by calculating the enrichment of the increase (i.e., gain) and decrease (i.e., loss) of local structural and functional properties in the set of disease mutations relative to the unlabeled variants (FPR of 1%). We recapitulated previous observations that protein structural changes are common mechanisms of Mendelian disorders (Fig. 3)[36,37]. However, we also observed enrichment for macromolecular binding sites, in agreement with recent work on protein–DNA and protein–protein interactions (PPIs)[38], as well as several PTM types. We observed depletion for properties associated with flexible and disordered regions, potentially owing to the enrichment of enzymes involved in metabolic processes[39]. In addition, we found that substitutions affecting residues involved in metal binding and allosteric regulation were also enriched in the set of Mendelian disease mutations.

Metals, unlike PTMs, freely form coordinate bonds without enzyme involvement and are perhaps more ubiquitous in nature. One metal ion can also be in competition with another at one or more sites in a protein due to their similar chemical properties[40]. In terms of allosteric regulation, to the best of our knowledge, a sequence-based predictor of allosteric residues does not exist, and only one structure-based study has systematically investigated the role of allosteric regulation in monogenic diseases[36]. Contrary to the findings of that study, we predict that the disruption of allosteric sites is an active mechanism in such diseases. Allosteric regulation and metal binding are treated as both structural and functional properties in our ontology (Fig. 1b). This is supported by the fact that metals are known to play important roles in

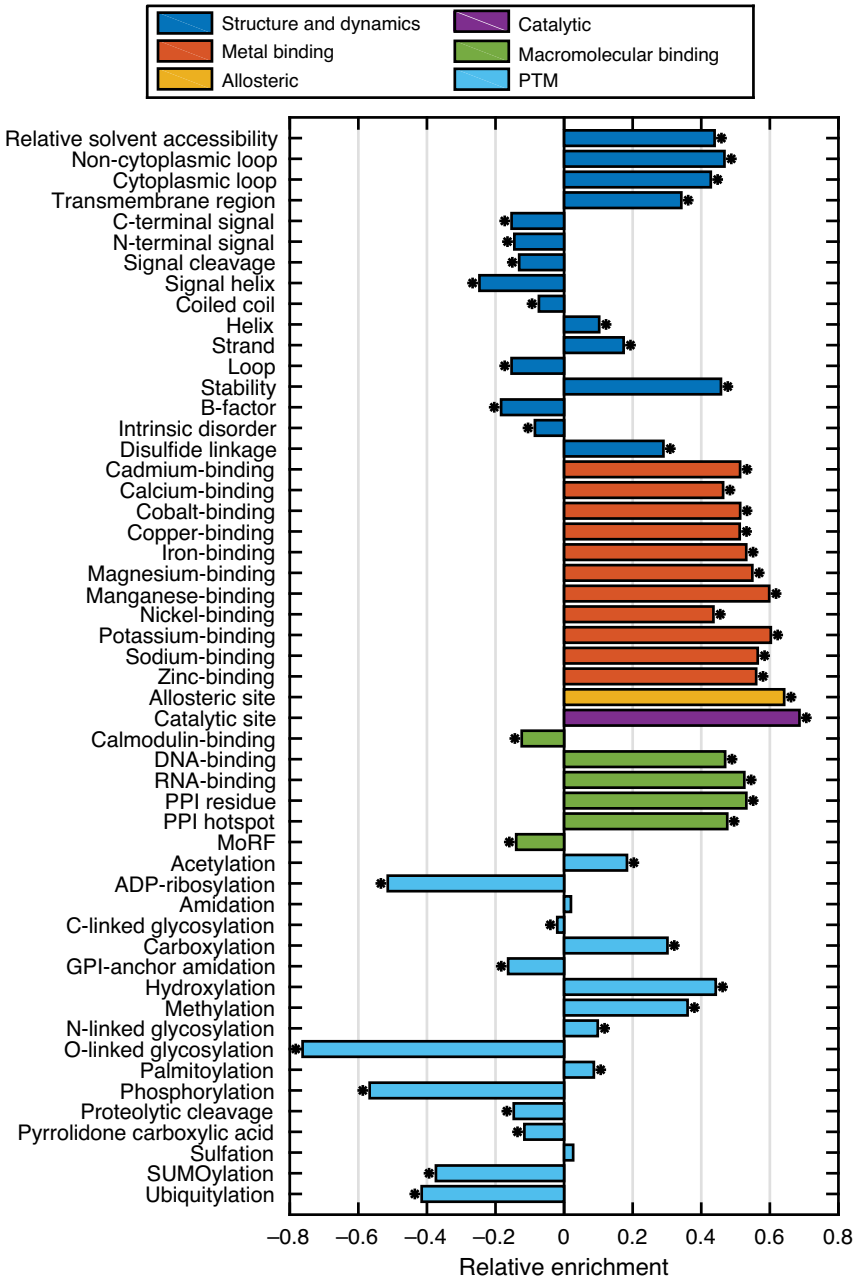

Fig. 3 Significantly enriched and depleted pathogenic mechanisms predicted by MutPred2. The data set consisted of 53,180 inherited disease mutations and an unlabeled set of 205,303 variants. Losses and gains are plotted together by considering the maximal effect for a given mutation position. An asterisk indicates significance at the 0.05 level with Benjamini–Hochberg correction, as computed by a one-sided Fisher's exact test.

stabilizing both protein structure and macromolecular interactions[40]. Further details of the enrichment analysis are provided in Supplementary Notes (Supplementary Fig. 6).

