Delineation of functionally essential protein regions for 242 neurodevelopmental genes

Iqbal, S

2023-02

Iqbal , S , Brunger , T , Perez-Palma , E , Macnee , M , Brunklaus , A , Daly , MJ , Campbell , AJ , Hoksza , D , May , P & Lal , D 2023 , ' Delineation of functionally essential protein regions for 242 neurodevelopmental genes ' , Brain : a journal of neurology , vol. 146 , no. 2 , pp. 519-533 . https://doi.org/10.1093/brain/awac381

http://hdl.handle.net/10138/355366
https://doi.org/10.1093/brain/awac381

BRAIN
ORIGINAL ARTICLE

Delineation of functionally essential protein regions for 242 neurodevelopmental genes

Sumaiya Iqbal,[1,2,3,†] Tobias Brünger,[4,†] Eduardo Pérez-Palma,[5] Marie Macnee,[4] Andreas Brunklaus,[6,7] Mark J. Daly,[2,3,8] Arthur J. Campbell,[1,2] David Hoksza,[9] Patrick May[10] and Dennis Lal[2,4,11,12]

[†]These authors contributed equally to this work.

See Costain and Andrade (https://doi.org/10.1093/brain/awad011) for a scientific commentary on this article.

Neurodevelopmental disorders (NDDs), including severe paediatric epilepsy, autism and intellectual disabilities are heterogeneous conditions in which clinical genetic testing can often identify a pathogenic variant. For many of them, genetic therapies will be tested in this or the coming years in clinical trials. In contrast to first-generation symptomatic treatments, the new disease-modifying precision medicines require a genetic test-informed diagnosis before a patient can be enrolled in a clinical trial. However, even in 2022, most identified genetic variants in NDD genes are 'variants of uncertain significance'. To safely enrol patients in precision medicine clinical trials, it is important to increase our knowledge about which regions in NDD-associated proteins can 'tolerate' missense variants and which ones are 'essential' and will cause a NDD when mutated. In addition, knowledge about functionally indispensable regions in the 3D structure context of proteins can also provide insights into the molecular mechanisms of disease variants.

We developed a novel consensus approach that overlays evolutionary, and population based genomic scores to identify 3D essential sites (Essential3D) on protein structures. After extensive benchmarking of AlphaFold predicted and experimentally solved protein structures, we generated the currently largest expert curated protein structure set for 242 NDDs and identified 14 377 Essential3D sites across 189 gene disorders associated proteins. We demonstrate that the consensus annotation of Essential3D sites improves prioritization of disease mutations over single annotations. The identified Essential3D sites were enriched for functional features such as intermembrane regions or active sites and discovered key inter-molecule interactions in protein complexes that were otherwise not annotated. Using the currently largest autism, developmental disorders, and epilepsies exome sequencing studies including >360 000 NDD patients and population controls, we found that missense variants at Essential3D sites are 8-fold enriched in patients.

In summary, we developed a comprehensive protein structure set for 242 NDDs and identified 14 377 Essential3D sites in these. All data are available at https://es-ndd.broadinstitute.org for interactive visual inspection to enhance variant interpretation and development of mechanistic hypotheses for 242 NDDs genes. The provided resources will enhance clinical variant interpretation and *in silico* drug target development for NDD-associated genes and encoded proteins.

1 The Center for the Development of Therapeutics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
2 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
4 Cologne Center for Genomics, University of Cologne, 50923 Köln, Germany
5 Universidad del Desarrollo, Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana, 7610658 Las Condes, Santiago de Chile, Chile
6 The Paediatric Neurosciences Research Group, Royal Hospital for Children, Glasgow G12 8QQ, UK
7 School of Health and Wellbeing, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Received June 29, 2022. Revised August 12, 2022. Accepted September 04, 2022. Advance access publication October 18, 2022

8 Institute for Molecular Medicine Finland (FIMM), Centre of Excellence in Complex Disease Genetics, University of Helsinki, 00100 Helsinki, Finland
9 Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, 110 00 Staré Město, Czechia, Czech Republic
10 Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg
11 Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA
12 Genomic Medicine Institute, Lerner Research Institute Cleveland Clinic, Cleveland, OH 44106, USA

Correspondence to: Sumaiya Iqbal
415 Main St, Cambridge, MA 02142, USA
E-mail: sumaiya@broadinstitute.org

Correspondence may also be addressed to: Dennis Lal
9500 Euclid Ave, NE-50, Cleveland, OH 44195, USA
E-mail: lald@ccf.org

Keywords: neurodevelopmental disorder; genetics; bioinformatics

Introduction

Neurodevelopmental disorders (NDDs) are a group of congenital or early-onset conditions that affect about 2–5% of children worldwide.[1,2] NDDs are characterized by neurocognitive deficits with symptoms ranging from mild impairments, allowing those affected to live reasonably everyday lives, to severe disorders that require lifelong care.[3,4] Diverse factors such as gestational infection and maternal alcohol consumption contribute to NDDs.[5,6] However, inherited genetic variants that disrupt genes, encoding instructions for neuronal development and functioning, are major contributors to individual risk for NDDs and can, in fact, be causal for the disorder.[3,7] A few hundreds of such genes have been reported,[7-13] but most of them have only recently been identified. This novelty opens an avenue to extend the frontier of knowledge about these NDD-associated genes and their disease mechanisms; for example, identifying regions in the corresponding human proteins that are conserved for their molecular functions and should be constraint against deleterious mutations.

Previous studies have estimated that around 42–48% of patients with a severe developmental disorder carry a pathogenic *de novo* mutation in a protein-coding gene, with missense *de novo* mutations (i.e. a single nucleotide change leading to a single amino acid substitution) being more common compared to protein-truncating *de novo* mutations (PTVs; nonsense, frameshift and essential splice site variants[13]). In contrast to PTVs, interpretation of missense variants is challenging due to their variegated functional outcomes depending on the amino acid substituted and the protein domain affected. Missense variants in the same NDD-associated gene can possess a range of pathogenicity,[3] causing mild-to-severe phenotypes and often leading to multiple clinically distinct disorders due to differences in the protein's altered molecular function. For example, different pathogenic *SCN1A* variants can lead to Dravet syndrome, a severe epilepsy syndrome, or generalized epilepsy with febrile seizures plus (GEFS+), a milder epilepsy manifestation.[14,15] Molecular effects of missense variants such as gain- (GoF) or loss-of-function (LoF) can further determine the phenotype observed and even affect pharmacological treatment. For example, in *SCN2A*-related epileptic encephalopathies, GoF missense variants are associated with an earlier seizure onset and respond to sodium channel blockers. In contrast, LoF missense variants in *SCN2A* are associated with autism and do not benefit from antiepileptic drugs.[16] This varying effect of missense variants in the same NDD-associated gene complicates their clinical interpretation, which cannot fully be addressed by existing bioinformatic tools that have been fairly successful in classifying variants into discrete categories, e.g. benign and pathogenic.[17] Instead, evolutionarily and structurally informed methods for identifying important regions in protein 3D structures would be a critical help to selectively nominate positions for further experimental assays, deepening the insights into varying molecular effects of mutations, and subsequently apply the outcome to patient stratification and precision care.

It has been previously shown that 3D structural information of proteins helps prioritize mutational hotspots and can unveil perturbed biological pathways by missense variants.[18-22] For example, distinct positional clustering of missense variation on the 3D structure of sodium channel and calcium channel proteins are found to be associated with LoF and GoF mechanisms underlying different NDDs.[14,16] 3D clusters of NDD-associated missense mutations in the GTP-binding domain of the GNAO1 protein highlight the importance of G-protein signalling in neurodevelopment.[23] Mutations in the catalytic and regulatory 3D sites of the CDKL5 structure, affecting the phosphorylation signalling pathways, are shown to be implicated in CDKL5-related NDDs.[24,25] This evidence, along with missense variants being predominantly causal for NDDs,[13] suggests that it will be paramount to generate a resource of protein structures with annotation of important 3D sites, underpinned by genomic indications (e.g. known positions of pathogenic variants), for a comprehensive collection of NDD-associated genes, which is precisely the aim of our study.

Recently, DeepMind's neural network-based method, AlphaFold,[26] was shown to be able to predict the 3D structure of proteins at an accuracy matching experimental methods, and subsequently, predicted structures for the entire human proteome were deposited in the AlphaFold protein structure database.[27] Facilitated by this unprecedented resource and the experimental structures available in the Protein Data Bank,[28] here, we generated a new consensus annotation of protein residues in 3D—referred to as '3D essential sites' (Essential3D)—that are conserved across human gene paralogues, intolerant of missense variants and enriched for pathogenic variants. We made all experimental and predicted structures of 242 NDD-associated proteins annotated with Essential3D sites publicly available (https://es-ndd.broadinstitute.org), which will facilitate studies, both on individual NDD target genes and on a large scale, to generate testable hypotheses on the perturbed biological functions by *de novo* variants and their mechanism of action in NDDs.

Materials and methods

Protein structure selection and filtering

Experimentally solved structures of 185 NDD proteins (out of 242) were collected from the Protein Data Bank (PDB)[28] and were filtered based

on two criteria: (i) coverage of >30% of the full protein sequence; and (ii) >100 amino acid residues in the structure. We analysed 2461 structures of 154 NDD proteins out of 2715 structures of 185 proteins that met these criteria. Predicted structures obtained from the AlphaFold database were analysed only when the structure of the full-length protein was available as a single file, precluding proteins longer than 2700 amino acid residues. Structures of 230 NDD proteins (out of 242) satisfied this criterion and were selected for this study. Three sets of structures were generated from the available structures: (i) monomeric-PDB: monomeric structures from PDB; (ii) multimeric-PDB: protein complexes from PDB; and (iii) monomeric-AlphaFold: monomeric structures from AlphaFold (Supplementary Table 1).

Missense variant collection

Canonical transcripts for the 242 NDD genes were accessed from the UniProt database.[29] All protein-coding missense ariants were collected for these canonical transcripts. All variants refer to the human reference genome GRCh37.p13/hg19.

Selection of variants for the generation of Essential3D site annotation

The general 'population' missense variants from the genome aggregation Database (gnomAD, public release 2.0.2)[30] were downloaded as Variant Call Format (VCFs)[31] files (http://gnomad.broadinstitute.org/downloads). The extraction of missense variants was performed with vcftools (filter = 'PASS') using the pre-annotated 'CSQ' field. The pathogenic and likely-pathogenic missense variants were collected from the ClinVar, release July 2021.[32] Additionally, high confidence disease mutations were collected from the Human Gene Mutation Database (HGMD, version 2020 professional release),[33] with filters hgmd_confidence = 'HIGH' and hgmd_variant_Type = 'DM'. Variants from both ClinVar and HGMD databases were combined to generate the set of 'pathogenic' variants. Altogether, we obtained 87 028 'population' missense variants in 242 NDD genes and 9241 'pathogenic' missense variants in 207 NDD genes (Fig. 1) and used these variants to generate the Essential3D site annotation (Fig. 2 and Supplementary Fig. 1).