Impact of de novo mutations in neurodevelopmental disorders. Neurodevelopmental disorders have a strong genetic component[41–43]. Recent whole-genome and whole-exome sequencing in neurodevelopmental diseases has identified thousands of de novo mutations in patients with such phenotypes. However, distinguishing between pathogenic and benign de novo mutations remains challenging. We applied MutPred2 to a data set of 4324 de novo missense mutations obtained through the exome sequencing of families impacted by four different disorders (autism spectrum disorder (ASD), intellectual disability, schizophrenia, and epileptic encephalopathy), and 1316 de novo missense mutations from the healthy siblings of the patients from these families as a control ("Methods").

We first examined whether pathogenicity scores alone were sufficient to distinguish between cases and controls. We found that MutPred2 predicted significantly higher proportions of pathogenic mutations in the case set than in the control set at 10 and 5% FPR score thresholds with odds ratios of 1.44 and 1.56, respectively (Fig. 4a). Statistically significant odds ratios exceeding 1.22 were observed starting at a score threshold of 0.45 (Supplementary Fig. 7). Given the fact that the overall mutational load for de novo missense mutations is similar in the cases and the healthy controls[44,45], the higher fraction of predicted pathogenic missense mutations in the cases suggests the good discriminative ability of MutPred2. Low odds ratios are not unexpected, as missense variants are likely to be less disruptive to

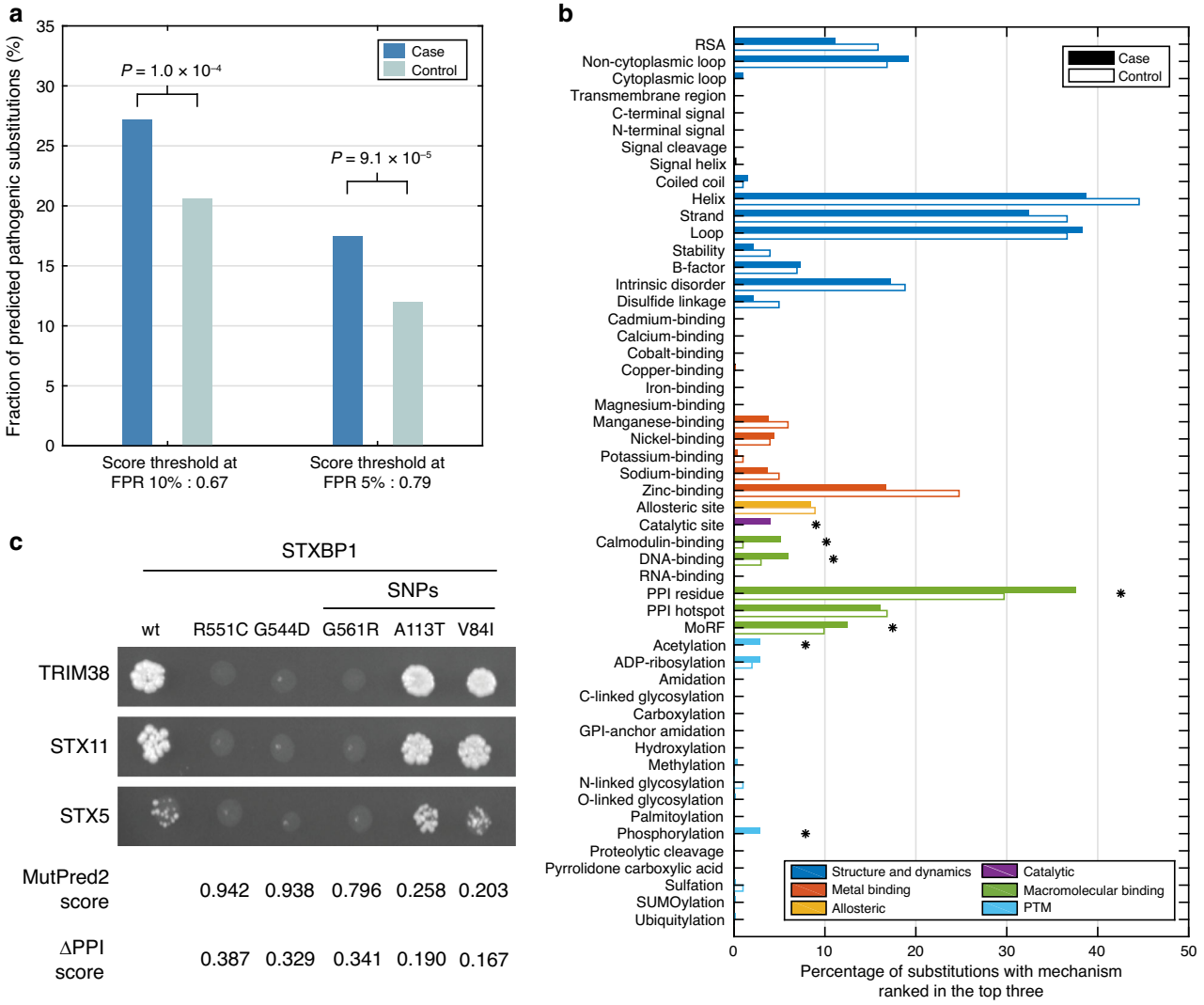

Fig. 4 Summary of MutPred2 predictions on de novo missense mutations from four neurodevelopmental disorders. After removing genes with mutations shared by both case and control sets, subsequent analyses were based on 2986 and 844 mutations from the case and control sets, respectively. **a** Proportions of case and control mutations predicted to be pathogenic by MutPred2 at thresholds corresponding to false-positive rates (FPR) of 10% and 5%, respectively. *P* values were computed using a two-tailed Fisher's exact test. Odds ratios and *P* values for other thresholds are shown in Supplementary Fig. 7. **b** Enrichment of structural and functional signatures of case mutations versus the control group. Only those mutations considered to be pathogenic at the 5% FPR threshold were included in this analysis. Properties are grouped based upon their broader classes as described in the ontology (Fig. 1b). Statistical significance was assigned at $\alpha = 0.05$ using a one-sided binomial test with Benjamini–Hochberg correction and represented by asterisks. **c** Representative images of 3AT selection plates with the interaction profiles of STXBP1 against TRIM38, STX11, and STX5. Pathogenicity scores and probability of alteration of protein–protein interactions (PPIs) corresponding to each mutation are shown. The PPI alteration probability of >0.5 × (1 − 0.5) = 0.25 is considered to be high scoring.

protein structure and function than loss-of-function (stop, frameshifting indel, and splice site) variants, for which a twofold enrichment in cases versus controls has been previously observed for autism[44].

Next, we examined whether mutations in neurodevelopmental disorders preferentially alter specific protein structural and functional properties. We asked which molecular mechanisms were frequently ranked among the top three in the set of predicted pathogenic mutations (at the 5% FPR threshold) from the case and control sets (Supplementary Data File 1). In contrast to the set of Mendelian disorders, we observed a statistically significant enrichment of the majority of macromolecular binding features (e.g., calmodulin binding, DNA binding, and protein binding), catalytic sites, and two types of PTMs (acetylation and phosphorylation) among the case mutations (Fig. 4b). The most significant enrichment was observed for the PPI residue feature, in agreement with previous studies demonstrating the loss of PPIs, as a result of mutations associated with human Mendelian diseases[46–48]. At the same time, well-defined structure and metal binding were not significantly enriched among case mutations, despite high proportions of secondary structure elements (helix, strand, and loop) being affected by the mutations in both cases and controls (Fig. 4b).

Experimental validation of the impact of de novo mutations. High pathogenicity scores for a given mutation provide hypotheses of disruption of protein structure or function that could lead to a disease. We used the yeast two-hybrid (Y2H) system to test the impact of our high-scoring mutations on PPIs, with corresponding binding partners[46,47].

We selected three genes with de novo missense mutations: *STXBP1*, *ZBTB18*, and *DNMT3A*, and introduced the high-scoring mutations into the open reading frame (ORF) clones of these genes. We then tested both wild-type (wt) and mutant variants for interactions with protein partners that were available in the human hORFeome collection[49] ("Methods"). We also tested common single-nucleotide polymorphisms (SNPs) from dbSNP adjacent to the de novo mutations, as experimental controls. In total, we tested ten mutations from these three genes (Supplementary Table 13).

Two mutations, R551C[50] and G544D[51], and one SNP (G561R) in the *STXBP1* gene disrupted interactions between STXBP1 and three protein partners (TRIM38, STX11, and STX5), whereas two other SNPs (A113T and V84I) did not alter interaction patterns compared to the wt protein (Fig. 4c). These results are in agreement with the MutPred2 predictions that assigned high scores to both R551C (score 0.942) and G544D (score 0.938), and also to the G561R (score 0.796) SNP. Interestingly, although the G561R variant is present in dbSNP and has unknown clinical significance, it is a rare variant[52] (Supplementary Table 13). STXBP1 is a syntaxin-binding protein that plays a role in the release of neurotransmitters via regulation of syntaxin, a transmembrane attachment protein receptor. A recent study demonstrated that the heterozygous loss of *STXBP1* in human neurons lowers the level of its protein product together with its binding partner syntaxin-1, emphasizing the importance of this PPI[53]. STXBP1 is highly expressed in the brain, and other mutations in *STXBP1* are associated with early infantile epileptic encephalopathy[51], intellectual disability[54], and developmental delay[55]. Our results suggest that the underlying mechanism of the mutations tested here could be attributable to loss of binding to protein partners.

We then examined the effect of the R486G mutation[54] and three SNPs (G416R, A502T, and T507A) on the binding of ZBTB18 with two partners: CTBP1 and CTBP2 (Supplementary Fig. 8a). Although the experimental results for R486G (score 0.932), G416R (score 0.668), and T507A (score 0.069) were in agreement with predictions, the MutPred2 score for A502T (score 0.208) did not match the observed loss of binding (Supplementary Fig. 8a). A502T may therefore be an undiscovered pathogenic variant missed by MutPred2 or alternatively it is a PPI-altering variant that does not lead to a disease phenotype. Finally, we tested the R635W mutation[56] in the *DNMT3A* gene (Supplementary Fig. 8b). The prediction result for R635W (score 0.886) agreed with experimental observation of the loss of interaction with the TCL1A partner.

Overall, MutPred2 predictions agreed well with experimental observations. It is, however, important to interpret these results with caution, because the loss of PPIs may not be the primary mechanism for pathogenicity, which could also result from the loss of stability, aberrant folding, increased degradation, and from other relevant mechanisms of protein structure disruption by a mutation.

Discussion

An individual person's genome contains ~10,000 amino acid-altering variants when compared to the reference genome[35]. The first step in connecting this information with phenotype, and particularly with disease, involves prioritizing variants that affect protein structure and function. The current generation of pathogenicity prediction methods has enabled the reduction of a large number of hypothetically deleterious variants to hundreds of possible candidates. However, these numbers still remain prohibitively large for subsequent experimental characterization, even considering high-throughput studies in vitro or new CRISPR/Cas9 technologies. To address this challenge, we developed MutPred2, a tool for the inference of the structural, functional, and phenotypic consequences of coding variants. By modeling the effects of variants on local protein structure and function, MutPred2 improves pathogenicity prediction and assigns putative molecular alterations, using the proposed ranking approach. This additional information can be used to accelerate experimental validation. The assignment of specific molecular impact also allows one to quantify molecular signatures within different data sets; e.g., classes of disease, a specific disease, a healthy population, a subpopulation, among others.

Previous studies have developed predictors trained on the data sets that were filtered based on MAFs and/or source of data[23,57]. Our results suggest that such filtering is beneficial only when it is directly relevant to the prediction task at hand (Supplementary Table 7); in fact, we observe that the model trained on the entire data set without filtering performs well across all prediction tasks, thereby reducing the need for specialized models. We attribute this good generalization performance to the availability of more data (the data set size decreases drastically when frequency-based filtering is applied) and reduced ascertainment bias.