Selection of variants for the validation of Essential3D site annotation

The *de novo* variants were retrieved from the denovo-db database (http://denovo-db.gs.washington.edu; June 2019)[34] and were filtered for missense variants. All variants flagged as 'validated' and 'unknown' were collected. We obtained 848 variants labelled with one of the NDD-associated phenotypes within the 'Primary Phenotype' flag. The phenotypes that were included are: 'schizophrenia', 'developmental disorder', 'autism', 'intellectual disability', 'Tourette-syndrome', 'epilepsy', 'early-onset-Parkinson', 'cerebral palsy', 'sporadic infantile spasm syndrome' and 'amyotrophic lateral sclerosis'. Of 848 variants, 322 were not exclusive of those present in ClinVar and HGMD databases, and therefore, were used as an independent validation set (Fig. 3 and Supplementary Fig. 2). In addition, 23 650 benign population variants were collected from the DiscovEHR (June 2019) database.[35] Of these, 7854 were independent of the variants in gnomAD and were used as the control group in validation.

Selection of variants for testing the Essential3D site annotation

For testing, we collected data from exome sequencing studies on autism spectrum disorder (ASD; 35 584 individuals[7]) and developmental disorders (DD; 31 058 individuals[13]), and epilepsy (Epi25 collaborative; 9170 individuals[12]). From the epilepsy cohort, we selected the developmental and epileptic encephalopathy (DEE; $n = 1476$) subset since this early onset drug-resistant from of epilepsy is considered as NDD with the highest clinical genetic test diagnostic rate among epilepsies. We obtained 498, 2370 and 2865 variants associated with ASD, DD and DEE, respectively. Two hundred and twenty-eight ASD variants, 879 DD variants and 464 DEE variants were independent of the pathogenic variants used for generating Essential3D site annotation and were used as an independent set of NDD variants for testing Essential3D sites (Table 1). As a control group, we collected 53 989 variants from the UK Biobank (UKBB; 281 852 individuals[36]), of which 25 116 were exclusive of variants used for generating Essential3D site annotation and were used for testing. The gene-wise counts of variants are summarized in Supplementary Table 1. All variant counts reported above and in the Supplementary material correspond to unique amino acid substitutions.

Selection of the radius of 3D window

The 3D window around each amino acid residue (target residue) in the protein structure was defined by the residues located within a certain radius (r) from it (neighbouring residues), i.e. residues that meet the criteria:

$$d = \sqrt{(x_t - x_n)^2 + (y_t - y_n)^2 + (z_t - z_n)^2} < r \quad (1)$$

Here, d is the Euclidian distance between the 3D coordinates of the Cα-atom of the target residue (x_t, y_t, z_t) and neighbouring residues (x_n, y_n, z_n). We adopted this approach based on the hypothesis that a structurally and functionally important and variant-intolerant residue (i.e. Essential3D site) is likely to be proximal to other residues forming direct and higher-order contacts[37,38] and that are with known pathogenic variants.[19]

We tuned the value of r to maximize the enrichment of pathogenic variants compared to population variants around Essential3D sites in experimental structures (monomeric-PDB set; Fig. 1). For this analysis, we selected 50% of available pathogenic (ClinVar and HGMD) and population (gnomAD) variants by random bootstrapping and separated them out as a test set. With the remaining 50% variants, we generated Essential3D site annotation and evaluated the enrichment on the test set (two-sided Fisher's exact test). We repeated this analysis for r-values ranging from 6 to 18 Å and observed the highest enrichment of pathogenic variants around Essential3D sites for $r = 12$ Å [odds ratio (OR) = 12.6, $P = 4.3 \times 10^{-74}$; Supplementary Fig. 3]. Subsequently, $r = 12$ Å was used to define the 3D window for the generation of all structure-based scores across all structure sets (Fig. 2 and Supplementary Fig. 1).

Collection of features from UniProt

The UniProt database[29] was mined to collect the annotations of regions or sites of interest in proteins in terms of 26 features related to protein function (referred to as functional features; https://www.uniprot.org/help/sequence_annotation). These features were: active site, metal-binding site, binding site, site, zinc finger, DNA binding domain, nucleotide phosphate-binding region, calcium-binding region, region of interest, repeat, a coiled-coil region, motif, domain, topological domain, transmembrane domain, intramembrane domain, peptide, transit peptide, signal peptide, propeptide,

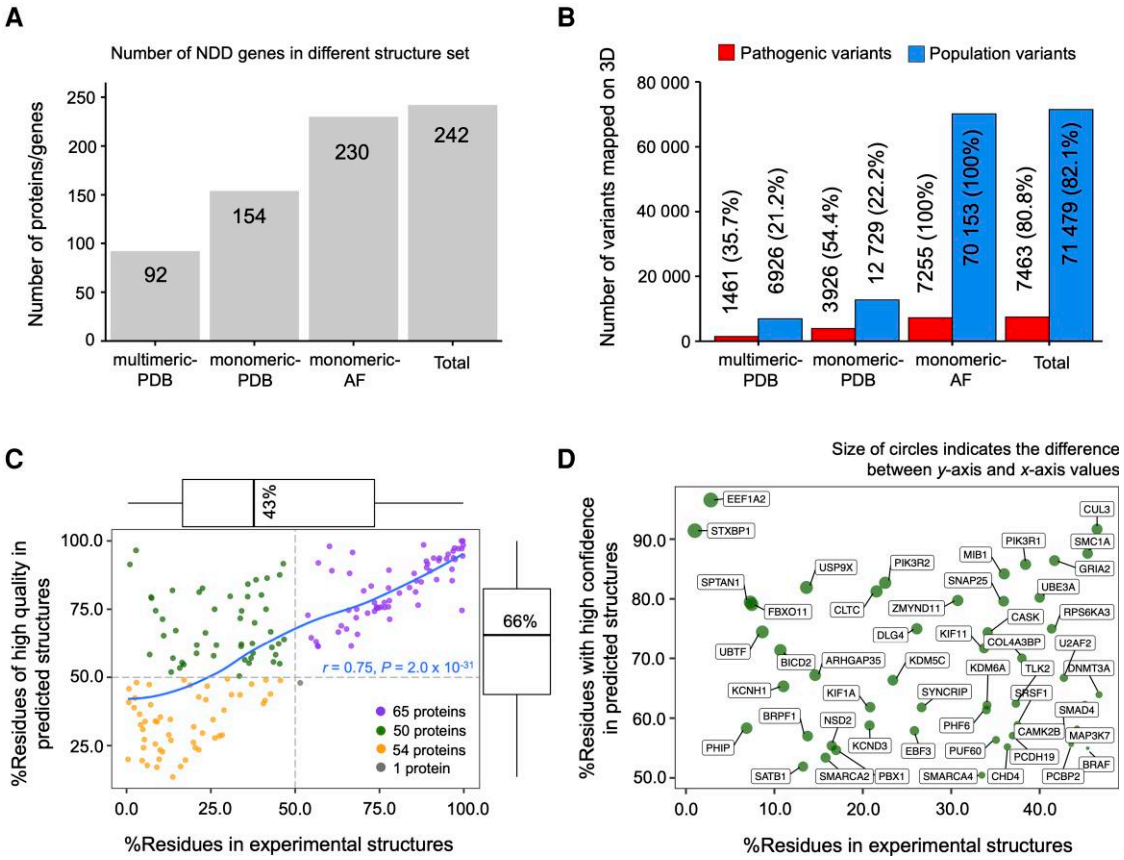

Figure 1 Overview of structure and variant level data in our dataset and related statistics. (**A**) Number of NDD genes with available protein structures in three different structure sets, and in total 3D structures. (**B**) Number of population (gnomAD) and pathogenic (ClinVar and HGMD) variants that were mappable onto different structure sets, and in total 3D structures. The number and percentage values correspond to variants mapped to 3D out of the total number of variants in genes included in the respective structure sets. (**C**) Relationship between per-protein sequence coverage (%residues) in experimental structures and fraction of high-quality residues in the corresponding AlphaFold-predicted structures. The locally fitted line (polynomial regression fitting) indicates that the greater the residue coverage in the experimental structure, the higher the fraction of high-quality residues in the predicted structure [r = 0.75 (95% CI: 0.67–0.81), $P = 2.0 \times 10^{-31}$, Pearson's product-moment test]. (**D**) Fifty NDD genes (second quadrant in **C**) for which the sequence coverage (%residues) in experimental structures (*x*-axis) was <50%, while AlphaFold-predicted structures had over 50% of the protein residues (*y*-axis) predicted with a high quality.

modified residues, lipidation, disulphide bond, cross-link, glycosylation and compositional bias.

Statistical analysis

Statistical enrichment of pathogenic variants versus population variants was computed using two-tailed Fisher's exact tests (Fig. 3, Table 1 and Supplementary Fig. 2). The same test was performed to find functional features associated with Essential3D sites (Fig. 5). We corrected for multiple testing by Bonferroni correction.

Implementation of the ES-NDD browser

The server is implemented as a JavaScript application running over the data generated by the pipeline. These include both residue-level annotations and predicted 3D structures. To visualize the annotations, the essential sites-NDD (ES-NDD) browser relies on the MolArt tool,[39] which enables visualization of sequence annotations in the context of available structural data. The annotations for the selected gene and structure are passed on to MolArt, displaying the annotations as colour overlays over the structure. In the case of experimental structures, the structure definition is fetched on the fly from the PDB. The data are fetched from the server for predicted structures and passed to MolArt together with the residue-level sequence-structure mapping.

Code availability

Our in-house scripts for 3D-score generation are available on our GitHub page: https://github.com/dlal-group/Essential3D_NDD.

Data availability

Precalculated annotations are available on our GitHub page (https://github.com/dlal-group/Essential3D_NDD).