There has been debate about the precise fraction of missense variants in an individual's genome that contributes to disease. Estimates for the proportion of missense variants in a genome that have deleterious effects on protein function, and hence phenotype, have ranged between 10 and 25%, depending on the operationalization of pathogenicity[58–60]. However, other works have suggested that this proportion could be much lower[61–63]. In general, the calculation of this fraction on real data has been confounded by the use of small and biased data sets, the accuracy of the underlying pathogenicity prediction method, the need to make estimates at a predetermined FPR, the limitations of simplifying assumptions on the parameters of the relevant theoretical framework, and differences in terminology.

The combined use of MutPred2 and AlphaMax[13] on exome-scale data allowed us to address these issues in a rigorous manner, with few assumptions. We established that, on average, ~100 heterozygous and 25 homozygous variants in an individual may cause disease, on a par with estimates derived using disease-causing variants from HGMD[63]. Although these numbers are large enough to yield disease phenotypes, our estimates do not account for epistatic interactions, such as compensatory variants[64]. However, our estimates are the lowest among those derived directly from data generated thus far, and support the views of early studies[62,64].

From a practical perspective, the extent of noise in current training sets for pathogenicity prediction is also important. Our results suggest that incorrectly labeled pathogenic variants constitute a small fraction of our training set and may not seriously impact predictive performance. Our estimates also suggest that ~10,000 pathogenic amino acid substitutions in dbSNP and UniProt may currently be unannotated. We believe this result is reasonable, considering that we did not filter out rare variants in our unlabeled set. However, it is important to note that our work does not address the issues of bias in current training sets[13,65].

Whole-exome sequencing studies in autism and other neurodevelopmental disorders demonstrated an excess of de novo mutations in individuals affected by these disorders[44,45,50,54,66–69]. Furthermore, several studies have demonstrated that the protein products of the genes with de novo mutations physically interact and form tightly connected protein interaction networks[70–72]. However, how specific mutations impact interactions between proteins, and which mutations are pathogenic remains an open question.

MutPred2 predicts more pathogenic de novo missense mutations in cases than in controls. This is particularly remarkable

given that brain-relevant information that could increase the predictor's performance for this type of disease was not exploited. In fact, the only filtering step involved the removal of genes common to both the case and control sets. Although the log odds are moderate, each missense variant can now be associated with potential molecular alterations that should prompt further investigation (Supplementary Data File 1). For example, the L834P mutation in CHD8 identified in a patient with autism[73] is predicted to disrupt a catalytic site along with the allosteric site and PPI residue. The mutation K842R nearby has been shown to abolish the ATPase activity of CHD8 (ref. [74]), which is consistent with the MutPred2 prediction of catalytic site disruption by an adjacent L834P mutation. Likewise, the M2679T mutation in the RYR3 calcium channel identified in an autism patient[69] is predicted to disrupt calmodulin binding along with a loss of helical propensity. By similarity with other ryanodine receptors, RYR3 probably binds calmodulin at its C-terminus, and the prediction of the loss of calmodulin binding due to the M2679T C-terminal mutation concurs with the known function of this protein. Thus, MutPred2 predictions offer viable biological hypotheses that can be tested in the laboratory to improve our understanding of disease mechanisms.

Traditionally, researchers have adopted a top-down or disease-driven approach, where one starts with specific phenotypes and works one's way down toward causes at the molecular level. MutPred2 enables the adoption of a bottom-up or mechanism-driven approach toward understanding genetic disease. In this approach, one can envision experts specializing in molecular mechanisms studying germline and somatic variants across different diseases, and providing functional insights that can subsequently lead to hypotheses at the phenotypic level[75,76]. We loosely refer to this mechanism-driven approach as disease agnostic because the study and validation of impactful variants is determined by molecular mechanisms one is equipped to study, not necessarily the high-level phenotype.

By grouping disease classes together based on frequently affected molecular mechanisms in current data sets, one can consider the prospect of identifying common targets and repurposing drugs from one class of disease to its neighboring disease in this new space. For example, MutPred2 predicted a close relationship between the endocrine and immune systems at the molecular level (Supplementary Fig. 6). This agrees with observations related to the interactions between the two systems during ontogeny[77].

Methods

An input to MutPred2 is an amino acid sequence s; i.e., a wt protein sequence, and an amino acid substitution XiY, where X, the ith amino acid in s, is replaced by Y. We refer to the mutated (mt) sequence as s_{XiY}. The output of MutPred2 consists of a pathogenicity score, a number from [0,1], and a list of molecular mechanisms, each with its own score, that may be impacted by XiY. A pathogenicity score of 1 indicates near certainty that the variant is pathogenic, whereas a score of 0 indicates near certainty that the variant is benign. In the next several sections, we discuss data sets, data representation, and training of MutPred2. The details regarding classification models used to assess specific functional impacts are provided in Supplementary Methods.

Data sets. A data set of pathogenic amino acid substitutions was created by integrating HGMD[17] (June 2013; "DM"-annotated substitutions only), Swiss-Prot (release 2012_09 through SwissVar[18]), and dbSNP[19] (build 137). The set of unlabeled (putatively neutral) substitutions was compiled from Swiss-Prot and dbSNP, and then supplemented with additional variants in a way similar to the HumDiv training set for PolyPhen-2 (ref. [23]): for every human protein, pairwise alignments to other mammalian proteins were first extracted from a 46 species multiple sequence alignment, obtained from the UCSC Genome Browser[78]. Only those alignments where the two sequences shared at least 99% sequence identity were considered, and positions where a residue in the nonhuman sequence was replaced by a different one in the human sequence were identified.

Data representation and training. Given a sequence s and variant XiY, we extracted 1345 (including 20 optional) features. These features are subcategorized into six groups: (1) sequence-based features, (2) substitution-based features, (3) PSSM-based features, (4) conservation-based features, (5) homolog profiles (optional due to time necessary to compute), and (6) changes in predicted structural and functional properties. A detailed list of features and how they were extracted and encoded is provided in Supplementary Data File 2 and Supplementary Methods, respectively. Feature selection using a two-sample t test was performed and only those features that returned P values < 0.01 were retained. To remove (near) colinear features, z-score normalization and principal component analysis were performed on the selected features, with the retained variance set to 99%. An ensemble of 30 feed-forward neural networks was then trained on the resulting feature matrix. Each network consisted of a single hidden layer with four neurons and a single output neuron (the hyperbolic tangent activation function was used in both layers). A bootstrap aggregating (bagging) approach was adopted for training, where each network was trained on a balanced random sample (with replacement) of the original training set. To determine the number of iterations required for training, 25% of the training data were retained as a validation set. The final model was trained using the resilient propagation algorithm[79] and stopped when, either this optimal number of iterations was reached, or 1000 epochs were completed, or 500 validation checks were reached. Prediction scores were then calculated as the mean of all 30 scores.

Inferring molecular mechanisms of pathogenicity. The local effects of a variant on predicted structural and functional properties were modeled and utilized, both as features and for the assignment of putative molecular mechanisms. First, over 50 protein property predictors were developed within a unified positive-unlabeled learning framework (Supplementary Methods). The wt sequence s was first provided to these predictors to score the substitution site i and ±5 adjacent positions. The amino acid substitution was then introduced into the sequence in silico and the mt sequence s_{XiY} provided to all property predictors. The probabilities of changes in structural and functional propensities, given the substitution XiY, were calculated from the property predictors as follows[10]

$$\begin{aligned}\Pr(\text{loss of property } p|s, XiY) &= \Pr(\text{presence of property } p|s)\\ &\quad\cdot \Pr(\text{absence of property } p|s_{XiY})\\ &= \Pr(P=1|s)\cdot(1-\Pr(P=1|s_{XiY})),\end{aligned} \tag{1}$$

$$\begin{aligned}\Pr(\text{gain of property } p|S, XiY) &= \Pr(\text{absence of property } p|s)\\ &\quad\cdot \Pr(\text{presence of property } p|s_{XiY})\\ &= (1-\Pr(P=1|s))\cdot\Pr(P=1|s_{XiY}),\end{aligned} \tag{2}$$

where P is the random variable indicating presence (1) or absence (0) of property p. In the above equations, the wt residue at the ith position of the protein is X and the replacement amino acid is Y. Appropriate transformations were applied to ensure that the property predictors accurately approximate posterior distributions (Supplementary Methods). The posterior probability from the predictor for property p for the wt sequence can be interpreted as $\Pr(P=1|s)$, and the posterior probability for the substituted sequence can be interpreted as $\Pr(P=1|s_{XiY})$.

The property score was interpreted as the posterior probability of loss or gain, whichever was greater. Naively, if the wt posterior is 0.5 and the mutant posterior is 0.5 (i.e., no effect upon substitution), then the loss and gain probabilities will be 0.25, which we treated as a baseline threshold to implicate the property as a molecular mechanism in disease. It is important to note that the terms loss and gain are more appropriately interpreted as decreased and increased propensities for a certain property, respectively. Furthermore, in the case of properties that can be affected in both directions due to a single-residue change, interpretation becomes complicated. For instance, a substitution can increase a protein's propensity to bind one protein partner, but decrease its propensity for another. For simplicity, the term altered is reported in MutPred2 predictions for such properties along with the maximum of the loss and gain score.

In addition to posterior probabilities of loss and gain, we also provide empirical P values that the observed loss/gain score is as high or higher than the score randomly generated from the distribution of putatively neutral substitutions. The lower this P value, the more likely that the predicted property is involved in pathogenicity, under the assumption that nonpathogenic variants do not affect protein structure and/or function.

Predictor evaluation. All predictors were evaluated through per-protein cross-validation experiments. Unless otherwise noted, the training data for each predictor was first split into ten randomly generated partitions, such that all data points from a given protein were in the same partition. Then, in an iterative manner, each partition was treated as the test set and the remaining nine partitions constituted the training set. To avoid information leak, feature selection, normalization, and dimensionality reduction parameters were obtained on the training partition, and then applied to the test partition within each iteration. After ten iterations, every data point was assigned a prediction score. These scores were then used to estimate the accuracy of the model.

Evaluation of MutPred2 was performed in a similar manner, except that the cross-validation partitions were defined more stringently. Instead of a per-protein partition definition, a more stringent per-cluster partition definition was adopted as proposed by Calabrese et al.[31]. Protein sequences in the data set were first clustered using CD-HIT at the 50% sequence identity threshold. We then ensured that all substitutions from the same cluster were either entirely in the training set or entirely in the test set.

Independent validation and comparison. For the purposes of additional evaluation and comparison with other methods, an independent test set was compiled from mutations deposited in ClinVar[32] (March 5, 2015) and UniProt[80] ("humsavar.txt", release 2015_04). Fragments of length 25 residues centered at the mutation position were extracted and compared to similarly constructed fragments from the training sets of five methods (MutPred2, MutPred, PolyPhen-2 (both models), SNPs&GO[31], and FATHMM[27]), using CD-HIT-2D[81]. In the case of FATHMM, the additional "humsavar" data set that it was tested on (with similar performance) was used, because FATHMM training sets were not publicly available. All mutations whose local neighborhoods shared at least 50% sequence identity with those from at least one of the training sets were filtered out. Predictions for MutPred2, MutPred, and PolyPhen-2 were obtained using locally installed versions of the software. Scores for SNPs&GO were obtained through multiple queries to its web server. The FATHMM scores, along with predictions for other methods, such as CADD[26], SIFT[22], MutationTaster2 (ref. [29]), GERP++[28], and PhyloP[30] (20-way) were directly obtained from the dbNSFP database[82] (v3.0) of all possible single-nucleotide substitutions. In cases where the chromosomal positions in this database mapped to multiple protein positions, one-to-one correspondence of isoforms (and positions) was verified. Although other methods for the prediction of pathogenicity exist, we chose this representative set based on recommendations recently made by the ACMG/AMP[11]. The entire procedure remained the same for the threshold of 80% sequence identity.