Results

To identify and benchmark functionally essential sites in 3D structures of proteins, termed Essential3D sites, we searched the literature for genes known to have a significant exome-wide association with NDDs, specifically from three recently published epilepsy[9] and developmental disorder[11,13] studies, and retrieved 302 genes. Our study was focused on the analysis of missense variants, therfore we filtered this set of genes based on their constraints for missense variants (missense z-score[40] cut-off of 1.96) and obtained the final

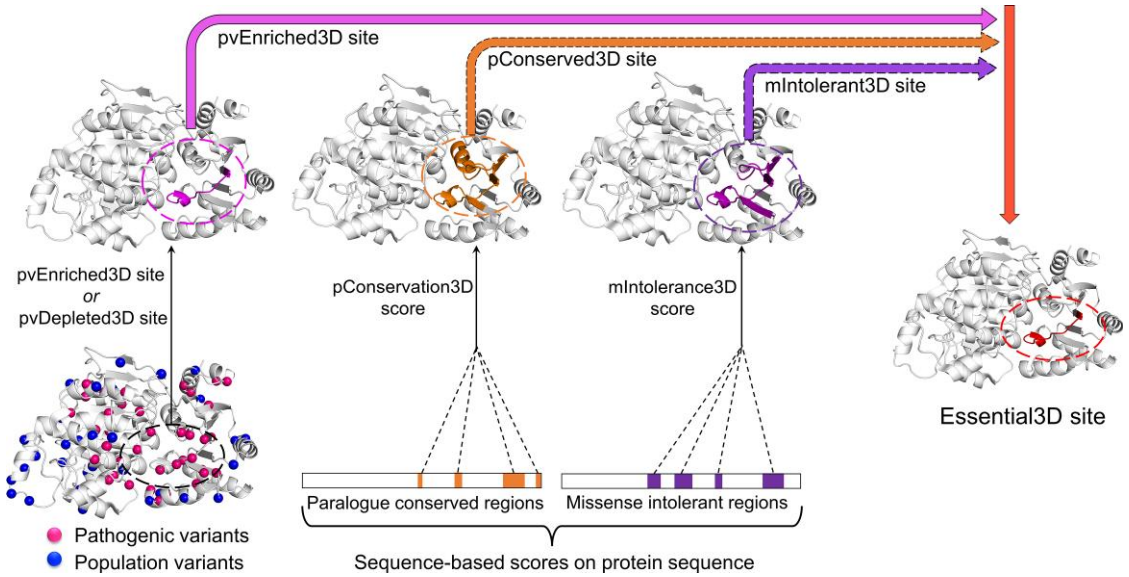

Figure 2 Schematic overview of the method to identify Essential3D sites. *Left*: Pathogenic (source: ClinVar and HGMD databases) and population (source: gnomAD database) missense variants are mapped onto protein structures. For each residue, a burden analysis (two-sided Fisher's exact test) is performed to detect 3D sites that are enriched for pathogenic variants and lack benign variants within a 12 Å 3D window (pvEnriched3D). *Middle*: Sequence-based scores quantifying conservation across human paralogues (coloured regions in the bar representing a protein sequence) and intolerance of missense mutations (coloured regions in the bar) are normalized for each residue within a 3D window (12Å radius), to compute pConservation3D and mIntolerance3D scores, respectively. Residues in 3D with pConservation3D > 0 and mIntolerance3D > 0 are designated as pConserved3D and mIntolerant3D sites. *Right*: 3D sites that are pvEnriched3D, pConserved3D and mIntolerant3D are identified as Essential3D sites (see the 'Materials and methods' section and Supplementary Fig. 1).

set of 242 NDD genes for this study (hereafter referred to as NDD genes or proteins; Supplementary Table 2). The majority of NDD-associated genes (73.2%, $n = 177$ genes) had a significant exome-wide association in only one of three epilepsy[9] and developmental disorder[11,13] studies, whereas 5% ($n = 14$ genes) and 21% ($n = 51$ genes) had a significant exome-wide association in two or all three studies, respectively (Supplementary Table 2).

Collection of protein structures and mapping of population and pathogenic variants

We collected structural data from two sources: (i) experimentally solved 3D structures from the Protein Data Bank (PDB)[28]; and (ii) highly accurate predicted 3D structures from the AlphaFold protein structure database.[27] For 185 of 242 NDD genes, at least one protein structure was available in the PDB (a total of 2715 structures). After filtering these structures based on coverage (see the 'Protein structure selection and filtering' section), we obtained 2461 experimentally solved structures representing 154 genes (64% of 242 NDD genes). From the AlphaFold database, we collected predicted structures of the full-length proteins representing 230 genes (95% of 242 NDD genes). In total, we generated three datasets from the available structures: two sets of monomeric structures (PDB and AlphaFold) and a set of full protein complexes—to first compare the quality of experimental and predicted and structures—and then separately analyse structures where an NDD protein is present as a single molecule and structures where an NDD protein is in complex with itself or other partner proteins, respectively. The two monomeric sets of structures were called: 'monomeric-PDB' set (154 proteins) and 'monomeric-AlphaFold' (230 proteins), and the set of protein complexes was named 'multimeric-PDB' (homomeric or heteromeric structures of 92 proteins) (Fig. 1A). Altogether, these three sets covered structural data for 242 NDD genes (Fig. 1A and Supplementary Table 1).

Next, to detect 3D sites that are enriched for pathogenic missense variants compared to those that are benign, we collected missense variants in these 242 NDD genes and mapped them onto their corresponding 3D protein structures. We obtained 87 028 'population' missense variants from the gnomAD database[30]; 82% of these variants were mappable onto 3D (71 479/87 028; Fig. 1B). At the same, we collected 9241 'pathogenic' missense variants in 207 (of 242) genes from ClinVar[32] (pathogenic and likely-pathogenic) and HGMD[33] databases, and could map about 81% of them onto 3D (7463/9241; Fig. 1B). The lack of pathogenic variants in 15% of known NDD genes can be explained because these NDD genes have only been identified recently[7,11] and are not yet routinely screened or reported clinically, therefore absent from clinical databases. Approximately 50% and 20% of pathogenic and population variants, respectively, could be mapped onto experimental structures (monomeric-PDB and multimeric-PDB sets; Fig. 1B), while 100% of all variants were mappable onto AlphaFold-predicted structures of full-length proteins (monomeric-AlphaFold set; Fig. 1B). The uniform mappability of both types of variants onto AlphaFold-predicted structures per protein (Supplementary Fig. 1A), enabled us to perform an unbiased identification of 3D sites that are enriched for pathogenic variants, accounting for the presence of population variants as the comparison group.

Next, we examined the AlphaFold-predicted protein structures with the analogous experimental structures to identify those NDD genes, for which the use of AlphaFold structures provided significant advantage in terms of high-quality predicted residues (pLDDT or predicted local-distance difference test >70; per-residue estimates of reliability generated by the AlphaFold neural network[26]). Overall, a positive correlation [$r = 0.75$ (95% CI: 0.67–0.81), $P = 2.0 \times 10^{-31}$, Pearson's product-moment test; Fig. 1C] was observed between per-protein sequence coverage (%residues) in monomeric-PDB and high-quality predicted residues in

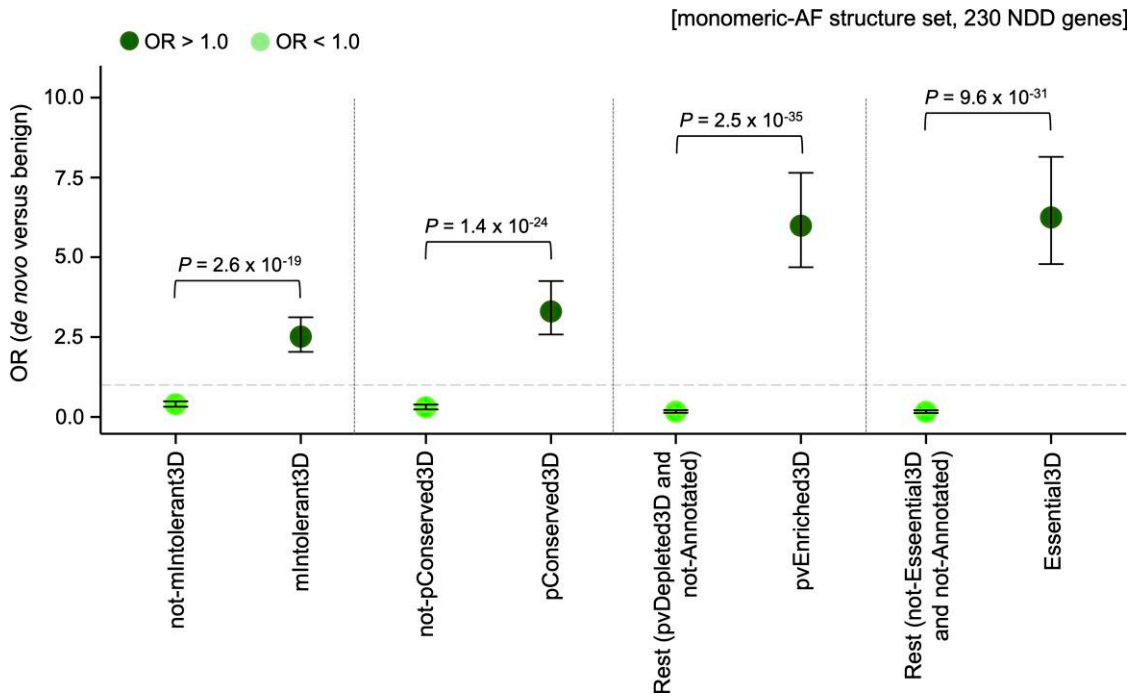

Figure 3 Residues annotated as Essential3D sites by a consensus approach are more enriched in *de novo* variants compared to the individual criterion, used to derive the consensus. OR values (*y*-axis) were calculated using two-sided Fisher's exact test with 432 NDD-associated *de novo* variants and 7957 benign variants. Filled circle colours indicate OR > 1.0 (enrichment) and OR < 1.0 (depletion). Error bars indicate 95% CIs and *P*-values represent the significance of test results. The horizontal dashed line represents OR = 1.0.

Table 1 Enrichment of Essential3D sites in independent sets of NDD variants

Sequencing studies	189 (of 242) NDD genes with identified Essential3D sites						
	M[a] (number of variants)	Essential3D site *m* of M variants (%M)	Rest of the protein sequence[b] *m* of M variants (%M)	Enrichment [OR (P)] of Essential3D sites in NDD variants versus control[c] (UKBB)			
ASD[7]	211 (93 genes)	52 (24.6%)	159 (73.4%)	8.8 (6.9×10^{-28})	–	–	7.7 (1.2×10^{-127})
DD[13]	750 (168 genes)	221 (29.5%)	529 (70.5%)	–	11.3 (1.6×10^{-123})	–	
Epilepsy (DEE) (Epi25 Collaborative[12])	392 (131 genes)	29 (7.4%)	363 (92.6%)	–	–	2.2 (3.2×10^{-4})	
UKBB[36]	22 943 (187 genes)	821 (3.6%)	22 122 (96.4%)	–	–	–	–

NDD missense variants were obtained from sequencing studies on ASD, known DD,[13] and DEE as part of the Epi25 collaborative.[12] Control variants were collected from the UKBB. The enrichment (two-sided Fisher's exact test) of Essential3D sites is reported separately for ASD, DD and DEE variants and for all NDD variants compared to those from UKBB as the control group.

[a]M indicates the number of variants identified in *n* genes in the corresponding study, out of 189 NDD genes with identified Essential3D sites. %M indicates the fraction of M variants that mutate Essential3D sites in the protein or the rest of the protein (not-Essential3D and not-Annotated sites).

[b]Not-Essential3D and not-Annotated sites.

[c]UKBB data are considered as the 'control' group for enrichment analysis against ASD, DD and DEE variants separately from three sequencing studies and total NDD variants from all three studies.

monomeric-AlphaFold structures. This observation supports the hypothesis that the availability of residue coordinates in experimental structures increases the possibility of a high-quality residue predictions in AlphaFold structures. However, for 50 proteins, over 50% of residues in predicted structures were of high quality (pLDDT > 70) despite their analogous experimental structures had less than 50% sequence coverage (Fig. 1C). Two notable examples include syntaxin-binding protein (*STXBP1*) and elongation factor protein (*EEF1A2*), for which high-quality predicted residue coordinates of almost the full-length proteins (over 90% sequence coverage) could be obtained from the AlphaFold database (Fig. 1D), while the available experimental structures for these proteins had less than 10% sequence coverage. To further compare the similarity between experimental and predicted structures, we superimposed the corresponding 3D structures from the monomeric-PDB and monomeric-AlphaFold sets using TM-align[41] and found that AlphaFold-predicted structures are highly similar to experimentally derived structures (median similarity score of 0.94; Supplementary Fig. 4B). This supports the reliability

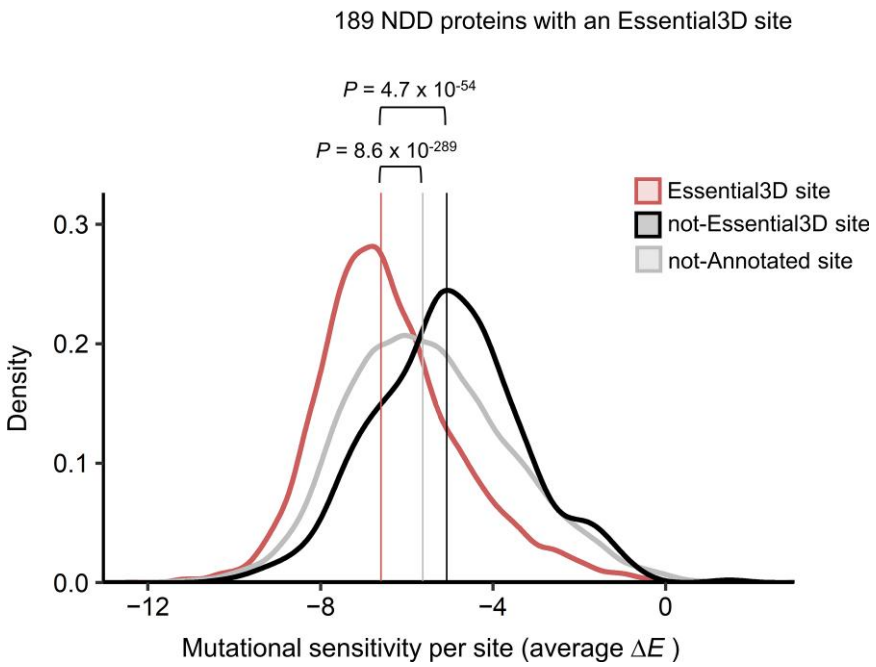

Figure 4 Essential3D sites highlight positions in proteins with high mutational sensitivity (average Δ*E*) as quantified by EVmutation.[46] Average Δ*E* is the mean change in energy for substituting the reference amino acid in the wild type protein with 19 alternate amino acids. The median mutational sensitivity for Essential3D, not-Essential3D and not-Annotated sites were −6.60, −5.07 and −5.63, respectively (Essential3D versus not-Annotated sites, $P < 8.6 \times 10^{-289}$; Essential3D versus not-Essential3D sites, $P < 4.7 \times 10^{-54}$; Mann-Whitney U-test).

of AlphaFold-predicted structures and warrants their use in our study to increase the power of statistical analyses and identify essential sites across many NDD genes. All experimental and predicted structures of the 242 NDD proteins annotated with variants and generated scores are made publicly available (https://es-ndd.broadinstitute.org).

Identification of Essential3D sites in 189 neurodevelopmental disorder-associated genes

Distant amino acid residues in the linear protein sequence are often in close proximity in 3D to form conserved structural folds,[42,43] which are essential for protein's stability and function (e.g. interaction, recognition, signalling). Based on this fundamental principle of protein folding, we developed a structure-guided, consensus approach to identify 'Essential3D' sites in protein structures using three lines of evidence: (i) residues that are enriched for pathogenic variants compared to population variants; (ii) residues that are evolutionarily conserved across human paralogues[44]; and (iii) that are intolerant of missense variants[45] (Fig. 2; see the 'Materials and methods' section for details on methodology and Supplementary Fig. 1). Each line of evidence and the consensus annotation of Essential3D sites were derived separately for 3D structures in three structure sets (Fig. 1A).

To detect pathogenic variant enriched 3D sites, we performed burden analysis (two-tailed Fisher's exact test) of pathogenic versus population missense variations within a 3D window of 12 Å radius around each residue in the 3D structure (see the 'Selection of the radius of 3D window' section and Supplementary Fig. 3). For the monomeric structures (monomeric-PDB and monomeric-AlphaFold), the spatial window included only residues of the selected NDD protein chain itself, whereas, for the multimeric structures, the window may also include residues from any interacting protein chain in the complex within the defined 12 Å radius. Residues with significant enrichment of pathogenic variations within their 3D window (OR > 1.0 and $P < 0.05$) were categorized as pathogenic variant enriched 3D sites ('pvEnriched3D'; Fig. 2). Additionally, residues with significant depletion of pathogenic variations within their 3D window (OR < 1.0 and $P < 0.05$) were categorized as pathogenic variant depleted 3D sites ('pvDepleted3D').

To generate two additional lines of evidence for capturing essential sites, we first annotated the amino acid residues of proteins with two sequence-based scores: a score quantifying the conservation of residues across human paralogue proteins[46] and a missense variant intolerance score (missense tolerance ratio or MTR). Then, we normalized these scores using the z-score function for each residue in protein structures within the 3D window of a predefined radius, to compute the 3D representation of these two sequence-based scores: per-residue 'pConservation3D' (paralogue conservation 3D score) and 'mIntolerance3D' (missense intolerance 3D score) (Fig. 2). Residues with pConservation3D > 0 and mIntolerance3D > 0 were identified as pConserved3D and mIntolerant3D sites, respectively (see the 'Materials and methods' section and Supplementary Fig. 1).

Finally, we derived the consensus annotation of residues—termed Essential3D sites—that are pvEnriched3D, pConserved3D and mIntolerant3D sites (Fig. 2). Similarly, residues that are pvDepleted3D, not-pConserved3D and not-mIntolerant3D sites were annotated as not-Essential3D (see the 'Materials and methods' section and Supplementary Fig. 1). Residues that did not meet the criteria to be Essential3D or not-Essential3D sites were kept 'not-Annotated'. Altogether, 14 377 Essential3D sites were identified in predicted or experimental structures of 189 (of 242) NDD proteins. Structure set-wise and gene-wise counts of identified pEnriched3D, pConserved3D, mIntolerant3D, and Essential3D sites are reported in Supplementary Fig. 5 and Supplementary Tables 3–6, respectively.

Essential3D sites improve variant prioritization over single criteria

After identifying Essential3D sites by taking a consensus from three criteria (Fig. 2; pEnriched3D, mIntolerant3D and pConserved3D sites), we wanted to validate the utility of employing a consensus approach over a single criterion. To check this, we selected an independent set of NDD-associated *de novo* missense variants from the denovo-db database[34] ($n = 432$, after excluding all variants that overlapped with pathogenic variants in ClinVar and HGMD; see the 'Materials and methods' section). Additionally, we generated an independent set of benign population variants from the DiscovEHR database[37] ($n = 7957$, after excluding all variants that are present in gnomAD). Then we measured the burden of Essentian3D sites in *de novo* variants compared to benign variants (two-sided Fisher's exact test). We observed a 6.2-fold burden of Essential3D sites in *de novo* variants [95% CI (4.8–8.1), $P = 9.6 \times 10^{-31}$] for predicted structures (monomeric-AlphaFold set; Fig. 3). By repeating the analysis for pEnriched3D, mIntolerant3D and pConserved3D sites, we found that *de novo* variants had the highest burden of Essential3D sites, with a gradual decrease in effect size for pvEnriched3D sites (OR = 5.9), pConserved3D sites (OR = 3.3), and mIntolerant3D sites (OR = 2.5; Fig. 3). This observation supports the strategy of using multiple lines of evidence in identifying the functionally essential 3D sites in proteins. Similar investigations were performed for experimental structures, which showed comparable results (Supplementary Fig. 2).

Essential3D sites are enriched for rare missense variants in large exome sequencing studies

After demonstrating that Essential3D sites are enriched in *de novo* missense variants, we attempted to further validate whether these sites in NDD genes present vulnerable protein positions for rare missense variants associated with NDDs. For this, we used data from three rare variant sequencing studies on NDDs, including the largest ASD[7] (35 584 cases), DD[13] (31 058 cases) and Epi25 collaborative[12] (9170 cases) studies. Since several of these studies used variants from gnomAD as the control group, which was also part of our initial Essential3D site annotation, we selected population variants from the UK-Biobank (UKBB[36]) as the control dataset. Essential3D sites were 8.8-fold ($P = 6.9 \times 10^{-28}$) enriched for ASD patient variants, 11.3-fold ($P = 2.6 \times 10^{-123}$) enriched for DD patient variants, and 2.2-fold ($P = 3.2 \times 10^{-4}$) enriched for variants associated with DEE, compared to population variants from UKBB. The relatively lower enrichment of Essential3D sites in DEE variants in the Epi25 study could be explained because of a higher ration of likely benign variants. In contrast to the other studies, the Epi25 study only includes cases and no data from parents. Since the information about inheritance is missing, we were not able to filter for *de novo* variants—a powerful filter to enrich a variant set for pathogenic variants. A combined analysis of all NDD missense variants from three exome studies of ASD, DD and DEE versus missense variants from UKBB as control, showed about an 8-fold enrichment ($P = 1.2 \times 10^{-127}$, two-sided Fisher's exact test; Table 1) of Essenetial3D sites in NDD variants.

Essential sites capture mutation-intolerant locations for yet unseen variants

Although Essential3D sites were found enriched for NDD missense variants (Fig. 3), about 76% of identified sites in 189 NDD proteins were not annotated in any patient variant databases (e.g. ClinVar, HGMD, denovo-db) and was also not found in a rare variant sequencing studies discussed in Table 1. Additionally, 94% of these Essential3D sites also are not present in gnomAD population database. To assess the mutational effects of all Essential3D sites, we collected the ΔE mutational effect, i.e. the mean change in statistical energy for substituting the reference amino acid in the wild-type protein with all other amino acids, computed by EVmutation[47] for all residues in the 189 NDD proteins with Essential3D sites. The more negative the energy change, the more deleterious the effect of mutating that residue is, accounting for co-evolution and epistasis.[47] Expectedly, Essential3D sites in proteins showed significantly different mutational sensitivity compared to not-Essential3D sites and not-Annotated sites (Mann-Whitney U-test; Fig. 4), with the median ΔE for Essential3D sites being relatively more negative. Moreover, the median ΔE for Essential3D sites with and without currently a known pathogenic mutation (ClinVar and HGMD databases) on them were comparable, −6.9 and −6.4, respectively. This indicates that Essential3D sites represent potential positions for novel deleterious mutations and will be valuable in prioritizing missense variants and important residues in proteins.

Essential3D sites prioritize features important for the function of the protein

By definition, Essential3D sites are enriched for pathogenic variants within a local 3D window in the protein structure, conserved across paralogs, and depleted for population missense variants (Fig. 2), which makes it likely that these sites represent residues important for specific protein function. For testing this, we used the curated annotation of 26 position-specific features related to protein function from the UniProt database[29] (e.g. active sites, ligand binding sites, functional domain, referred to as 'functional features'; see the 'Materials and methods 'section for all 26 features). Overall, 12 083 of 14 377 (84%) Essential3D sites were overlapping a functional feature. Additionally, we identified 12 features that were significantly enriched for Essential3D sites (two-sided Fisher's exact test; Fig. 5 and Supplementary Table 7) in the 189 NDD genes. The remaining features were either depleted for Essential3D sites or showed no significant association.