Score distributions on genomes. Two individuals from each of the five super populations represented in the 1000 Genomes Project[35] (phase 3) were randomly selected, such that they came from different populations. In total, variants for ten genomes were extracted from the integrated variant call format files: NA19026, HG02014, HG02002, HG01075, HG02384, NA18632, NA12829, HG01615, HG04206, and HG02651. ANNOVAR[83] was used to identify and retain non-synonymous single-nucleotide substitutions, map their coordinates to amino acid positions in protein sequences, extract their zygosity information, and obtain MAFs from the ExAC browser[84]. The "coding_change.pl" program in ANNOVAR was used to obtain protein sequences for MutPred2. Both MutPred2 and PolyPhen-2 (both models) were locally installed and run on this data set. Scores for SIFT, FATHMM, and CADD were obtained from the dbNSFP database and were transformed to ensure the scores fit within the 0–1 range. For SIFT, all scores were subtracted from one. For FATHMM and CADD, min–max transformation was performed. For FATHMM, signs of the scores were changed to ensure that higher scores indicated pathogenicity. The resulting scores were binned into fixed intervals for each individual separately. The mean fraction of variants within each bin and its standard error over all ten individuals were then plotted. Similarly, in the case of MAFs, the mean allele frequency and its standard error were plotted.

Enrichment of properties in the MutPred2 disease set. Frequently altered properties in the set of all disease variants in MutPred2's training set (Fig. 3) were identified by first deciding a threshold for loss and gain scores based on a predetermined FPR (here, 1%). Assuming that the vast majority of the nonpathogenic substitutions do not affect protein properties, one can use the fraction of these substitutions with a score greater than or equal to the threshold to approximate the rate. For instance, if an FPR of 5% is desired and if 5% of nonpathogenic variants have a loss score of 0.4 or greater, then the threshold would be 0.4. Based on this threshold, the numbers of disease-causing variants with and without the given mechanism were then counted. Thus, two proportion values were obtained: one for the fraction of disease-causing variants affecting the property (d_f), and one for the fraction of nonpathogenic variants affecting the property (n_f). Then, the enrichment E was calculated as:

$$E = \frac{d_f - n_f}{d_f + n_f}. \quad (3)$$

If E was positive, the property was considered to be enriched, and if it was negative, the property was treated as depleted in the disease set. Significance to these enrichment/depletion values was assigned using a one-sided Fisher's exact test with subsequent correction for multiple hypothesis testing, using the Benjamini–Hochberg method[85]. Since PTMs are known to occur on specific residues, these were further divided into two separate categories when generating the counts: when a substitution occurred at the predicted PTM site exactly, and when it occurred in its neighborhood. Note that although this data set was dominated by mutations from HGMD, the diseases covered are not strictly monogenic. Nevertheless, we refer to this set as the data set of Mendelian diseases.

Analyses on neurodevelopmental disorder mutations. A data set of 4324 de novo mutations identified through whole-exome or whole-genome sequencing of the individuals diagnosed with ASD, intellectual disability, epileptic encephalopathy, and schizophrenia[44,45,50,51,54,56,66,68,69,73,86–101], along with a control set of 1316 de novo mutations from the healthy siblings[44,45,54,66,69,91,95,97,99,102] was curated from the published literature. Genes with mutations shared by both case and control sets were removed from the analyses. Unlike the MutPred2 disease set, there was no prior knowledge of which mutations in the case and control sets were pathogenic or benign. Therefore, MutPred2 pathogenicity scores (at the 5% FPR threshold; score of 0.79) were used to divide each set into pathogenic and benign mutations. Only the mutations above this score threshold were considered for further analyses. To identify structural and functional signatures for each substitution, property scores were ranked in decreasing order. Then, the fractions of substitutions with a given property in the top three were compared between the case and control sets using a one-sided binomial test. The resulting P values were then FDR corrected using the Benjamini–Hochberg method.

Yeast two-hybrid assays. Candidate genes for experimental validation were selected based on their MutPred2 scores and the availability of the corresponding clones in the human ORFeome v8.1 collection[49]. An additional constraint was placed: a PPI-related mechanism had to be one of the top ten properties disrupted. The common SNPs in close proximity to the potential disease mutation were extracted from dbSNP. We also verified that the selected SNPs are not present in ClinVar. All genes are stored in ORFeome v8.1 in the pDONR223 vector.

Missense mutations were introduced into the ORFs by the site-directed mutagenesis using PCR overlap[103]. The M13 primers were used as the flanking primers for the PCR overlap reactions, and the sequences for the forward (5′–3′ direction) primers used for the site-directed mutagenesis were as follows:

STXBP1 R551C: GAGCCTGAATGAGATGTGCTGCGCCTACG AGGTG;
STXBP1 G544D: CATTTTCATCCTTGGGGATGTGAGCCTGA ATGAG;
STXBP1 G561R: GTGACCCAGGCCAACAGAAAGTGGGAG GTG;
STXBP1 A113T: CTGACTCTTGTCCAGATACCCTGTTTAA TGAACTG;
STXBP1 V84I: CATCCGAGAAGTCCATCCACTCTCTCATC;
ZBTB18 R486G: GTACAGCTCGGTGGTCTCGGAACTGGGCA TCTCC;
ZBTB18 G416R: GTGCTCGCTGTGTAGGAAGACTTTCTC;
ZBTB18 A502T: GGTCAAAAGCGAAACACTGAGCTTGCC;
ZBTB18 T507A: CTGAGCTTGCCTGCTGTCAGAGACTG;
DNMT3A R635W: GAGAAGAGGAAGCCCATCTGGGTGCTG TCTCTCTTTG.