Intramembrane regions of proteins located in a membrane without crossing it showed 15-fold enrichment of Essential3D sites among all investigated features ($P = 1.3 \times 10^{-227}$; Fig. 5), where the signal of enrichment was primarily contributed by the genes in the potassium voltage-gated channel family (*KCNA2*, *KCNB1*, *KCND3*, *KCNH1*, *KCNK3*, *KCNQ2*, *KCNQ3*; Supplementary Table 7). This enrichment was greater than the enrichment in the overall transmembrane region (6.3-fold, $P < 1.0 \times 10^{-300}$; Fig. 5), which has been previously reported as an important region in NDD associated ion channels.[14,15,48] Nevertheless, in 27 NDD proteins from voltage-gated calcium, sodium, and potassium channel families, GABA receptors, glucose transporters, and NMDA receptors (Supplementary Table 7), Essential3D sites were found enriched in the transmembrane region.

Interestingly, different types of binding sites and regions are found to be characteristic features of Essential3D sites (Fig. 5), such as nucleotide phosphate-binding regions (7-fold enriched, $P = 1.7 \times 10^{-113}$), active sites (7.4-fold enriched, $P = 1.7 \times 10^{-5}$), binding sites for physiological ligand, co-enzymes (6.4-fold enriched, $P = 1.9 \times 10^{-13}$), and metal-binding sites (5-fold enriched, $P = 1.2 \times 10^{-15}$). This result was driven by Essential3D sites being ubiquitously overlapping with these binding sites in NDD proteins that are kinases, phosphatases, helicases, etc (Supplementary Table 7). In addition, DNA binding domains and short conserved 'motifs' of

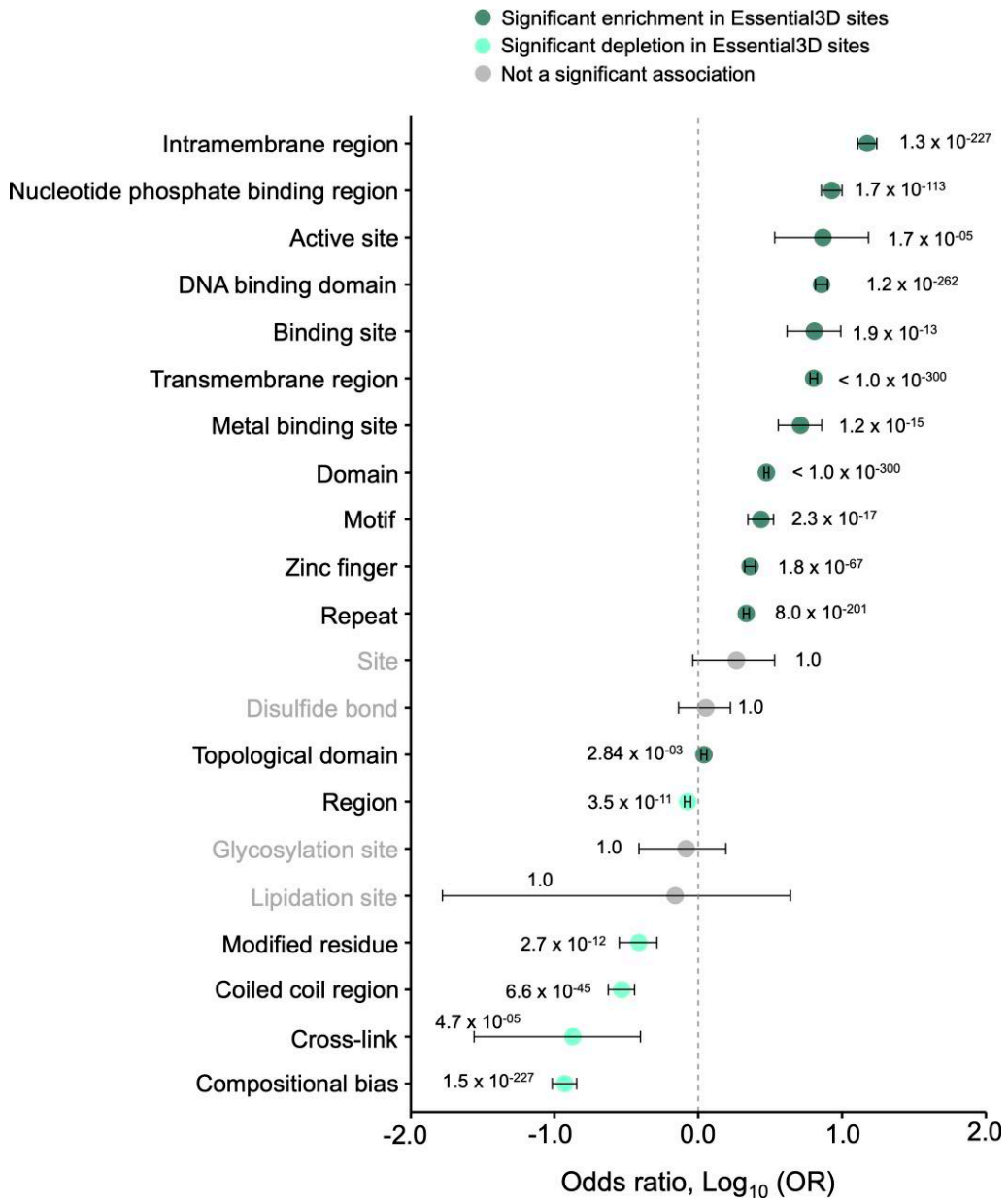

Figure 5 Association of Essenetial3D sites with function features from the UniProt database. The plot shows the results of two-sided Fisher's exact tests of association between Essential3D sites and the remaining protein residues with the features for the 189 NDD proteins with an Essential3D site. The plot includes results for the 21 (of 26) functional features that overlapped with at least one Essential3D site (see the 'Materials and methods' section for all 26 features). Log_{10} (OR) > 0 and Log_{10} (OR) < 0 along with $P < 0.05$ (after Bonferroni correction; see the 'Materials and methods' section for details) indicates that the corresponding feature (*y*-axis) has significant enrichment and depletion, respectively, in Essential3D sites. Error bars indicate 95% CIs. The vertical dashed line at Log_{10} (OR) = 0 indicates no association. For non-significant associations ($P \geq 0.05$), the circle and feature name are grey.

biological significance (e.g. DEXX, LXXLL motifs) are found to be preferential locations of Essential3D sites (Fig. 5).

Essential3D site annotation use case examples

Identification of protein complex inter-molecule interactions as disease mechanism

In large protein complexes, variants at protein–protein, protein–ligand interaction surfaces can alter protein function (recognition, signalling, etc.), and can cause diseases. To illustrate the utility of Essential3D sites in identifying such mechanisms of missense variants, we now describe two case studies of experimentally solved protein complexes associated with NDDs: the guanine nucleotide-binding proteins (G protein) complex[23,49,50] and the PP2A complex.[51]

G proteins have been associated with multiple NDDs, including DEE and involuntary movement disorders.[23,49,50] As a disease mechanism, perturbation of G protein signalling has been proposed.[23] For our investigation, we selected a hetero-tetrameric protein structure (PDB ID: 6G79, Fig. 6A), where the α-subunit of the G protein (Gα, *GNAO1*) is in a complex with G(I)/G(S)/G(T) subunit β_1 ($G\beta_1$, *GNB1*) and G(I)/G(S)/G(O) subunit γ_2 ($G\gamma_2$, *GNG2*), and the