The conditions for the PCR reaction were as follows: 94 °C for 5 min, 30 cycles; 94 °C for 30 s, 55 °C for 30 s, 68 °C for 1 min in the PCRs 1 and 2, and for 2 min and 30 s in the PCR 3; and lastly 68 °C for 7 min. The resulting mutant ORFs were Gateway cloned into the pDONR223 vector, and verified by Sanger sequencing. Then, the mutant ORFs were Gateway cloned into the pDB_DEST vector and transformed into yeast for pairwise interactions testing. Briefly, miniprep plasmid DNA of all DB-X clones, both *wt* and mutant constructs, were transformed into the Y2H strain MATαY8930. The interacting partners of the *wt* proteins were extracted from BIOGRID[104]; only the partners that are present in the human ORFeome v8.1 collection were subsequently tested for interactions. The binary PPI Y2H screens of all DB-X baits against AD-Y preys (i.e., partners) were performed[70,105]. Briefly, the DB-X and AD-Y clones were mated in YEPD media for 24 h, and then plated on Sc-Leu-3AT and Sc-Leu-His-CHX plates (i.e., test for autoactivation) for selection. Only colonies that grew on the Sc-Leu-3AT plate, but not on the Sc-Leu-His-CHX plates were counted as positives. All the pairwise Y2H screens were repeated three times in independent experiments, and only interactions that scored as positives at least twice were considered as positives (Supplementary Fig. 9).

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

Data availability

A modified version of the training data for MutPred2 that does not contain variants exclusively from HGMD is available at http://mutpred.mutdb.org/#dload. Since we used the professional version of HGMD, restrictions apply to the availability of these data. These variants are however available under the appropriate license agreement at http://www.hgmd.cf.ac.uk/ac/index.php.

Code availability

The source code for MutPred2 is available at https://github.com/vpejaver/mutpred2. Trained models and other data files relevant to the source code are hosted at http://mutpred.mutdb.org. A web server and standalone version of MutPred2 are also available at http://mutpred.mutdb.org and http://mutpred.mutdb.org/#dload, respectively.

Received: 23 July 2019; Accepted: 23 October 2020;
Published online: 20 November 2020

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. 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).
3. Niroula, A. & Vihinen, M. Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.* **37**, 579–597 (2016).
4. Kumar, S., Sanderford, M., Gray, V. E., Ye, J. & Liu, L. Evolutionary diagnosis method for variants in personal exomes. *Nat., Methods* **9**, 855–856 (2012).
5. Miosge, L. A. et al. Comparison of predicted and actual consequences of missense mutations. *Proc. Natl Acad. Sci. USA* **112**, E5189–E5198 (2015).
6. Rost, B., Radivojac, P. & Bromberg, Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* **590**, 2327–2341 (2016).
7. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
8. Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C. & Friedberg, I. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol.* **9**, e1003063 (2013).
9. Olsen, J. V. & Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell Proteom.* **12**, 3444–3452 (2013).
10. Li, B. et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744–2750 (2009).
11. 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).
12. Menon, A. K., van Rooyen, B., Ong, C. S. & Williamson, R. C. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37, 125–134 (ICML, Lille, France, 2015).
13. Jain, S., White, M. & Radivojac, P. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, 2693–2701 (NIPS, Barcelona, Spain, 2016).
14. Denis, F., Gilleron, R. & Letouzey, F. Learning from positive and unlabeled examples. *Theor. Comput. Sci.* **348**, 70–83 (2005).
15. Latinne, P., Saerens, M. & Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing. In *Proceedings of the Eighteenth International Conference on Machine Learning,* 298–305 (ICML, Williamstown, MA, USA, 2001).
16. Vihinen, M. Variation Ontology for annotation of variation effects and mechanisms. *Genome Res.* **24**, 356–364 (2014).
17. Stenson, P. D. et al. The Human Gene Mutation Database (HGMD(®)): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**,1197–1207 (2020).
18. Mottaz, A., David, F. P., Veuthey, A.-L. & Yip, Y. L. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* **26**, 851–852 (2010).
19. Sherry, S. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
20. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
21. Jain, S., White, M. & Radivojac, P. Recovering true classifier performance in positive-unlabeled learning. In *AAAI Conference on Artificial Intelligence,* 2066–2072 (AAAI, San Francisco, CA, USA, 2017).
22. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
23. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
24. Chen, W. H., Zhao, X. M., van Noort, V. & Bork, P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput. Biol.* **9**, e1003073 (2013).
25. Hsiao, T. L. & Vitkup, D. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet.* **4**, e1000014 (2008).
26. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
27. 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).
28. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
29. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
30. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
31. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30**, 1237–1244 (2009).
32. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
33. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Nucleic Acids Res.* **32**, 358–368 (2011).
34. 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).
35. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
36. Wang, Z. & Moult, J. SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270 (2001).
37. Khan, S. & Vihinen, M. Spectrum of disease-causing mutations in protein secondary structures. *BMC Struct. Biol.* **7**, 56 (2007).
38. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
39. Xie, H. et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6**, 1882–1898 (2007).
40. Dudev, T. & Lim, C. Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. *Chem. Rev.* **114**, 538–556 (2014).
41. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol.* **14**, 1109–1120 (2015).
42. Brandler, W. M. & Sebat, J. From de novo mutations to personalized therapeutic interventions in autism. *Annu. Rev. Med.* **66**, 487–507 (2015).
43. Iakoucheva, L. M., Muotri, A. R. & Sebat, J. Getting to the cores of autism. *Cell* **178**, 1287–1298 (2019).
44. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
45. Xu, B. et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
46. Sahni, N. et al. Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* **23**, 649–657 (2013).
47. Zhong, Q. et al. Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
48. Das, J. et al. Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. *Hum. Mutat.* **35**, 585–593 (2014).
49. The ORFeome Collaboration. The ORFeome Collaboration: a genome-scale human ORF-clone resource. *Nat. Methods* **13**, 191–192 (2016).
50. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
51. Allen, A. S. et al. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
52. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
53. Patzke, C. et al. Analysis of conditional heterozygous STXBP1 mutations in human neurons. *J. Clin. Investig.* **125**, 3560–3571 (2015).
54. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
55. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
56. Jiang, Y. H. et al. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
57. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**(Suppl 3), S3 (2013).
58. Sunyaev, S. et al. Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
59. Chasman, D. & Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706 (2001).

60. Lohmueller, K. E. et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–997 (2008).
61. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
62. Ng, P. C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12**, 436–446 (2002).
63. Xue, Y. et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).
64. Kondrashov, A. S. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J. Theor. Biol.* **175**, 583–594 (1995).
65. Scott, C., Blanchard, G. & Handy, G. Classification with asymmetric label noise: consistency and maximal denoising. *J. Mach. Learn. Res.* **30**, 489–511 (2013).
66. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
67. Buxbaum, J. D. et al. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052–1056 (2012).
68. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
69. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
70. Corominas, R. et al. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat. Commun.* **5**, 3650 (2014).
71. Sakai, Y. et al. Protein interactome reveals converging molecular pathways among autism disorders. *Sci. Transl. Med.* **3**, 86ra49 (2011).
72. Lin, G. N. et al. Spatiotemporal 16p11.2 protein network implicates cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. *Neuron* **85**, 742–754 (2015).
73. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
74. Thompson, B. A., Tremblay, V., Lin, G. & Bochar, D. A. CHD8 is an ATP-dependent chromatin remodeling factor that regulates beta-catenin target genes. *Mol. Cell Biol.* **28**, 3894–3904 (2008).
75. Christopoulos, A. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat. Rev. Drug Discov.* **1**, 198–210 (2002).
76. Wenthur, C. J., Gentry, P. R., Mathews, T. P. & Lindsley, C. W. Drugs for allosteric sites on receptors. *Annu. Rev. Pharmacol. Toxicol.* **54**, 165–184 (2014).
77. Besedovsky, H. & Sorkin, E. Network of immune-neuroendocrine interactions. *Clin. Exp. Immunol.* **27**, 1–12 (1977).
78. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
79. Riedmiller, M. & Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*, 586–591 (IEEE, San Francisco, CA, USA, 1993).
80. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
81. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
82. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
83. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
84. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
85. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
86. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project & 986 Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–370 (2014).
87. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
88. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
89. Girard, S. L. et al. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat. Genet.* **43**, 860–863 (2011).
90. Guipponi, M. et al. Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. *PLoS ONE* **9**, e112745 (2014).
91. Gulsuner, S. et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
92. Hashimoto, R. et al. Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *J. Hum. Genet.* **61**, 199–206 (2016).
93. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
94. McCarthy, S. E. et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol. Psychiatry* **19**, 652–658 (2014).
95. O'Roak, B. J. et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).
96. O'Roak, B. J. et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
97. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
98. Turner, T. N. et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
99. Xu, B. et al. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* **43**, 864–868 (2011).
100. Yuen, R. K. et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* **21**, 185–191 (2015).
101. Yuen, R. K. et al. Genome-wide characteristics of de novo mutations in autism. *NPJ Genom. Med.* **1**, 16027 (2016).
102. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
103. Heckman, K. L. & Pease, L. R. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat. Protoc.* **2**, 924–932 (2007).
104. Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).
105. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

Acknowledgements

We thank Prof. Matthew W. Hahn for helpful comments. This work was supported by the US National Institutes of Health (NIH) grants R01 LM009722 (S.D.M.), R01 MH105524 (L.M.I. and P.R.), R01 MH104766 (L.M.I.), R01 MH109885 (L.M.I.), Simons Foundation Grant #345469 (L.M.I.), R01 MH076431 (J.S.), K99 LM012992 (V.P.), the Indiana University Precision Health Initiative (P.R.), and Washington Research Foundation Fund for Innovation in Data-Intensive Discovery and the Moore/Sloan Data Science Environments Project (V.P.).

Author contributions

V.P. developed MutPred2, performed computational experiments, and wrote the first draft of the manuscript. J.U. performed experimental testing of predictions and analyzed the results. J.L.-M. and K.A.P. provided multiple data sets for training and validation purposes. M.M. and D.N.C. provided the HGMD data set and related metadata, and assisted with study design. G.N.L., L.M.I., and J.S. provided case and control data sets for neurodevelopmental disorders. V.P., G.N.L., and H.-J.N. designed and produced figures. V.P., J.L.-M., G.N.L., and H.-J.N. contributed to the design of experiments. L.M.I. and J.S. provided guidance on experiments and analyses related to the neurodevelopmental disorders data sets, and helped supervise the project. P.R. and S.D.M. conceived MutPred2, supervised its development. All authors contributed to the review and writing of the manuscript.

Competing interests

The authors declare the following competing interests. D.N.C. and M.M. acknowledge Qiagen Inc. for their financial support through a License Agreement with Cardiff University. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19669-x.

Correspondence and requests for materials should be addressed to L.M.I., S.D.M. or P.R.

Peer review information *Nature Communications* thanks Maricel Kann and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.