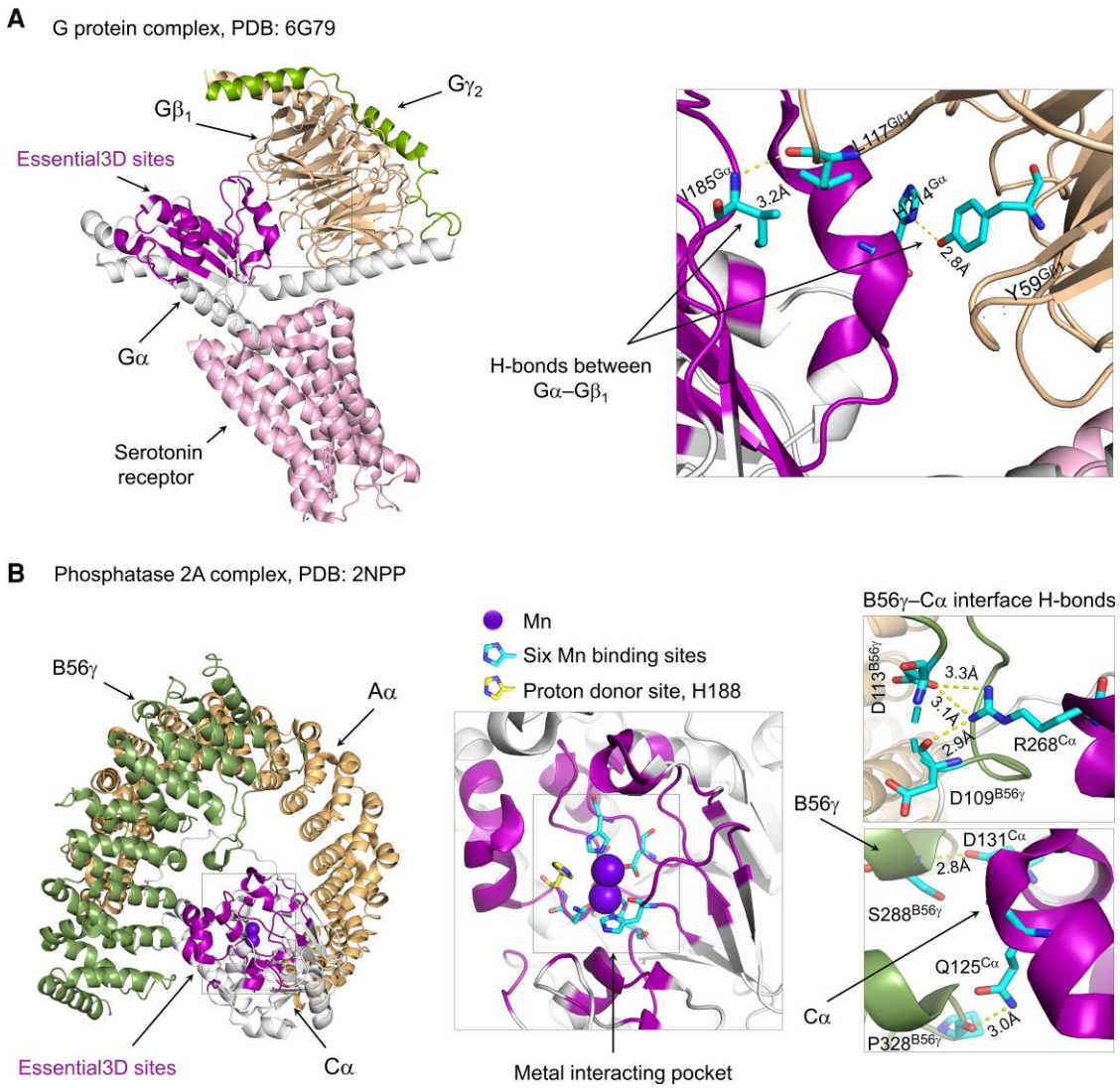

Figure 6 Essential3D site annotates important protein-protein interactions in protein complexes. (**A**, *left*) In the G-protein complex structure comprising multiple NDD genes (*GNAO1*: Gα subunit, *GNB1*: $G\beta_1$ subunit), Essential3D sites highlight two potential mechanisms: Gα-Gβ1 and Gα-GTP interactions. *Right*: In the Gα-Gβ1 interprotein interface, two Gα-residues: $I185^{G\alpha}$ and $H214^{G\alpha}$, forming hydrogen bonds with two $G\beta_1$-residues: $L117^{G\beta1}$ and $Y59^{G\beta1}$, were identified as Essential3D sites. Additionally, three known GTP-binding regions recorded by UniProt were identified as Essential3D sites. (**B**, *left*) Essential3D sites in the phosphatase 2A protein complex of catalytic Cα (*PPP2CA*), scaffolding Aα (*PPP2R1A*), and regulatory B56γ (*PPPR5C*) highlight two mechanisms: Cα-catalytic pocket and Cα-B56γ interactions. (*middle*) Six Mn binding residues (D57, H59, D85, N117, H167, H241) and one proton donor residue (H118), forming a metal interacting catalytic pocket around the manganese were captured as Essential3D sites. *Right*: In the Cα-B56γ interaction surface, two hydrogen bond networks were captured by Essential3D sites: (*top*) $R268^{C\alpha}$ (an Essential-3D site) with $D109^{B56\gamma}$ and $D113^{B56\gamma}$; and (*bottom*) $Q125^{C\alpha}$ and $D131^{C\alpha}$ (Essential3D sites) with $P328^{B56\gamma}$ and $S288^{B56\gamma}$, respectively.

G-protein coupled receptor (*HTRB1*) for serotonin. Eighty-five residues in the Gα structure were identified as Essential3D sites (Fig. 6A, left), including all three GTP-binding regions (G40–S47, D201–Q205, N270–D273) and one metal-binding site (S47). We further investigated the Essential3D sites on Gα-Gβ1 interaction surface and found two sites in Gα: I185, H214, forming hydrogen bonds (H-bond) with two residues in Gβ1: Y59, L117 (Fig. 6A, right), which are potentially important for Gα-Gβ1 recognition as well as the overall stability of the protein complex.

As a second example, we checked the type 2A protein phosphatase (PP2A) heterotrimeric complex (PDB id: 2NPP; Fig. 6B), comprising three subunits: a catalytic Cα subunit (*PPP2CA*), a regulatory B56γ subunit (*PPP2R5C*) and a scaffolding Aα subunit (*PPP2R1A*). PP2As are highly expressed in the brain and regulate neuronal signalling by catalysing phospho-Ser/Thr dephosphorylations in diverse substrates. Both haploinsufficiency and other mechanisms (i.e. dominant-negative) of *PPP2CA* have been suggested to cause intellectual disability and developmental delay.[51] We identified 110 Essential3D sites in two PP2A subunits, Aα (26 residues) and Cα (84 residues). In the catalytic Cα, all six Mn binding sites (D57, H59, D85, N117, H167, H241; in cyan) and the proton donor site H118 (in yellow) have been detected as Essential3D sites, forming a metal-interacting catalytic pocket around the manganese (Fig. 6B, middle). Additionally, the Essential3D site annotation captured residues on the Aα–Cα interface, forming inter-protein H-bonds between the regulatory and catalytic subunit of the complex (Fig. 6B, right), which are essential for the overall stability and functionality of macromolecular protein complexes.[52] These observations suggest that Essential3D sites can effectively annotate functionally important residues in the protein-protein interface of protein complexes.

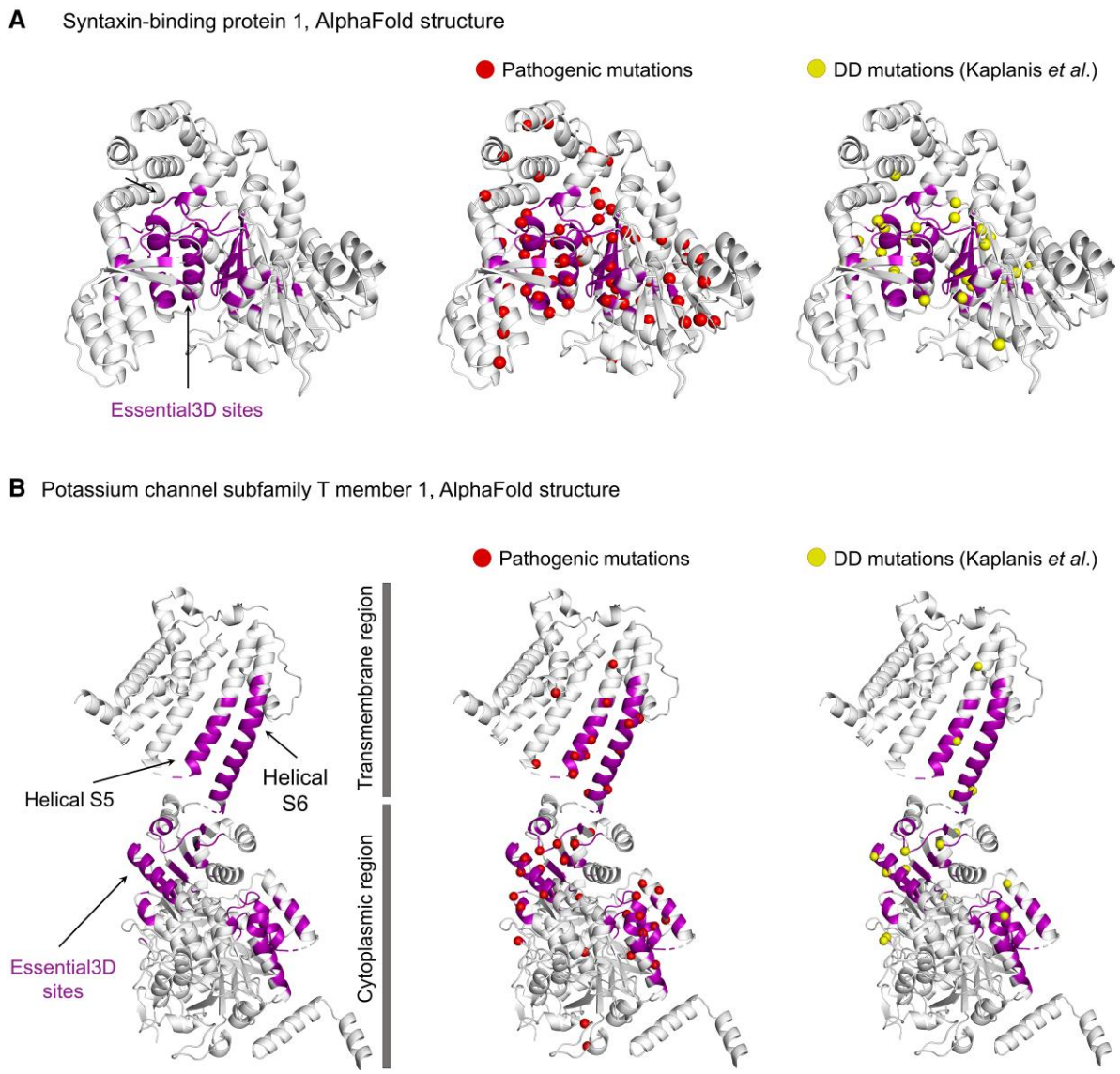

Figure 7 Use of AlphaFold-predicted structures enabled identification of Essential3D sites in NDD proteins with little to no experimental structure available and these Essential3D sites capture *de novo* DD-associated variants. (**A**, *left*) Predicted structure of syntaxin-1 protein (*STXBP1*) with annotated Essnetial3D sites. Pathogenic *STXBP1* variants from ClinVar and HGMD databases (*middle*) and *de novo* DD variants (*right*) are mapped onto the structure of Syntaxin-1. Sixty-five per cent of all DD variants affected Essential3D sites. (**B**) Predicted structure of potassium channel protein, subfamily T member 1 (*KCNT1*) with annotated Essnetial3D sites. Pathogenic *KCNT1* variants from ClinVar and HGMD databases (*middle*) and *de novo* DD variants[13] (*right*) are mapped onto the structure. Sixty-five per cent of all DD variants affected Essential3D sites. Of 18 *KCNT1* variants associated with DD, 10 affected Essential3D sites.

In addition to these case studies, we observed that overall Essential3D sites capture a higher fraction of *de novo* missense variants in protein complexes (multimeric-PDB) than in monomeric structures (monomeric-PDB set) of the NDD proteins (15% versus 7%; Supplementary Fig. 6B). Annotations for all protein complexes representing 92 NDD proteins are available through the resource developed as part of the study (https://es-ndd.broadinstitute.org).

Prioritization of essential functional unites in genes with previously missing experimental structure

Here, we present two case studies of predicted structures of NDD proteins: STXBP1 and KCNT1, for which the use of AlphaFold-predicted structures provided a significant advantage in terms of sequence and variant coverage in 3D.

STXBP1 is a key component of the SNARE complex, a mediator of vesicle fusion and exocytosis. Multiple mutations in STXBP1 protein reportedly are associated with early infantile epileptic encephalopathy.[53] The experimentally resolved structure of STXBP1 covered only a small, phosphorylated six amino acid residue long region of the protein, whereas the AlphaFold-predicted structure covered almost the full STXBP1 protein (91% of 594 residues are of high-quality, pLDDT > 70). Essential3D sites in STXBP1 were located in the compact core of the protein structure (Fig. 7A, left). While the pathogenic mutations (ClinVar and HGMD databases) were observed both in the core and at the surface of the structure (Fig. 7A, middle), the majority of the DD-associated mutations (17 of 26) found in an independent cohort study,[13] were located on Essential3D sites (Fig. 7A, right). Additionally, the mapping of autism ($n = 25$; Satterstrom *et al.*[7]) and DEE variants ($n = 9$; Epi25 collaborative[12]) onto the structure showed

specific 3D patterns of mutations associated with these two different NDDs (Supplementary Fig. 7). Specifically, 52% of autism versus 22% of DEE mutations affected Essential3D sites, indicating that Essential3D site annotation may preferably capture variants associated with ASD and other DDs over DEE.

Finally, we studied the predicted structure of the potassium channel subfamily T member 1 (KCNT1; Fig. 7B, left, only the residues predicted with a high-quality are shown). Variants affecting KCNT1 have been found to cause DEE and reported in multiple studies.[54,55] KCNT1 is composed of six helical transmembrane segments (S1–S6), two N- and C-terminal cytoplasmic regions, and extracellular and intercellular regions. Three visible clusters of Essential3D sites were found in the structure: one in the transmembrane region (spanning S5 and S6) and two in the C-terminal cytoplasmic regions. Noticeably, these three regions harboured most pathogenic mutations available in ClinVar and HGMD databases as well as the *de novo* DD-associated mutations[13] (Fig. 7B, middle and right), indicating that Essential3D site annotation could be used as a tool for prioritizing these DD variants for functional assays. Note that, no experimental structure is available for the KCNT1 protein, but the AlphaFold-predicted structure covered 850 of 1230 residues with high confidence and enabled us to identify Essential3D sites.

Discussion

We present a resource of protein structures corresponding to NDD genes, where residues in 3D are annotated with multiple genomic indications (mutational hotspots, conservation, etc.) and functional features (e.g. DNA binding site, catalytic pocket), that could serve as a resource for the translational neuroscience community. In particular, the interpretation of missense variants within these NDD genes is challenging. We focused on the study of missense variants. A recent study analysing *de novo* mutations in over 31 000 individuals identified 28 novel genes associated with developmental disorders.[13] Reportedly, 54% of these genes had only missense *de novo* mutations with an alteration of protein function as the most plausible disease mechanism.[13]

New high-quality structure predictions enable large scale identification of functionally essential protein sites

A significant advancement in the field of structural biology occurred in 2021 when DeepMind's artificial intelligence-based method AlphaFold predicted protein structure models of a quality approaching that of experimental determination.[26,27] But to fully utilize the potential of this resource in medical research, i.e. for clinical variant interpretation, prioritization of variants for functional assay and uncovering how perturbation of the structure affects the protein's mechanism of function, these structures need to be further annotated with clinical and genomic data.[56] In this study, we generated such a rich annotation resource of functionally important residues in 3D (Essential3D sites) within 242 NDD genes; these Essential3D sites are enriched for pathogenic variants in 3D, conserved across human paralogues, and constrained for missense variants in the general population (Fig. 2 and Supplementary Fig. 1). Use of AlphaFold-predicted structures noticeably increased the sample size and residue coverage in our study: we could analyse 76 additional NDD genes (Fig. 1A) for which no experimental structure was available and identified 11 136 additional Essential3D sites compared to those identified only in experimental structures (Supplementary Fig. 5). Additionally, in a supplemental analysis, we observed that Essential3D sites in predicted structures capture 3.5-times more *de novo* NDD mutations compared to experimental structures of the same number of proteins ($n = 142$; Supplementary Fig. 6A), which supports the benefit of using AlphaFold-predicted structures with a higher sequence coverage. Moreover, >90% of all Essential3D sites identified in AlphaFold-predicted structures are of high quality (pLDDT > 70; Supplementary Fig. 8). This could be driven by the fact that AlphaFold is highly accurate (pLDDT > 70) for protein regions with defined secondary structure, and by our 3D window-based annotation method, we expect to capture Essential3D sites in the compact structured regions of the protein.

Structure-guide methods are complementary to sequence-based methods

As part of our method to identify 3D essential sites, we transformed two sequence-based scores, i.e. gene-family or paralogue conservation[57] and missense intolerance scores,[45] into 3D using a local 3D neighbourhood-based normalization method, and computed pConservation3D (paralogue conservation in 3D) and mIntolerance3D (missense intolerance in 3D) scores (Supplementary Fig. 2). Upon assessing the correlation between sequence-based and 3D scores as a function of number of residues (n) within the local 3D window, we observed that as n increases, the correlation decreases (Pearson's $r = 0.89$ for $n = 4$ to $r = 0.51$ for $n = 52$, MTR score versus mIntolerance3D; $r = 0.85$ for $n = 4$ to $r = 0.57$ for $n = 52$, sequence-based paralogue conservation score versus pConservation3D; Supplementary Fig. 9A). This suggests that sequence-based and structure-informed scores differ for residues with many spatially proximal residues around them (quantified by n), meaning within the compact protein core and at the crowded protein–protein interaction surfaces. An additional comparison of correlations between sequence-based and 3D scores for residues with different prediction quality provided by the AlphaFold method[26] showed a drop in Person's r from 0.82 to 0.65 as the prediction quality (pLDDT) decreased from >90 to <50 (Supplementary Fig. 9B). It has been shown that residues with pLDDT < 50 often are part of unstructured protein regions whereas high-quality predictions predominantly represent structured domains.[27,58] Combined with these existing study outcomes,[27,58] our results indicate that the local 3D window based, structure-guided pConserved3D and mIntolerant3D scores are complementary to sequence-based methods in detecting essential 3D sites that represent residues located far apart in sequence but are spatially proximal in the 3D structure. In fact, we found that about 44% of all Essential3D sites have neighbouring residues within 12-Å radius window that are on average ≥50 residues (median = 102 residues, maximum = 895 residues) apart in the sequence (Supplementary Fig. 10).

A computational approach to identifying mutational hotspots and critical functional domains

According to American College for Clinical Genetics and Genomics (ACMG) guidelines, a missense variant located in a mutational hotspot or critical functional domain is a moderate evidence criterion for pathogenicity. Previous bioinformatic resources and tools have attempted to identify such mutational hotspots through clustering of pathogenic and likely pathogenic missense variants reported in clinical variant databases[59-63] or by scoring the absence of population variants along the protein sequence (MTR[45]). Other approaches that identify critical functional domains have attempted to bridge genomics with structural biology to develop resources and tools that map and visualize missense variants on protein structures. Such tools include mutation3D,[59] COSMIC-3D,[64] PhyreRisk[65] and VarMap.[66] Spatial neighbourhood-based methods using 3D protein structural

models have also been developed to generate pathogenicity predictions (PIVOTAL[67]) and combine genomic data with structures and network models (Bio-node 3D[68]). Our essential site annotation combines all these approaches while using the latest version of AlphaFold predicted structures and the largest NDD patient variant datasets. Combining all available types of variant functional features and mapping these on consensus scores on structure has the benefit to provide molecular insights that help predict clinical phenotypes. Most bioinformatic tools focus on scores, which allow for benign versus pathogenic variant discrimination. However, for many disease-associated genes, pathogenic missense variants can cause mild to severe disorders or even lead to multiple clinically distinct disorders due to differences in the protein's altered molecular function. However, in medical genetics practice, better interpretable information than variant pathogenicity predictions alone are needed to adjust a patient's management.[17] For example, the gene *GJA1* is associated with highly pleiotropic inheritable diseases affecting a variety of organ systems.[69] The severity of a phenotype can also differ across missense variants in the same gene. Pathogenic variants in the *SCN1A* gene can lead to Dravet syndrome, a catastrophically severe epilepsy syndrome; or GEFS+, a milder benign form.[70]

Development of a web portal to facilitate variant interpretation

Effective synthesis of available information into meaningful conclusions for a novel variant represents a significant challenge. When and where to make effective use of the overwhelming number of independent bioinformatic resources, scores, and currently available tools requires knowledge and experience in multiple specialized domains beyond the scope of a single user's expertise.[17,71-74] Most people interested in missense variant interpretation are clinicians, biologists, genetic counsellors and structural chemists, who may not have extensive experience using bioinformatic methods. To address this issue, tools with interactive mechanisms that apply to human-centred design methods and principles[75,76] can grant end-users more agency in guiding the interpretation and can be used for critical decision-making purposes beyond a score.[75,77] Therefore, we developed ES-NDD (https://es-ndd.broadinstitute.org), a user-friendly web application that includes rich sources of annotations—including our essential 3D sites—that can be interactively explored in 3D visualizations of protein structure models. We believe, this way, the user will find more trust in the results, the agency to hypothesis-test and apply their domain knowledge, while simultaneously leveraging the benefits of automation. This will also help determine whether a mutation of a site is likely to alter a specific function of the protein, which often could be hypothesized from position-specific features of Essential3D sites. A use case example: we identified that residues lining the central pore in the GLUT1 transporter structure including those that are interacting with the monosaccharide are Essential3D sites (PDB ID: 4PYP; Supplementary Fig. 11). Patient variants in GLUT1 have been associated with the GLUT1 deficiency characterized by severe early-onset epilepsy, developmental delay and movement abnormalities.[78,79] It could be hypothesized that a mutation on these sites with the 'ligand-binding' feature, may cause an alteration of protein function by changing the shape of the pore or perturbing the protein-ligand interaction.[80] Using similar feature annotations from the UniProt database[29] (Fig. 5), we identified 12 features that are statistically associated with Essential3D sites in the 189 NDD genes. We postulate that the presence of a novel mutation on an Essenetial3D site with a certain feature will aid in hinting to the mechanism of the function that would be perturbed upon mutation of these sites.

Limitations

There are important challenges in assessing missense variants that are not covered in this report. First is the heterogeneity of disease presentation. Different variants of the same gene can lead to different disease severity or even different diseases,[14,15] aspects that will be masked by using simple discrete pathology categories. Currently, Essential3D sites are calculated based on the aggregation of all available pathogenic variants for each gene, and we expect that our method can be applied to predict disease-specific Essential3D sites that may be useful to discriminate different disease presentations. Second, we studied variants from the canonical transcripts only; hence, Essential3D sites annotations were identified for the canonical protein isoforms only. However, every gene has an average of 3.5 transcripts,[81] and it is not always known which transcript is expressed in which cell type or in the brain at all. Third, we validated Essential3D sites by evaluating their burden in patients compared to variants from UKBB across all populations. In the future, it will be useful to perform this evaluation in a population-specific manner to further understand the clinical utility of Essential3D sites in different populations. Fourth, for many protein structures, different assemblies of protein subunits exist. As we could show by studying Essential3D sites in protein complexes, such assemblies affect the calculation of 3D scores and can provide insights into important protein-protein interaction sites. We choose one protein complex per gene, based on the structure coverage of the protein, for our analysis. Still, we precalculated Essential3D sites for all protein complexes for each of the 242 NDD-associated genes, which can be explored in our ES-NDD browser.

Conclusion

To summarize, we developed a new method to annotate functionally essential 3D sites in proteins accounting for structural and genomic (i.e. conservation, constraints) context. We applied this method to experimental and predicted structures of 242 NDD-associated proteins, demonstrated that de-novo missense variants in the largest NDD sequencing studies are highly enriched at Essential3D sites and created a resource of 3D structures annotated with population and pathogenic variants and the essential 3D sites (ES-NDD; https://es-ndd.broadinstitute.org). We believe that our resource will enable a wide range of communities involved in neuroscience research, including clinicians, biologists and genetic counsellors without the bioinformatics training, to perform a structure-based prioritization of missense variants in a large set of NDD proteins, especially in the 76 genes with no experimental structure available, for which we generated the essential 3D sites in AlphaFold-predicted structures.

Funding

This work was supported by the National Institute of Health (grants: NIH 5U54NS108874, 1R01NS112499-01A1), the Center Without Walls on ion channel function in epilepsy ('Channelopathy-associated Research Center', grant U54 NS108874) and Dravet Syndrome Foundation research grant to D.L., the Fonds National de la Recherche Luxembourg (Research Unit FOR-2715, FNR grant INTER/DFG/21/16394868 MechEPI2) to P.M. and the German Federal Ministry for Education and Research (BMBF, Treat-ION, 01GM1907D) to D.L., T.B. and P.M. P.M. was supported by Treat-ION2 BMBF 01GM2210B.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Emerson E. Deprivation, ethnicity and the prevalence of intellectual and developmental disabilities. *J Epidemiol Community Health*. 2012;66:218-224.
2. Parenti I, Rabaneda LG, Schoen H, Novarino G. Neurodevelopmental disorders: From genetics to functional pathways. *Trends Neurosci*. 2020;43:608-621.
3. Thapar A, Cooper M, Rutter M. Neurodevelopmental disorders. *Lancet Psychiatry*. 2017;4:339-346.
4. Morris-Rosendahl DJ, Crocq MA. Neurodevelopmental disorders-the history and future of a diagnostic concept. *Dialogues Clin Neurosci*. 2020;22:65-72.
5. Jarmasz JS, Basalah DA, Chudley AE, Del Bigio MR. Human brain abnormalities associated with prenatal alcohol exposure and fetal alcohol spectrum disorder. *J Neuropathol Exp Neurol*. 2017;76:813-833.
6. Goeden N, Velasquez J, Arnold KA, *et al*. Maternal inflammation disrupts fetal neurodevelopment via increased placental output of serotonin to the fetal brain. *J Neurosci*. 2016;36:6041-6049.
7. Satterstrom FK, Kosmicki JA, Wang J, *et al*. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*. 2020;180:568-584.e23.
8. Sanders SJ, He X, Willsey AJ, *et al*. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron*. 2015;87:1215-1233.
9. Heyne HO, Singh T, Stamberger H, *et al*. De novo variants in neurodevelopmental disorders with epilepsy. *Nat Genet*. 2018;50:1048-1053.
10. Singh T, Walters JTR, Johnstone M, *et al*. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat Genet*. 2017;49:1167-1173.
11. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*. 2017;542:433-438.
12. Epi25 Collaborative. Ultra-rare genetic variation in the epilepsies: A whole-exome sequencing study of 17,606 individuals. *Am J Hum Genet*. 2019;105:267-282.
13. Kaplanis J, Samocha KE, Wiel L, *et al*. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*. 2020;586(7831):757-762.
14. Heyne HO, Baez-Nieto D, Iqbal S, *et al*. Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci Transl Med*. 2020;12(556):eaay6848.
15. Escayg A, Goldin AL. Sodium channel SCN1A and epilepsy: Mutations and mechanisms. *Epilepsia*. 2010;51:1650-1658.
16. Sanders SJ, Campbell AJ, Cottrell JR, *et al*. Progress in understanding and treating SCN2A-mediated disorders. *Trends Neurosci*. 2018;41:442-456.
17. Richards S, Aziz N, Bale 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*. 2015;17:405-424.
18. Sivley RM, Dou X, Meiler J, Bush WS, Capra JA. Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am J Hum Genet*. 2018;102:415-426.
19. Kamburov A, , Lawrence MS, Polak P, *et al*. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A*. 2015;112:E5486-E5495.
20. Iqbal S, Pérez-Palma E, Jespersen JB, *et al*. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc Natl Acad Sci U S A*. 2020;117:28201-28211.
21. Pandurangan AP, Blundell TL. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Sci*. 2020;29:247-257.
22. Tang Z-Z, Sliwoski GR, Chen G, *et al*. PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biol*. 2020;21(1):217.
23. Kelly M, Park M, Mihalek I, *et al*. Spectrum of neurodevelopmental disease associated with the GNAO1 guanosine triphosphate-binding region. *Epilepsia*. 2019;60:406-418.
24. Olson HE, Demarest ST, Pestana-Knight EM, *et al*. Cyclin-dependent kinase-like 5 deficiency disorder: Clinical review. *Pediatr Neurol*. 2019;97:18-25.
25. Katayama S, Sueyoshi N, Inazu T, Kameshita I. Cyclin-dependent kinase-like 5 (CDKL5): Possible cellular signalling targets and involvement in CDKL5 deficiency disorder. *Neural Plast*. 2020;2020:6970190.
26. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.
27. Tunyasuvunakool K, Adler J, Wu Z, *et al*. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021;596:590-596.
28. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res*. 2000;28:235-242.
29. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res*. 2018;46:2699.
30. Karczewski KJ, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434-443.
31. Danecek P, Auton A, Abecasis G, *et al*. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156-2158.
32. Landrum MJ, Lee JM, Benson MB, *et al*. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46:D1062-D1067.
33. Stenson PD, Mort M, Ball EV, *et al*. The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. 2017;136:665-677.
34. Turner TN, Yi Q, Krumm N, *et al*. denovo-db: A compendium of human de novo variants. *Nucleic Acids Res*. 2017;45:D804-D811.
35. Dewey FE, Murray MF, Overton JD, *et al*. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*. 2016;354:aaf6814.
36. Sudlow C, Gallacher J, Allen N, *et al*. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.
37. Yuan C, Chen H, Kihara D. Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics*. 2012;13:292.
38. Adhikari B, Cheng J. Protein residue contacts and prediction methods. *Methods Mol Biol*. 2016;1415:463-476.
39. Hoksza D, Gawron P, Ostaszewski M, Schneider R. MolArt: A molecular structure annotation and visualization tool. *Bioinformatics*. 2018;34:4127-4128.
40. Lek M, Karczewski KJ, Minikel EV, *et al*. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285-291.
41. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33:2302-2309.
42. Hocker B. Design of proteins from smaller fragments-learning from evolution. *Curr Opin Struct Biol*. 2014;27:56-62.

43. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*. 2009;10:709-720.
44. Perez-Palma E, May P, Iqbal S, *et al*. Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res*. 2020;30:62-71.
45. Traynelis J, Silk M, Wang Q, *et al*. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res*. 2017;27:1715-1729.
46. Lal D, May P, Perez-Palma E, *et al*. Gene family information facilitates variant interpretation and identification of disease-associated genes. *Genome Med*. 2020;12(1):28.
47. Hopf TA, Ingraham JB, Poelwijk FJ, *et al*. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. 2017;35:128-135.
48. Miceli F, Soldovieri MV, Ambrosino P, *et al*. Molecular pathophysiology and pharmacology of the voltage-sensing module of neuronal ion channels. *Front Cell Neurosci*. 2015;9:259.
49. Scheffer IE, Berkovic S, Capovilla G, *et al*. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia*. 2017;58:512-521.
50. Muir AM, Gardner JF, van Jaarsveld RH, *et al*. Variants in GNAI1 cause a syndrome associated with variable features including developmental delay, seizures, and hypotonia. *Genet Med*. 2021;23:881-887.
51. Reynhout S, Jansen S, Haesen D, *et al*. De novo mutations affecting the catalytic Calpha subunit of PP2A, PPP2CA, cause syndromic intellectual disability resembling other PP2A-related neurodevelopmental disorders. *Am J Hum Genet*. 2019;104:139-156.
52. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. *J Mol Biol*. 2013;425:3919-3936.
53. Al Mehdi K, Fouad B, Zouhair E, *et al*. Molecular modelling and dynamics study of nsSNP in STXBP1 gene in early infantile epileptic encephalopathy disease. *Biomed Res Int*. 2019;2019:4872101.
54. McTague A, Nair U, Malhotra S, *et al*. Clinical and molecular characterization of KCNT1-related severe early-onset epilepsy. *Neurology*. 2018;90:e55-e66.
55. Parrini E, Marini C, Mei D, *et al*. Diagnostic targeted resequencing in 349 patients with drug-resistant pediatric epilepsies identifies causative mutations in 30 different genes. *Hum Mutat*. 2017;38:216-225.
56. Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med*. 2021; 27:1666-1669.
57. Lal D, May P, Perez-Palma E, *et al*. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Med*. 2020;12:28.
58. Akdel M, Pires DEV, Porta Pardo E, *et al*. A structural biology community assessment of AlphaFold 2 applications. *bioRxiv 391185*. 2021. doi: 10.1101/2021.09.26.461876
59. Meyer MJ, Lapcevic R, Romero AE, *et al*. mutation3D: Cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum Mutat*. 2016;37:447-456.
60. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol*. 2017;18:225.
61. Geisheker MR, Heymann G, Wang T, *et al*. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat Neurosci*. 2017;20:1043-1051.
62. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH. Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics*. 2010;11:11.
63. Poole W, Leinonen K, Shmulevich I, Knijnenburg TA, Bernard B. Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression. *PLoS Comput Biol*. 2017;13:e1005347.
64. Jubb HC, Saini HK, Verdonk ML, Forbes SA. COSMIC-3D provides structural perspectives on cancer genetics for drug discovery. *Nat Genet*. 2018;50:1200-1202.
65. Ofoegbu TC, David A, Kelley LA, *et al*. PhyreRisk: A dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants. *J Mol Biol*. 2019;431:2460-2466.
66. Stephenson JD, Laskowski RA, Nightingale A, Hurles ME, Thornton JM. VarMap: A web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics*. 2019;35:4854-4856.
67. Liang S, Mort M, Stenson PD, Cooper DN, Yu H. PIVOTAL: Prioritizing variants of uncertain significance with spatial genomic patterns in the 3D proteome. *bioRxiv 2020.06.04.135103*. 2020. doi: 10.1101/2020.06.04.135103
68. Segura J, Sanchez-Garcia R, Sorzano COS, Carazo JM. 3DBIONOTES v3.0: Crossing molecular and structural biology data with genomic variations. *Bioinformatics*. 2019;35:3512-3513.
69. Paznekas WA, Boyadjiev SA, Shapiro RE, *et al*. Connexin 43 (GJA1) mutations cause the pleiotropic phenotype of oculodentodigital dysplasia. *Am J Hum Genet*. 2003;72:408-418.
70. Brunklaus A, Du J, Steckler F, *et al*. Biological concepts in human sodium channel epilepsies and their relevance in clinical practice. *Epilepsia*. 2020;61:387-399.
71. Bellazzi R, Masseroli M, Murphy S, Shabo A, Romano P. Clinical bioinformatics: Challenges and opportunities. *BMC Bioinformatics*. 2012;13(Suppl 14):S1.
72. Mangul S, Martin LS, Eskin E, Blekhman R. Improving the usability and archival stability of bioinformatics software. *Genome Biol*. 2019;20:47.
73. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet*. 2017;100:267-280.
74. Amendola LM, Jarvik GP, Leo MC, *et al*. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. *Am J Hum Genet*. 2016;99:247.
75. Babione JN, Ocampo W, Haubrich S, *et al*. Human-centred design processes for clinical decision support: A pulmonary embolism case study. *Int J Med Inform*. 2020;142:104196.
76. Bates DW, Kuperman GJ, Wang S, *et al*. Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*. 2003;10:523-530.
77. Cai CJ, Reif E, Hegde N, *et al*. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery; Paper 4:1-14. https://doi.org/10.1145/3290605.3300234
78. Castellotti B, Ragona F, Freri E, *et al*. Screening of SLC2A1 in a large cohort of patients suspected for Glut1 deficiency syndrome: Identification of novel variants and associated phenotypes. *J Neurol*. 2019;266:1439-1448.
79. Nickels KC, Zaccariello MJ, Hamiwka LD, Wirrell EC. Cognitive and neurodevelopmental comorbidities in paediatric epilepsy. *Nat Rev Neurol*. 2016;12:465-476.
80. Deng D, Xu C, Sun P, *et al*. Crystal structure of the human glucose transporter GLUT1. *Nature*. 2014;510:121-125.
81. Tung K-F, Pan C-Y, Chen C-H, Lin W-C. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. *Sci Rep*. 2020;10:16245.